Rethinking Algorithm Design for Modern Challenges in Data Science

by

Sitan Chen

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2021

Author . . . .
Department of Electrical Engineering and Computer Science
August 27, 2021

Certified by . . . .
Ankur Moitra
Norbert Wiener Professor of Mathematics
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Rethinking Algorithm Design for Modern Challenges in Data Science

by

Sitan Chen

Submitted to the Department of Electrical Engineering and Computer Science
on August 27, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Heuristics centered around gradient descent and function approximation by neural networks have proven wildly successful for a number of fundamental data science tasks, so much so that it is easy to lose sight of how far we are from understanding *why* they work so well.

Can we design learning algorithms with rigorous guarantees to either match, outperform, or augment these heuristics? In the first part of this thesis, we present new provable algorithms for learning rich function classes like neural networks in natural learning settings where gradient-based methods provably fail. Our algorithms are based on a new general recipe that we call *filtered PCA* for dimensionality reduction in multi-index models.

Asking for rigorous guarantees not only helps uncover general mechanisms that make learning tractable, but also lets us be certain that our algorithms are resilient to the demands of modern data. In the second part of this thesis, we study challenging settings where even a constant fraction of data may have been corrupted and develop new *iterative reweighing schemes* for mitigating corruptions in the context of distribution estimation, linear regression, and online learning. A distinctive feature of many of our results here is that they make minimal assumptions on the data-generating process.

In certain situations however, data may be difficult to work with not because it has been corrupted, but because it comes from a number of heterogeneous sources. In the third part of this thesis, we give improved algorithms for two popular models of heterogeneity, mixtures of product distributions and mixtures of linear regressions, by developing novel ways of using *Fourier approximation*, the *method of moments*, and *combinations thereof* to extract latent structure in the data.

In the final part of this thesis, we ask whether these and related ideas in data science can help shed light on problems in the sciences. We give two such applications, one to rigorously pinning down the much-debated *diffraction limit* in classical optics, and the other to showing memory-sample tradeoffs for *quantum state certification.*

Thesis Supervisor: Ankur Moitra
Title: Norbert Wiener Professor of Mathematics

Acknowledgments

First, I'd like to thank Ankur for being such a fantastic advisor over the last five years. On top of being an endless source of technical wisdom, he's also been so helpful with his incisive big-picture advice and quick to tailor his advising style to suit his students' needs. Having gone into grad school with zero background in ML or statistics, I couldn't have asked for a more patient and encouraging advisor to help me make the jump into learning theory.

I'm also indebted to the incredible guidance I've received outside of MIT. In the summer of 2018 and much of the subsequent academic year, I was fortunate to be hosted by Prasad Raghavendra who graciously taught me all about SDP extension complexity and with whom I very much look forward to collaborating again in my postdoc. This foray into lower bounds also introduced me to Raghu Meka, from whom I've learned many tricks of the trade thanks to his excellent mentorship and our collaborations over the last three years, and whose leaps of insight during research meetings and mathematical aesthetic remain a continual source of inspiration. Working with Raghu in turn introduced me to Adam Klivans, who has been so generous with his ideas and time as a mentor and close collaborator over the last year, whose perspectives on learning have heavily influenced my own research direction, and who gave a ton of helpful feedback on an earlier draft of this thesis. I also want to thank Jerry Li for hosting me during my internship at MSR Redmond in the summer of 2019, teaching me the ropes of robust statistics and answering all my dumb questions about matrix SoS, taking the plunge with me into the world of quantum, and being one of my most frequent collaborators ever since.

I've been so fortunate to be able to work with so many absurdly talented and inspiring coauthors over the last five years. Apart from the ones already mentioned above, this includes Sebastien Bubeck, Jordan Cotler, Michelle Delcourt, Hsin-Yuan Huang, Frederic Koehler, Xiaoxiao Li, Ryan O'Donnell, Guillem Perarnau, Luke Postle, Zhao Song, Runzhou Tao, Morris Yau, Ruizhe Zhang, and Danyang Zhuo.

Going further back, the idea of studying TCS in grad school might never have entered my mind had I not been lucky enough to spend two wonderful summers in undergrad working under the amazing guidance of Thomas Steinke, Salil Vadhan, and Les Valiant. Going even further back, I owe the mentorship of Jesse Geneson and Tanya Khovanova at the Research Science Institute for convincing me to pursue a career in research.

Outside of research, playing the piano has been a huge part of my life at MIT. I want to thank David Deveau, Eileen Huang, Natalie Lin Douglas, and Marcus Thompson for an incredible job running the Emerson Fellows seminar. I especially want to thank David; our weekly lessons have been a consistent source of joy and fulfillment throughout my last decade in Cambridge.

A big factor in my decision to go to MIT was its famously tight-knit theory group, and I'm so grateful to have been a part of such a warm and vibrant community. Thank you to Debbie, Joanne, and Rebecca for always making ToC a fun group to be a part of and for organizing so many great social events for the group even in pandemic times. And to all my friends at MIT and elsewhere, thanks for making the last five years so enjoyable. I'll miss the notsogitcs Asian food trips, the always entertaining and opinionated conversations with the G5 lunch crowd, crossword puzzles every Friday, ping pong matches and tennis outings, and hangouts around the G6 espresso machine. I hope we can all stay in touch.

Lastly, I want to thank all the loved ones who have supported me along this journey:

To Chunsheng Chen and Bingbing Yuan, thank you for all the holiday meals over the last nine years and for always making me feel at home in Cambridge.

To Heeyoon, my lodestar, I imagine dating an academic can be a pretty rough deal at times, and yet by some miracle you take the whole ordeal in stride. You fill my heart to the brim with light and I am so stupidly blessed to have you in my life.

Finally, to my beloved parents and grandmother, I don't know what I did in my previous life to deserve you, but thank you for sacrificing everything to raise me and giving me every opportunity to be the best version of myself. 谁言寸草心，报得三春晖.

Contents

List of Figures

Chapter 1

Introduction

1.1 Algorithmic Opportunities in Data Science

Given data generated from some unknown process, what can we learn about that process? This question lies at the heart of modern data science, and the techniques that machine learning practitioners have developed over the last decade to answer it have had and will continue to have a profound impact on almost every facet of society. Given the astounding empirical successes of these techniques however, it is easy to lose sight of the fact that we don't actually understand why they work so well. What is more, the sobering truth is that these techniques are largely centered around a small toolbox of decades-old heuristics, and the overwhelming majority are variations on a theme– run gradient descent on a nonconvex objective and hope for the best.

In this thesis we take the opposite tack and ask: by limiting ourselves to this algorithmic toolbox, are we leaving something on the table? After all, algorithmic creativity has been the workhorse behind success stories in so many other areas of computer science: Reed-Solomon decoding for digital storage and communication, interior point methods for convex optimization, shortest path algorithms for network routing protocols, Cooley-Tukey for digital signal processing, and the list goes on. What if we seriously considered the possibility there might be new and better algorithms for tackling modern challenges in data science?

In the coming chapters, we will see the following picture emerge from asking this question. For a wide range of tasks that involve discerning complex structure from data, from a practical standpoint there might be an “obvious” or “popular” choice of heuristic that one might expect to do the job. But as it turns out, these heuristics can fail horribly in very natural settings. Our main contribution will be to devise genuinely new algorithmic primitives that outperform existing techniques by circumventing these failure modes and *provably* solving the task at hand. To name two notable examples:

- **Deep neural networks**: In practice, the algorithm of choice for training a neural network on data to get good out-of-sample performance would be to run gradient descent. Surprisingly however, there are no-go theorems showing this can fail even if there exists a network that achieves zero test error and the input vectors are sampled from a benign distribution like a Gaussian [GGJ+20, DKKZ20]! In Chapter 3, we give a new algorithm to provably learn deep neural networks in this setting (see Theorem 3.4.2). Not only is this the first provable guarantee of its kind to handle *arbitrary depth*, but it also identifies the *first example of a neural network class which is efficiently learnable, but provably not via gradient descent.*

- **Robust regression**: One of the go-to algorithms for performing regression on a corrupted dataset is to minimize a loss function, e.g. the Huber loss, which is less sensitive to outliers than the square loss. As we show in Chapter 6, such an approach is provably suboptimal even in the well-studied Huber contamination model for linear regression (see Theorem 6.9.1). Instead, we devise the first algorithm for this problem to run in polynomial time and essentially achieve the information-theoretically optimal error guarantee (see Theorem 6.1.1). In contrast to practically all recent works in robust statistics, our algorithm makes *no assumptions on the distribution generating the data.*

What do we learn from these results? Apart from providing new recipes for solving these particular problems, results like these offer general *prescriptions* for how to reason about when these and related tasks are tractable and in particular which approaches can or cannot work.

For instance, our neural network result teaches us that there are sophisticated wrappers we can build on top of tried and tested subroutines in data science that allow us to solve a much wider family of supervised learning problems than previously believed, and the only way to arrive at these wrappers was by stress-testing existing techniques from practice and asking for provable end-to-end guarantees. Our techniques ultimately provide a way around a recurring bottleneck in supervised learning by suggesting a new answer to a basic question: apart from gradient queries, *are there more powerful statistics we can leverage about the dataset to extract rich structure?*

Similarly, stress-testing algorithms even for age-old questions like linear regression on which many more complex learning systems are built teaches us new ways to be robust in challenging settings *where even the uncontaminated parts of the data are misbehaved.* Indeed, while a number of algorithms like Huber regression and more modern robust statistics approaches work quite well when the data is evenly spread, this turns out not to be the right assumption to work with in many realistic settings. For example, one can imagine kernelized settings where infinite-dimensional data gets passed through a complicated feature map or online settings where data arrives in a dynamic fashion. In these cases, our approaches based on fundamentally new ways of reweighing the data to mitigate outliers are not only helpful but, as we will see, even necessary.

Finally, another benefit of honing in on new algorithms for data science is that the tools we end up developing can also teach us about problems in other fields. In this thesis, we also give a number of applications to *inverse problems in the sciences.* Here, algorithmic and statistical thinking are important not only in supplying concrete approaches, but even in suggesting the right ways to rigorously frame problems in the first place. After all, much of scientific discovery revolves around extracting signal from data, but if we limit how we work with and reason about data to heuristics, there's a real risk that we're not getting the full picture.

To name one example from our results in this vein (see Chapters 9 to 11), it turns out that some of the ideas that go into our new algorithms for learning from heterogeneous data in Chapters 7 and 8 help shed new light on an old debate from optics dating back to work of 19th century physicists like Lord Rayleigh and Ernst Abbe:

- **Diffraction limit in optical systems**: In classical optical systems like telescopes, the physics of diffraction makes point sources of light appear blurred. It has been a subject of fierce debate (see Section 9.9) whether this imposes fundamental limits on how well one can resolve closely spaced point sources. Nowadays, introductory optics textbooks [Hec15, Ken08] commonly cite the so-called *Abbe limit* as the critical level of separation below which resolution becomes impossible. In Chapter 9, we frame this question in the language of learning *mixture models* and rigorously prove for the first time that *the Abbe limit is actually not the right limit*! We complement this with various algorithms above and below the true diffraction limit.

This is a case in point that statistical thinking has the potential to clarify even well-established scientific debates, and new algorithms for learning have the potential to make new discoveries.

We now describe the contributions of this thesis in greater detail. In Sections 1.1.1 to 1.1.3, we survey the general themes and motivations for the data science tasks we consider, and in Section 1.2 we informally overview the results proved in this thesis.

1.1.1 Learning Rich Function Classes

In this section we focus on the well-studied setting of **supervised learning**. Here, we get access to samples from a distribution $\mathcal{D}$ which generates pairs (x, y), where x is some *feature vector* sampled from a distribution $\mathcal{D}_{\mathbf{x}}$, the *label* y is a (possibly noisy) function of x, and the goal is to output a function $\widehat{f}$ approximating the conditional expectation $\mathbb{E}[y|x]$ under some metric. For instance, x might be a picture of a traffic sign, y might be 0 or 1 depending on whether or not x depicts a stop sign, and the figure of merit might be the *misclassification error*, i.e. the probability that $\widehat{f}(x) \neq y$ for x sampled from $\mathcal{D}_{\mathbf{x}}$.

In practice, the algorithm of choice for this problem would be to run stochastic gradient descent to train a large and deep neural network on enough examples and take $\widehat{f}$ to be the resulting network. This heuristic has proven capable of achieving superhuman performance on a host of image classification benchmarks [HZRS15, THK^{+}21].

These empirical results suggest that 1) there exist neural networks which can closely approximate $\mathbb{E}[y|x]$ for real-world supervised learning tasks, and 2) gradient descent can efficiently find these networks. How might we try to rigorously justify this phenomenon? For starters, we could adopt 1) as a hypothesis and attempt to prove 2) as a consequence.

In fact, we could make life even easier and assume that $\mathbb{E}[y|x]$ is closely approximated by an extremely simple neural network, say, a linear separator. That is, we might assume that there exists a vector w^* such that

$$\Pr_{(x,y)\sim\mathcal{D}}[y \neq \mathrm{sgn}(\langle w^*, x\rangle)] = 0.00001\%. \tag{1.1}$$

In the absence of further assumptions however, one can design distributions $\mathcal{D}$ over (x, y) for which gradient descent fails to find a good classifier. In fact the situation is far worse: for such distributions, *no efficient algorithm* can find *any classifier*, even a highly nonlinear one, that achieves better than 50.00001% accuracy [Dan16]![1] In learning-theoretic parlance, it is hard to agnostically learn halfspaces, even improperly.

The reader might rightfully wonder whether such hardness results stem from the fact that we aren't assuming anything about the "missing 0.00001%," i.e. about the specific structural reasons why the data is not perfectly captured by a linear separator. So what if we made our lives *even* easier and strengthened the assumption (1.1) by assuming that there exists a vector w^* that achieves *zero* misclassification error? This is the so-called *realizable* case, as the data is perfectly realized by a linear separator. In the realizable case, it has been known for some time that gradient descent with respect to a a suitable convex surrogate loss more or less suffices [Ros58, Byl94, BFKV98].

But what about more sophisticated functions than linear separators, e.g. deep neural networks? Practice would suggest this is a problem perfectly suited for training a neural network via gradient descent. After all, in the realizable case, we are literally promised that there exists a neural network achieving zero train and test error, and all we have to do is find it! Yet without making further assumptions on distribution $\mathcal{D}_{\mathbf{x}}$ over feature vectors, learning neural networks even in the realizable case is known to be NP-hard [BR89, Vu06]. This begs the following question which will be the focus of the first part of this thesis:

> ***Question 1:*** *Are there natural settings where one can prove that rich classes of functions, e.g. neural networks, are learnable in the realizable case?*

[1] This holds under a plausible complexity-theoretic conjecture regarding random constraint satisfaction problems.

As mentioned above, and as we will describe in Section 1.2.1 and subsequently in greater detail in Chapters 2 and 3, the main contribution of the first part of this thesis will be to give the first natural setting where efficiently learning neural networks and related function classes in the realizable case is possible, but *provably not via gradient descent.*

1.1.2 Learning From Untrustworthy Data

As the discussion in Section 1.1.1 suggests, asking for provable guarantees helps shed light on the mechanisms that make efficient learning possible which in turn leads to the development of new algorithmic primitives. In this section we turn to another motivation for targeting provable guarantees: without a principled understanding for how practical heuristics seem to work so well, there is a genuine danger in deploying them without discretion. Here we highlight common pitfalls that motivate the algorithms we develop in the next part of this thesis:

Adversarial corruptions From a security standpoint, it is well-known that neural networks are susceptible to data poisoning attacks in which small adversarial corruptions to the training data can dramatically skew the behavior of the resulting classifier [SKL17, CLL$^+$17]. The threat these attacks pose is particularly salient in the context of learning from data obtained in a decentralized fashion. For instance, data labels obtained from crowdsourcing platforms are notoriously dubious in quality [KCB$^+$20, CK20], making such platforms an easy target for data poisoning [TXSS20]. Similar vulnerabilities apply to federated learning [BEMGS17] and crowd sensing [MLX$^+$18] where data is distributed across many users or servers, some fraction of which might be compromised.

Even in settings where there might not be an explicit attacker, there are myriad other challenges posed by datasets in the wild; below, we discuss the ones most pertinent to this thesis.

Heavy-tailed behavior The distributions generating real-world data are often **heavy-tailed**; in supervised learning contexts, this applies both to the distribution over features $\mathcal{D}_{\mathbf{x}}$ and to the conditional distribution of y given x. As examples of the former, in standard scene recognition benchmarks, there are many object classes which appear in a very small fraction of images [ZAR14, WRH17], and a similar phenomenon applies to frequencies of different alleles and haplotypes in population genetics [SLG$^+$15]. As examples of the latter, in financial markets, sharp price fluctuations are much more likely to occur than a Gaussian model would suggest [BT03, Rac03], and in microarray data, gene expression intensities typically follow a power law behavior [PH05].

Dynamic, non-i.i.d. data In some situations like bandit or reinforcement learning, data may even be generated in an online fashion, rather than i.i.d. from some fixed distribution, further compounding the problems introduced by corrupted and highly noisy data. For example, corruptions due to intermittent loss of power or Internet are a known issue in mobile health deployments of algorithms for contextual bandits [HATC$^+$19, PPG18], one of the prototypical models for sequential decision-making. Corruptions due to fraudulent clicks pose a similar problem for deployments of these algorithms for ad recommendations [IJMT05, DGZ12].

In light of these challenges, the second part of this thesis will focus on designing algorithms to answer the following question:

> ***Question 2:*** *How do we mitigate the effect of noisy and untrustworthy data, especially in settings where even the "clean" parts of the data are heavy-tailed, dynamically generated, or otherwise misbehaved?*

Related questions have been studied for some time in the robust statistics literature, and in Chapters 4 to 6, we give a detailed overview of known results and clarify the ways in which they come up short. In this part of the thesis, we study Question 2 in the context of *distribution estimation*, *linear regression*, and *contextual bandits*, giving the first algorithms to obtain near-optimal statistical guarantees in the presence of corrupted data for these problems. We overview the relevant models and our results in Section 1.2.2 and provide the technical details in Chapters 4 to 6.

In the third part of the thesis, we ask how the algorithmic landscape for such problems changes if we assume additional structure on the noise inherent in real-world data. Specifically, we focus on structure that arises from *heterogeneity*.

Heterogeneity In many settings, fluctuations in the data primarily stem from the fact that the data comes from a number of heterogeneous sources. Returning to the example of microarray data above, a notorious confounding factor is the presence of so-called *batch effects*: it can be difficult to extract biological signals from data simply because it is often aggregated from experiments that were performed at different times, in different labs, or even on different microarray platforms [JLR07, CGB+11]. Similarly in association studies, as mentioned at the outset, heterogeneity can stem from *population stratification*, i.e. the presence of multiple genetic subpopulations in the data [CHRZ07,SAT96]. For these reasons, we ask:

> ***Question 3:*** *What are the most powerful algorithmic primitives for discerning subpopulation structure from heterogeneous data?*

In the third part of this thesis we answer Question 3 by giving faster algorithms for two well-studied *mixture models* (stylized models of data with subpopulation structure). Importantly, as we describe in Section 1.2.3 and subsequently in greater detail in Chapters 7 and 8, the algorithmic primitives we design help shed new light on the connections between two powerful techniques in learning theory for handling such problems: Fourier analysis and the method of moments.

1.1.3 Data Science and the Sciences?

In the final part of this thesis, we explore whether these ideas can say anything in other domains beyond learning theory. We give two such applications.

Heterogeneity and Optics If one points a telescope at the night sky, under ideal conditions one will discern bright, blurred disks in place of stars. Rather than some artifact of atmospheric turbulence or imperfections in the lens, these patterns, so-called *Airy disks*, emerge naturally from the physics of diffraction. Ever since the pioneering work of 19th century physicists like Lord Rayleigh, Ernst Abbe, and Sir George Biddell Airy, it has been widely believed that this blurring introduced by these disks imposes a fundamental limit on how well one can resolve nearby light sources [Air35,Ray79,Abb73]. Despite a century and a half of persistent debate (see Section 9.9) and numerous attempts to pin down the *diffraction limit*, many of which are covered in standard introductory texts on optics [Hec15, Ken08], a rigorous definition has remained elusive. That said, the consensus has largely been that it occurs at the *Abbe limit*. In fact, the press release for the 2014 Nobel Prize in Chemistry, awarded for the development of new optical systems that can resolve at much smaller distances, singled the Abbe limit out as the barrier being shattered [Nob14].

From the perspective of learning from heterogeneous data however, the diffraction limit has a natural interpretation. An image consisting of several Airy disks simply arises from a collection of photons, each sampled from one of a number of sources, striking the observation plane. By casting the problem of resolution as one of learning a mixture model and leveraging Fourier-analytic tools reminiscent of those we used to give algorithms for learning from heterogeneous data, we are able to rigorously pinpoint the diffraction limit as a statistical phase transition in the sample complexity for this problem. We give further details in Section 1.2.3 and defer the technical details to Chapter 9.

Dynamic Data and Quantum Learning With noisy intermediate-scale quantum computing looming on the horizon, it is timely to explore what ideas in machine learning on classical computers are transferable to the quantum setting, where analogues of even the most basic classical learning problems remain open. Take for instance the following popular spin on the guiding question we asked at the outset of this chapter: given samples from a distribution, what can we say about the distribution? Classically, one of the most elementary instantiations of this is *uniformity testing*: given i.i.d. samples from an unknown distribution over d elements, is the distribution uniform or far from being uniform? The quantum analogue of this, *mixedness testing*, is just as natural: given the ability to measure many copies of an unknown quantum state in d dimensions, is it the *maximally mixed state* or far from being maximally mixed? Whereas by now we have a deep understanding of the former question and practically any variant of it one could try to ask, the precise answer to the latter largely remains a mystery. This gap is all the more pressing given that certifying whether a state one has prepared satisfies certain properties seems to be a basic prerequisite for a variety of quantum data science tasks in the lab [dSLCP11, WGFC14, AGKE15].

There are a number of subtleties that arise in the quantum setting. For instance, whereas in uniformity testing we can only interact with the distribution by receiving samples, in mixedness testing we can choose the kinds of measurements we conduct, and in particular our choices of measurements can depend *adaptively* on previous measurement outcomes. This should be reminiscent of the discussion above on learning from dynamic, non-i.i.d. data like in bandit settings, and indeed, as we will see in Section 1.2.4 and Chapters 10 and 11, techniques for proving bandit lower bounds will be key to shedding light on the sample complexity of mixedness testing and related tasks like *quantum state certification*.

1.2 Our Contributions

In this section we elaborate on the specific models we study and our contributions:

1. In Section 1.2.1, we answer Question 1 by giving a new algorithmic primitive, *filtered PCA*, for learning rich function classes in high dimensions, notably in settings where standard approaches like gradient descent provably fail. This is the subject of Chapters 2 and 3, based on the following two works:

   - S. Chen, R. Meka. Learning Polynomials of Few Relevant Dimensions. *Proceedings of the 33rd Annual Conference on Learning Theory (COLT 2020).*
   - S. Chen, A.R. Klivans, R. Meka. Learning Deep ReLU Networks Is Fixed-Parameter Tractable. *Proceedings of the 62nd Annual IEEE Symposium on Foundations of Computer Science (FOCS 2021).*

2. In Section 1.2.2 we answer Question 2 by describing new approaches based on novel ways of iteratively reweighting the data that obtain the first near-optimal recovery guarantees for a number of basic statistical questions like distribution estimation, regression, and contextual bandits in the presence of corruptions and heavy-tailed noise. This is the subject of Chapters 4 to 6, based on the following three works:

   - S. Chen, J. Li, A. Moitra. Efficiently Learning Structured Distributions from Untrusted Batches. *Proceedings of the 52nd Annual ACM Symposium on Theory of Computing (STOC 2020).*

   - S. Chen, J. Li, A. Moitra. Learning Structured Distributions from Untrusted Batches: Faster and Simpler. *Advances in Neural Information Processing Systems (NeurIPS 2020).*

   - S. Chen, F. Koehler, A. Moitra, M. Yau. Online and Distribution-Free Robustness: Regression and Contextual Bandits with Huber Contamination. *Proceedings of the 62nd Annual IEEE Symposium on Foundations of Computer Science (FOCS 2021).*

3. In Section 1.2.3 we answer Question 3 by giving improved algorithms for learning two popular mixture models, *mixtures of product distributions* and *mixtures of linear regressions*. Notably, our techniques give new ways to exploit the Fourier transform and the method of moments, in some cases simultaneously, in distribution learning. This is the subject of Chapters 7 and 8, based on the following two works:

   - S. Chen, A. Moitra. Beyond the Low-Degree Algorithm: Mixtures of Subcubes and Their Applications. *Proceedings of the 51st Annual ACM Symposium on Theory of Computing (STOC 2019).*

   - S. Chen, J. Li, Z. Song. Learning Mixtures of Linear Regressions in Subexponential Time via Fourier Moments. *Proceedings of the 52nd Annual ACM Symposium on Theory of Computing (STOC 2020).*

   As our first application to the sciences, we apply related ideas to give a rigorous interpretation for the *diffraction limit* in classical optics; surprisingly, we prove that this limit does not occur at the widely accepted *Abbe limit*. This is the subject of Chapter 9, based on the following work:

   - S. Chen, A. Moitra. Algorithmic Foundations for the Diffraction Limit. *Proceedings of the 53rd Annual ACM Symposium on Theory of Computing (STOC 2021).*

4. In Section 1.2.4, motivated by parallels between quantum and online learning tasks, we give a general framework for proving information-theoretic lower bounds for quantum testing problems. Using this, we give the first memory-sample tradeoffs for mixedness testing, as well as nearly instance-optimal bounds on the sample complexity of quantum

state certification. This is the subject of Chapter 10 and 11, based on the following two works:

- S. Bubeck, S. Chen, J. Li. Entanglement is Necessary for Optimal Quantum Property Testing. *Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science (FOCS 2020).*
- S. Chen, J. Li, R. O'Donnell. Towards Instance-Optimal Quantum State Certification With Independent Measurements. arxiv:2102.13098

1.2.1 Filtered PCA

In this part of the thesis, we describe a new algorithmic primitive for learning rich function classes in high dimensions.

Models and Assumptions We begin by clarifying the specific generative models we will study in this part of the thesis. We will consider the following standard setup for supervised learning:

Definition 1.2.1 (Distribution-specific PAC learning). *Let $\mathcal{F}$ be some known class of functions $F : \mathbb{R}^d \to \mathbb{R}$, and let $\mathcal{D}_{\mathbf{x}}$ be some known distribution over $\mathbb{R}^d$. Given error parameter $\varepsilon > 0$, we will say that an algorithm* (properly) PAC learns $\mathcal{F}$ over $\mathcal{D}_{\mathbf{x}}$ *in time T and sample complexity N if the following holds: given samples $(x_1, y_1), \ldots, (x_N, y_N)$ for some unknown $F \in \mathcal{F}$, where $x_1, \ldots, x_N$ are i.i.d. samples from $\mathcal{D}_{\mathbf{x}}$ and $y_i = F(x_i)$, the algorithm runs in time T and outputs $\widehat{F} \in \mathcal{F}$ for which*

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathbf{x}}}\left[(F(x) - \widehat{F}(x))^2\right] \le \varepsilon^2$$

with high probability over the randomness of the samples and the algorithm. We will sometimes refer to the distribution over (x, y) samples as $\mathcal{D}$.

Motivated by the question of learning rich function classes like neural networks, we will focus on function classes $\mathcal{F}$ consisting of functions of the following form:

Definition 1.2.2 (Multi-index model). *Let $k, d \in \mathbb{N}$ be parameters satisfying $k \le d$ (typically* *we think of k as much smaller than d). Given a matrix $V \in \mathbb{R}^{k\times d}$ and a function $h : \mathbb{R}^k \to \mathbb{R}$, the function $F : \mathbb{R}^d \to \mathbb{R}$ given by*

$$F(x) = h(Vx)$$

is a multi-index model *with* link function h *and* relevant subspace *given by the row span of V.*

Multi-index models have had a long history of study in statistics [DH18, PVY17, NWL16, Bri12, Li92, PV16, YBL17, BB$^+$18, Li91, HJS01, HJP$^+$01, DJS08]. From our perspective, they represent an expressive testbed for developing algorithmic and analytic tools for PAC learning. For one, because multi-index models depend on a low-dimensional projection of the input (that is, we typically think of k as much smaller than d), they represent an appealing class of functions for which one could hope to avoid the curse of dimensionality. Additionally, they capture the following two classes of rich function families which will be the focus of the algorithms we develop (we refer to Section 2.1.2 for a discussion of other commonly studied function classes that can be realized as multi-index models).

Definition 1.2.3 (Low-rank polynomials)**.** *Fix parameters $m, k, d \in \mathbb{N}$ with $k < d$. Consider the class of link functions consisting of all polynomials $p : \mathbb{R}^k \to \mathbb{R}$ of degree m. We refer to multiple-index models whose link function is of this form as* rank-k polynomials of degree m in d dimensions*, or* low-rank polynomials *for short when the parameters are clear from context.*

This is a significant generalization of the *phase retrieval* problem (where the underlying function is given by $F(x) \triangleq \langle v, x\rangle^2$ for some $v \in \mathbb{R}^d$), which has been the subject of a long line of work in the signal processing and machine learning communities, see e.g. [CSV13, CLS15, CEHV15, NJS13, NWL16] and the references therein. Over inputs from the Boolean hypercube $\{\pm 1\}^d$, rank-k polynomials of degree k can also encode arbitrary *k-juntas*, i.e. Boolean functions which only depend on k bits of their output [MOS03]. Furthermore, as we discuss immediately preceding Section 2.1.1, PAC learning low-rank polynomials can also be thought of as a variant of tensor decomposition.

Note that if the goal were simply to output any degree-m polynomial (rather than a low-rank one) that approximates F, one trivial baseline would simply be to run *polynomial* *regression.* But this would take $O(d^m)$ time and samples, whereas information-theoretically the sample complexity for this problem should only need to scale *linearly* in the ambient dimension d when k is bounded.

We now turn to the second class of functions we will work with:

Definition 1.2.4 (Neural networks)**.** *Given weight matrices*

$$\mathbf{W}_0 \in \mathbb{R}^{k_0 \times d}, \mathbf{W}_1 \in \mathbb{R}^{k_1 \times k_0}, \dots, \mathbf{W}_L \in \mathbb{R}^{k_L \times k_{L-1}}, \mathbf{W}_{L+1} \in \mathbb{R}^{1 \times k_L},$$

consider the function

$$F(x) \triangleq \mathbf{W}_{L+1}\sigma\left(\mathbf{W}_L\sigma\left(\cdots\sigma(\mathbf{W}_0 x)\cdots\right)\right),$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is some activation applied entrywise. In this work, we focus on $\sigma(z) = \mathsf{ReLU}(z) \triangleq \max(0, z)$. We say that F is computed by a (feedforward) ReLU network with depth $L + 2$ and size $S \triangleq \sum_{i=0}^{L} k_i$. *Note that F is a multiple-index model with relevant subspace of dimension equal to the rank of $\mathbf{W}_0$.*

As we discuss in Section 3.1.1, the problem of PAC learning neural networks has been the subject of intense study in the learning theory literature. In the last few years alone there have been many papers giving provable results for learning restricted classes of neural networks under various settings [JSA15, ZLJ16, ZSJ+17, BG17, GKKT17, LY17, ZPS17, Tia17, GKM18, DLT18, GLM17, GKLW18, MR18, BJW18, GK19, AZLL19, VW19, ZYWG19, DGK+20, GMOV18, LMZ20, DK20], though in the learning setting we consider, all relevant prior work only pertains to neural networks of depth 2, that is, functions of the form $F(x) = \mathbf{W}_1\phi(\mathbf{W}_0 x)$.

Similar to low-rank polynomials, neural networks when restricted to inputs from $\{\pm 1\}^d$ can implement arbitrary *k-juntas*, where k is the dimension of the relevant subspace (see Appendix 3.5.2). As it is widely conjectured to be impossible to PAC learn general k-juntas in time better than $O(d^k)$ when $\mathcal{D}_{\mathbf{x}}$ is the uniform distribution over $\{\pm 1\}^d$, it is necessary for us to restrict our attention to more "benign" choices of $\mathcal{D}_{\mathbf{x}}$. In this part of the thesis, we therefore work with the following standard distributional assumption:

Assumption 1 (PAC learning under Gaussian inputs– Chapters 2 and 3)**.** *In Definition 1.2.1, we will take $\mathcal{D}_{\mathbf{x}}$ to be the standard Gaussian distribution in d dimensions, which we will denote by $\mathcal{N}(0, \mathrm{Id})$.*

We note that the vast majority of the aforementioned works on phase retrieval and PAC learning neural networks use this assumption, in some sense the most "benign" high-dimensional distributional assumption one could hope to give provable guarantees for. We also note that it is straightforward to extend to the case where $\mathcal{D}_{\mathbf{x}}$ is Gaussian with *non-identity covariance*: simply estimate its covariance and "whiten" the dataset to simulate samples from $\mathcal{N}(0, \mathrm{Id})$ instead.

Existing Approaches In this section we survey existing approaches for PAC learning multi-index models and highlight their shortcomings. The starting point for all of these approaches is to note that because a multi-index model by definition only depends on the projection of its input to the relevant subspace, the main challenge is to approximately recover this subspace. Upon recovering this subspace, one can reduce the dimensionality of the problem to that of the relevant subspace by projecting, thereby avoiding runtime that scales prohibitively in the ambient dimension d which is often much larger than k. A natural way to exploit Assumption 1 to do this would be to consider the statistic

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y \cdot (xx^\top - \mathrm{Id})]. \tag{1.2}$$

The motivation for this is the following basic result:

Fact 1.2.5. *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ is a multi-index model with relevant subspace V and for which $\mathbb{E}_{\mathcal{D}}[y]$ is finite. Then under Assumption 1, if $v \in \mathbb{R}^d$ is orthogonal to V, then it lies in the kernel of $\mathbb{E}_{\mathcal{D}}[y \cdot (xx^\top - Id)]$.*

Proof. Without loss of generality suppose $\|v\| = 1$. Because $\mathcal{D}_{\mathbf{x}}$ is standard Gaussian in d dimensions, the random variable y is independent of $\langle v, x\rangle$ if v is orthogonal to V. Therefore

$$v^\top \mathbb{E}[y \cdot (xx^\top - \mathrm{Id})]v = \mathbb{E}[y \cdot (\langle v, x\rangle^2 - 1)] = \mathbb{E}[y] \cdot \mathbb{E}[\langle v, x\rangle^2 - 1] = 0,$$

where in the last step we used that $\langle v, x\rangle$ is distributed as a standard univariate Gaussian. □

Fact 1.2.5 suggests a simple algorithm for recovering the relevant subspace: estimate the matrix (1.2) from samples and output the span of all eigenvectors with non-negligible eigenvalue. Indeed, this is a standard approach in the multi-index model literature [PVY17, Bri12, Li92, PV16, YBL17, DKKZ20] and is also a standard preprocessing step in algorithms for phase retrieval [NJS13, CLS15, NWL16]. It can also be extended to higher-order statistics of the form $\mathbb{E}[y \cdot T(x)]$ for appropriate choices of *tensor-valued* function $T(x)$, which forms the basis for a number of provable algorithms for PAC learning depth-2 neural networks [JSA15, ZSJ+17, BJW18, DK20].

The key drawback of this technique is that the converse of Fact 1.2.5 need not hold! That is, the nontrivial part of the spectrum of $\mathbb{E}[y \cdot (xx^\top - \mathrm{Id})]$ may not reveal the full relevant subspace. For instance, what if $\mathbb{E}[y \cdot (xx^\top - \mathrm{Id})]$ or any of its tensor-valued analogues $\mathbb{E}[y \cdot T(x)]$ were simply zero? In fact, it was shown in [GGJ+20, DKKZ20] that there exist networks for which this is the case and that this poses a fundamental roadblock towards learning general depth-2 neural networks. In particular, they constructed families of neural networks for which any *correlational statistical query (CSQ) algorithm*, namely any algorithm that only looks at statistics of the form $\mathbb{E}[y \cdot g(x)]$ for arbitrary functions g, must essentially run in time $d^{\Omega(S)}$, where S is the size of the network.

Notably, the popular heuristic of simply training a (possibly overparametrized) student network by running noisy gradient descent to minimize the square loss is another prominent example of a CSQ algorithm and thus cannot be used to efficiently PAC learn ReLU networks under Assumption 1! On the one hand, this suggests that real-world data is fairly non-Gaussian. But from an algorithm designer's point of view, this also presents an exciting opportunity to design genuinely new algorithms that provably outperform ones used in practice, under a natural distributional assumption.

Our Results and Techniques We now turn to a rough overview of the new approach we develop for learning multi-index models, before informally stating the main theorems of this section. We will defer more detailed technical overviews and proofs to Chapters 2 and 3.

The idea to circumvent the impossibility result of [GGJ+20, DKKZ20] is to instead look at statistics of the form $\mathbb{E}[\phi(y) \cdot (xx^\top - \mathrm{Id})]$ for an appropriate choice of $\phi : \mathbb{R} \to \mathbb{R}$. It is easy to see that the proof of Fact 1.2.5 generalizes to matrices of this form, so the hope is still to recover some part of the relevant subspace by looking at the eigendecomposition of this matrix, and the main difficulty is how to select ϕ to ensure that the matrix $\mathbb{E}[\phi(y) \cdot (xx^\top - \mathrm{Id})]$ is actually nonzero. While it is tempting to consider analytic functions like a polynomial, exponential, or trigonometric function for ϕ, these turn out to be quite difficult to analyze. Instead, our key idea is to look at ϕ given by a *step function*. Specifically, for some appropriately chosen τ, we will work with

$$\phi(z) \triangleq \begin{cases} 1 & |z| \geq \tau \\ 0 & |z| < \tau. \end{cases}$$

To show that $\mathbb{E}[\phi(y) \cdot (xx^\top - \mathrm{Id})]$ is nonzero, it suffices to show that its trace $\mathbb{E}[\phi(y) \cdot (\|\Pi x\|^2 - k)]$ is nonzero, where Π is the projection to the relevant subspace. We can ensure this by selecting τ large enough that $\phi(y) = 1$ only if $\|x\|^2 > k$; this choice of τ is problem-dependent. One can interpret this algorithm as projecting to the nontrivial eigenvectors of the covariance of the conditional distribution on x after filtering out all data points (x, y) for which $|y|$ is small. For this reason, we refer to this approach as *filtered PCA*, which we summarize informally in the pseudocode below:

Algorithm 1: FilteredPCA($\{(x_1, y_1), \ldots, (x_N, y_N)\}, \tau$)

```
Throw out all points $(x_i, y_i)$ for which $|y_i| < \tau$.
Using all remaining points, form $\mathbf{M} = \sum_i x_i x_i^\top$.
return top principal components of $\mathbf{M}$
```

This is our general recipe for recovering at least one vector from the relevant subspace. It remains to show how to recover the rest of the subspace, and this turns out to be much more complicated and problem-dependent. At a high level, the idea will be to filter out data points (x, y) for which some other function of y is small, where this function is defined in terms of vectors in the relevant subspace that have been found so far.

Using this approach, we are able to prove the following guarantees for PAC learning low-rank polynomials and ReLU networks which, by the lower bounds of [GGJ+20, DKKZ20], provably cannot be obtained by gradient descent.

Theorem 1.2.6 (Informal, see Theorem 2.1.3)**.** *Under Assumption 1, there is an algorithm for PAC learning non-degenerate*[2] *rank-k polynomials of degree m in d dimensions to error ε with sample complexity $N = O_{k,m}\left(d\log(d)^{O(m)}\log^2(1/\varepsilon)\right)$ and runtime $T = \widetilde{O}_{k,m}(Nd)$.*[3]

Note that when the degree and rank of the polynomial are constants, the sample complexity and runtime of our algorithm specialize to $N = \widetilde{O}(d\log^2(1/\varepsilon))$ and $T = \widetilde{O}(d^2\log^2(1/\varepsilon))$. In particular, our dependence on the ambient dimension is near-optimal. We stress that this is in sharp contrast to the junta setting mentioned above, i.e. where $\mathcal{D}_{\mathbf{x}}$ is uniform over the hypercube. There, while information-theoretically it also suffices to take a number of samples that scales linearly in d, computationally it is conjectured that no algorithm can run in time better than the "polynomial regression baseline" of $d^{O(m)}$.

We also remark that the reason we are able to obtain a logarithmic rather than polynomial dependence on $1/\varepsilon$ is that we additionally give a gradient-based algorithm for refining an initial estimate that is suitably close to the ground truth (we obtain this initial estimate using the above filtered PCA approach). The gradient-based algorithm implements geodesic stochastic gradient on a suitable Riemannian manifold; the analysis is quite technical and we defer the overview and details to Chapter 2.

Finally, we turn to our guarantees for learning neural networks:

Theorem 1.2.7 (Informal, see Theorem 3.4.2)**.** *Under Assumption 1, there is an algorithm for PAC learning ReLU networks of size S of size S to error ε with sample complexity $N = O(d)\cdot\exp\left(\mathrm{poly}(S/\varepsilon)\right)$ and runtime $\widetilde{O}(d^2)\cdot\exp\left(\mathrm{poly}(S/\varepsilon)\right)$.*

Similar to Theorem 1.2.6, our guarantee scales near-optimally in the ambient dimension d. Note that for some absolute constant $c > 0$, this algorithm runs in time polynomial in d provided $S/\varepsilon = O(\log^c d)$. As we discuss in Chapter 3, such a result was not even known for general depth-2 networks, and to date this is the only known result for PAC learning neural networks of depth greater than 2 over Gaussian inputs. And to reiterate, thanks

[2] See Definition 2.3.1 for a precise definition of this condition, and the discussion preceding Definition 2.3.1 for an explanation of why such an assumption is needed.

[3] In this thesis, we use $\widetilde{O}(f)$ and $\widetilde{\Omega}(f)$ to denote $O(f\log^c(f))$ and $\Omega(f/\log^c(f))$ for an absolute constant $c > 0$. A subscript indicates that the hidden constant factor may depend arbitrarily on the quantities in the subscript.

to aforementioned lower bounds of [GGJ+20, DKKZ20], Theorem 1.2.7 provably cannot be achieved by noisy gradient descent.

1.2.2 New Iterative Reweighing Schemes

In this part of the thesis, we turn to the question of designing provable algorithms for learning under the additional constraint that the data may be untrustworthy. We begin by introducing the models we will study in this part of the thesis.

Model 1: Learning from Untrusted Batches The first, originally introduced in [QV17], is motivated by the discussion at the beginning of Section 1.1.2 about the effect of adversarial contaminations on data that was collected in a decentralized fashion, e.g. via a crowdsourcing platform or via users in a federated learning or crowdsensing setup. Suppose we aggregate a bunch of samples from users, each a draw from some unknown distribution $\mathcal{D}$. As a concrete example, perhaps we are interested in building a mobile spellcheck feature and would like to learn the distribution over misspellings of a particular word. It is reasonable to imagine that some small fraction of users supply faulty data, either because of device issues or for malicious reasons. How well can we learn $\mathcal{D}$ in this case?

There has been an explosion of progress in recent years on robust distribution learning, yielding the first algorithms for efficiently learning Gaussians [DKK+19a, LRV16], mixtures of Gaussians [BK20, Kan20, DHKK20, BDJ+20, LM21b, LM21a], and graphical models [DKSS21, PSBR20, CDKS18] from corrupted samples. Here, we instead focus on arbitrary distributions $\mathcal{D}$ over discrete domains.

The first thing to ask is how well one can learn such a distribution from a dataset of N i.i.d. samples, where some ηN of them have been arbitrarily corrupted. Unfortunately, regardless of how big N is, it is information-theoretically impossible to estimate $\mathcal{D}$ to error better than $O(\eta)$– the corruptions can be chosen in such a way that the dataset could equally plausibly be N uncorrupted i.i.d. samples from an arbitrary η-perturbation of $\mathcal{D}$ in total variation.

One useful feature of the above decentralized learning applications however is that the samples collected from users typically come in *batches*: rather than see a single independent sample from each compliant or malicious user, we might see several! This turns out to enable much better learning guarantees. First, let us formalize the setting just described:

Definition 1.2.8 (Learning from untrusted batches [QV17])**.** *Let $\mathcal{D}$ be an unknown distribution over $\{1, \ldots, n\}$. We are given m batches, each consisting of k samples. Each uncorrupted batch $i \in \{1, \ldots, m\}$ has the property that its samples were drawn i.i.d. from some unknown distribution $\mathcal{D}_i$ that is ω-close in total variation distance to $\mathcal{D}$. Moreover $(1 - \eta)m$ of the batches are uncorrupted, though the indexes of these batches are not known to the learner. The remaining ηm batches are arbitrarily corrupted. In fact, an adversary is allowed to choose the contents of the corrupted batches after observing all of the uncorrupted batches.*

[QV17] proposed an exponential-time algorithm for learning $\mathcal{D}$ to within $O(\omega + \eta/\sqrt{k})$ error in total variation distance. They also showed a matching information-theoretic lower bound showing that achieving $o(\omega + \eta/\sqrt{k})$ error is impossible in general, leaving as an open question whether one can match this lower bound with an *efficient* algorithm. As we describe below, we will answer this in the affirmative.

Beyond giving an efficient algorithm for learning general discrete distributions from untrusted batches though, one can also ask for more refined guarantees that take into account additional structure in $\mathcal{D}$. There is a long line of work in statistics on getting minimax rates for learning various classes of structured distributions from samples (see Section 4.1.3 for references). As we will show, our techniques can be extended to obtain guarantees of this nature even in the untrusted batches setting. We defer a formal definition of the class of structured distributions we consider to Definition 4.5.3 but note that they capture a wide array of examples including monotone, multi-modal, log-concave, and monotone hazard rate distributions, as well as mixtures thereof.

Model 2: Robust Regression We now turn our attention back to supervised learning, specifically the basic task of linear regression, and ask the same question as above: can we design algorithms that can tolerate a constant fraction of the data being corrupted?

Of course, this is by no means a new question. It has long been known that ordinary least-squares is highly sensitive to the presence of even a small number of outliers, and traditionally the fix, dating back to work of Huber [Hub64,Hub73] and even as far back as the 1700s [Bos57], is to minimize a loss function that is less sensitive than the square loss, e.g. the *Huber loss* or the absolute value loss [Chi20,L$^+$17,ZJS20,DT19]. One other classical approach originally proposed by Legendre [LS59] is that of *least trimmed squares*: run ordinary least squares, throw out points with large residual, and repeat on the remaining data [BJKK17, BJK15, SBRJ19]. Recent progress in high-dimensional robust statistics has also led to a number of new provable algorithms for robust regression in a variety of challenging settings [KKM18, BP20, ZJS20, CAT$^+$20].

The following model of robust regression is studied in many of these works (we discuss how it compares to other commonly studied models in Section 6.3):

Definition 1.2.9 (Huber-contaminated linear regression, informal– see Chapter 6)**.** *Fix noise parameter $\sigma > 0$ and dynamic range parameter $R > 0$. Let $\mathcal{D}_{\mathbf{x}}$ be a distribution over the unit ball, and let $\theta^* \in \mathbb{R}^d$ be an unknown vector satisfying $\|\theta^*\| \le R$. The learner receives a collection of samples $(x_1, y_1), \ldots, (x_N, y_N)$ generated via the following experiment:*

1. *Sample $x_1, \ldots, x_N \in \mathbb{R}^d$ independently from $\mathcal{D}_{\mathbf{x}}$.*

2. *Nature selects a random, unknown subset $S \subset \{1, \ldots, N\}$ of size $\eta \cdot N$.*

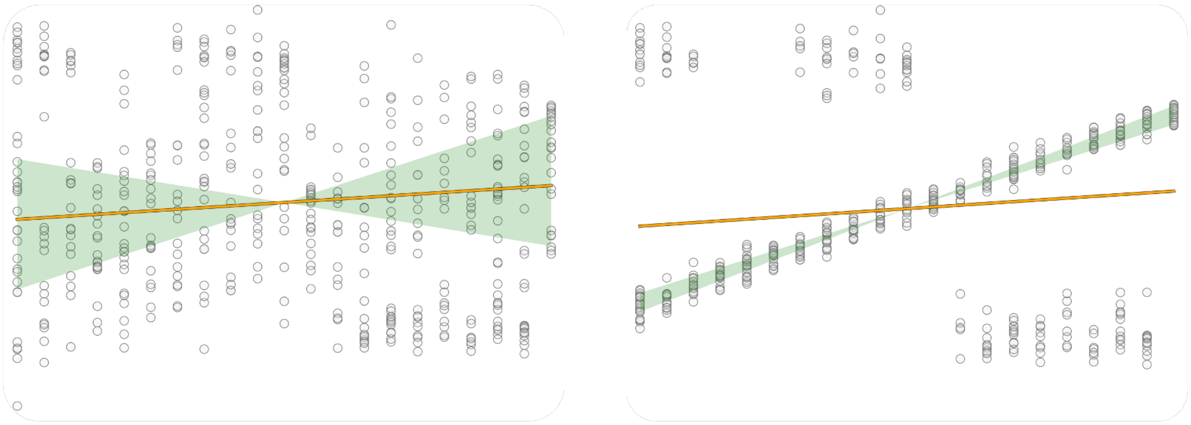
3. *For every $i \notin S$, define $y_i \triangleq \langle \theta^*, x_i \rangle + \xi_i$ for $\xi_i \sim \mathcal{N}(0, \sigma^2)$.*

4. *For every $i \in S$, an all-powerful adversary chooses the value of y_i.*

The goal of the learner is to produce a vector θ for which the clean mean squared error *$\mathbb{E}_{x \sim \mathcal{D}_{\mathbf{x}}}[\langle \theta^* - \theta, x \rangle^2]$ is small.*[4]

Note that simply running ordinary least squares on the dataset already achieves clean mean squared error $O(\eta R^2)$. On the other hand, by a simple reduction from one-dimensional robust mean estimation (see Example 6.4.3), no algorithm can achieve clean mean squared error better than $O(\eta^2 \sigma^2)$. In particular, when σ^2 is comparable to R^2, it is very easy to achieve the information-theoretically optimal level of error, but this is because not that much

[4] As will become clear in the analysis, it will be straightforward to extend our techniques to handle the *non-realizable* case where the dataset is not given by a linear ground truth, and we simply want to *compete* with the optimal linear predictor on the uncorrupted data.



(a) When σ^2 is comparable to R^2, many lines (range depicted in green), including the one found by ordinary least squares (orange), fit the data equally well (although the fit is not that good to begin with).

(b) When σ^2 is much smaller than R^2, then ordinary least squares (orange) fails, but in principle it should be possible to do much better even in high-dimensions.

Figure 1-1: Datasets with equal contamination rates but different levels of noise σ. The corruptions are located in the upper left and bottom right parts of both figures. The goal in robust regression is to achieve low square loss on the *uncorrupted points*. We depict in orange the ordinary least squares estimator and in green the range of linear predictors that would perform comparably to what our algorithms can achieve.

can be learned about θ^* in the first place (see the left panel of Figure 1-1 for an illustration). In contrast, we will be interested in the natural setting where σ is much smaller than R.

Unfortunately, as we discuss at length throughout Sections 6.1 and 6.3, one drawback of existing alternatives to ordinary least squares is that they make strong assumptions on the distribution $\mathcal{D}_{\mathbf{x}}$ over the covariates, ruling out the possibility of applying these guarantees to the settings discussed in Section 1.1.2 where the covariates can be heavy-tailed or dynamically generated.

As we describe below, in this thesis we give the first algorithms for Huber-contaminated linear regression that work *for arbitrary distributions* $\mathcal{D}_{\mathbf{x}}$ and obtain near-optimal clean mean squared error. In fact, our techniques are general enough to apply to fixed-design and online settings where the covariates can even depend on previous data and actions of the learner. We defer the specifics of these settings to Section 6.5.1 and Definition 6.4.1 respectively. We also remark that our techniques extend naturally to settings where $\mathcal{D}_{\mathbf{x}}$ is supported over an infinite-dimensional space and the x_i's are implicitly represented by a feature map, e.g. in kernel ridge regression.

Sum-of-Squares and Alternating Minimization Many of the recent algorithms for robust statistics are based on searching for a subset of the data which satisfies similar structural properties as the set of uncorrupted data. For instance, to robustly estimate the mean of a spherical Gaussian $\mathcal{N}(\theta, \mathrm{Id})$, [DKK+19a] give algorithms that exploit the fact that any large subset of the samples whose covariance is not too "spiky" in any one direction must have empirical mean close to the true mean of the distribution.

Implicit in many of these works is the following generic recipe for mitigating the effect of corruptions. First, let $\mathcal{A} : \mathcal{S} \to \Theta$ be a "non-robust" algorithm for solving a particular statistical task, where $\mathcal{S}$ is the universe of all possible datasets and Θ is the set of possible underlying parameters for the problem. For instance, if the task is estimating the mean of $\mathcal{N}(\theta, \mathrm{Id})$, Θ is the set $\mathbb{R}^d$ of all possible means, and given a dataset, $\mathcal{A}$ might simply compute its empirical mean.

Now comes the difficult and highly problem-dependent part. Given a corrupted dataset $\mathcal{Z} = \{z_1, \dots, z_N\}$, we need to design a *penalty function* $\ell : 2^{[N]} \times \Theta \to \mathbb{R}_{\geq 0}$ that takes a subset $S \subseteq [N]$ and an estimate $\theta \in \Theta$ and outputs a proxy for how "wrong" we would be if we guessed that the points in S were the uncorrupted points and took θ to be our estimate for the ground truth parameter. For one, we want *completeness*: when S is the true set of uncorrupted points and θ is the result of applying $\mathcal{A}$ to S, $\ell(S, \theta)$ should be small. Crucially, we also want *soundness*: when $\ell(S, \theta)$ is small and S is as big as the set of uncorrupted points, then θ is close to the ground truth.

Example 1.2.10. *For estimating the mean of a spherical Gaussian when an η-fraction of samples have been corrupted, [DKK+19a] takes $\ell(S, \theta)$ to be*

$$\left\| \frac{1}{|S|} \sum_{i \in S} (z_i - \theta)(z_i - \theta)^\top - \mathrm{Id} \right\|_2 . \tag{1.3}$$

If S were truly the set of uncorrupted samples and θ were their empirical mean, then (1.3) would be close to zero by standard concentration, so completeness holds. The key structural result shown in [DKK+19a] is that if $\ell(S, \theta)$ is small and $|S| \geq (1 - \eta)N$ then θ is close to the true mean, so soundness also holds.

Given $\mathcal{A}$, dataset $\mathcal{Z}$ of which an η-fraction has been corrupted, and ℓ, the generic recipe is then simply to solve

$$\min_{\substack{S \subseteq [N]: \\ |S| \geq (1-\eta)N}} \ell(S, \mathcal{A}(\{z_i\}_{i \in S})). \tag{1.4}$$

While there is no reason *a priori* for this to be efficiently solvable, in certain situations it is possible to relax (1.4) into a tractable optimization problem.

One existing approach is to form a convex relaxation via *sum-of-squares programming.* This forms the basis for some of the algorithms in this part of the thesis, but because the setup for this approach is rather technical, we defer the details to Chapter 4.

We now describe another common approach for relaxing (1.4) based on *alternating minimization.* First, note that we can relax the minimization over S in (1.4) to a minimization over a collection of *weights* $a = \{a_1, \dots, a_N\}$ for which $0 \leq a_i \leq 1$ for all i and $\sum_{i=1}^N a_i = (1-\eta)N$. In particular, one valid choice of weights would be to take $a_i = 1$ if sample z_i is uncorrupted and $a_i = 0$ otherwise; more generally, a_i can be thought of as a score for how confident we are that z_i is uncorrupted.

Letting $\Delta_{N,\eta}$ denote the set of such weights a, we can consider more general types of penalty functions ℓ over $\Delta_{N,\eta} \times \Theta$ instead of $2^{[N]} \times \Theta$. Likewise, we can consider more general types of non-robust algorithms $\mathcal{A}$ over $\Delta_{N,\eta} \times \mathcal{S}$ instead of $\mathcal{S}$ (for instance, in mean estimation, we could take $\mathcal{A}(a, \mathcal{Z})$ to be the *weighted empirical mean* $\frac{1}{(1-\eta)N} \sum_i a_i z_i$).

We then arrive at the following new optimization problem:

$$\min_{a \in \Delta_{N,\eta}} \ell(a, \mathcal{A}(a, \mathcal{Z})).$$

While this is a bilevel optimization problem and thus potentially still intractable, there is a natural heuristic for solving such a problem, *alternating minimization*:

```
Algorithm 2: AltMin($\eta, \mathcal{Z}, \mathcal{A}, \ell$)
Initialize $a$ arbitrarily.
for $1 \leq t \leq T$ do
    $\theta \leftarrow \mathcal{A}(a, \mathcal{Z})$.
    $a \leftarrow \min_{a \in \Delta_{N,\eta}} \ell(a, \theta)$.        // alternatively, take a gradient/MW step
return $\theta$
```

As the value of a is repeatedly updated by solving $\min_a \ell(a, \theta)$, taking a gradient step, or performing a multiplicative weight update, we can think of AltMin as *iteratively reweighing* the dataset to gradually dampen the influence of the corrupted data points on $\mathcal{A}$.

We remark that this idea has been used extensively in the literature on robust mean estimation, where there are by now a number of proofs that variants of AltMin converge to a good estimate for the mean [DKK+19a, ZJS20, CDGS20, HLZ20].

Our Results and Techniques– Learning From Untrusted Batches As we mentioned above, the specific details of how to implement AltMin (e.g. how to design ℓ) and how to analyze it are highly problem-dependent, and the main contribution in this part of the thesis will be to extend this framework to other natural statistical tasks, namely learning from untrusted batches and Huber-contaminated regression. We begin with our results for learning from untrusted batches, as they are closest in spirit to the ideas that go into robust mean estimation.

The starting point is to note that learning from untrusted batches amounts to robustly learning *multinomial distributions*. Given a batch of k i.i.d. draws from a distribution $\mathcal{D}$ over $\{1,\dots n\}$, consider the k-dimensional vector z of frequencies with which each element of the domain appears in the batch. This is by definition distributed as a sample from the multinomial distribution $\mathrm{Mul}_k(\mathcal{D})$ given by k draws from $\mathcal{D}$. Furthermore, as a distribution over $\mathbb{R}^n$, $\mathrm{Mul}_k(\mathcal{D})$ has mean $\theta^* \cdot k$, where θ^* is the vector whose i-th entry is the probability mass $\mathcal{D}$ assigns to element i of the domain.

As a result, learning from untrusted batches is *equivalent* to estimating the mean of a multinomial distribution in L_1 *norm* from corrupted samples. In light of this connection to robust mean estimation, it is natural to try to adapt (1.3) as follows. As the norm under which we need our estimate θ to be close to θ^* is L_1 instead of L_2, by the dual formulation of L_p norms we would like $\max_{v\in\{\pm 1\}^n}|\langle\theta-\theta^*,v\rangle|$ to be small. Consequently it would make sense to modify (1.3) by considering

$$\ell(a,\theta) \triangleq \max_{v:\{\pm 1\}^n} \sum_i a_i \langle z_i - \theta, v\rangle^2.$$

Actually, it turns out that to obtain near-optimal bounds with this approach, one should use higher-degree information by instead considering

$$\ell(a,\theta) \triangleq \max_{v:\{\pm 1\}^n} \sum_i a_i \langle z_i - \theta, v\rangle^t \tag{1.5}$$

for sufficiently large t. Incorporating this into a sum-of-squares relaxation, we are able to obtain the following warmup result:

Theorem 1.2.11 (Informal, see Theorem 4.4.1)**.** *Let $\mathcal{D}, k, n, \eta, \omega$ be as in Definition 1.2.8. There is an $(nk/\eta)^{polylog(1/\eta)}$-time algorithm for learning from untrusted batches that estimates $\mathcal{D}$ to within $O\left(\frac{\eta}{\sqrt{k}}\sqrt{\log 1/\eta} + \omega\right)$ in total variation distance.*

Note that this matches the information-theoretic lower bound of $\Theta(\omega + \eta/\sqrt{k})$ up to a $\sqrt{\log(1/\eta)}$ factor.

We then turn to the more challenging question of obtaining refined sample complexity guarantees when $\mathcal{D}$ is structured. Using a result of [ADLS17], it suffices to estimate $\mathcal{D}$ relative to a different norm than L_1, namely

$$\|\cdot\|_{\mathcal{A}_\ell} \triangleq \max_{\substack{v\in\{\pm 1\}^n \text{ with} \\ \leq 2\ell \text{ sign changes}}} |\langle \cdot, v\rangle|,$$

where the maximization is over bitstrings which, when read from left to right, change in sign at most 2ℓ times (see Section 4.5.1). The natural way to adapt (1.5) would be to constraint the maximization over $v \in \{\pm 1\}^\ell$ to vectors with bounded sign changes, but the key difficulty is that this is much harder to capture in an efficient way using a sum-of-squares relaxation. We show how to do so in Section 4.5 and 4.6 via a novel relaxation based on Haar wavelets.

It turns out however that by using an alternating minimization approach instead, we can achieve even better runtime and sample complexity. Specifically, we consider the following penalty function:

$$\ell(a,\theta) \triangleq \min_{\substack{v\in\{\pm 1\}^n \text{ with} \\ \leq 2\ell \text{ sign changes}}} \left\{ \langle vv^\top, \Sigma_{a,\theta}\rangle - \frac{1}{k}(\langle v, \mu_a\rangle - \langle v, \mu_a\rangle^2) \right\}, \tag{1.6}$$

where $\Sigma_{a,\theta} \triangleq \frac{1}{(1-\eta)N}\sum_{i=1}^N a_i(z_i - \theta)(z_i-\theta)^\top$ and $\mu_a \triangleq \frac{1}{(1-\eta)N}\sum_{i=1}^N a_i z_i$. More precisely, we consider a semidefinite relaxation of this penalty function, inspired by the aforementioned sum-of-squares relaxation based on Haar wavelets. The details are quite technically involved, and we defer a discussion of where (1.6) comes from and how to analyze it, as well as relevant prior and concurrent work, to Chapter 5.

Plugging this choice of penalty function into AltMin, we obtain a polynomial-time algorithm for learning structured distributions from untrusted batches to error that also matches the information-theoretic lower bound up to a log factor:

Theorem 1.2.12 (Informal, see Theorem 5.4.1)**.** *Let $\mathcal{D}, k, n, \eta, \omega$ be as in Definition 1.2.8. Suppose $\mathcal{D}$ is approximated by an s-part piecewise polynomial function with degree at most d. There is a polynomial-time algorithm for learning from untrusted batches that estimates $\mathcal{D}$ to within $O\left(\frac{\eta}{\sqrt{k}}\sqrt{\log(1/\eta)} + \omega\right)$ in total variation distance using $\widetilde{O}\left((s^2 d^2/\varepsilon^2) \cdot \log^3(n)\right)$ batches.*

Our Results and Techniques– Huber-Contaminated Regression We conclude this section by describing our contributions for regression in the presence of corruptions. To motivate our choice of penalty function $\ell(a, \theta)$, we first ask what distinguishing structural properties the uncorrupted part of the data satisfies. For instance, if the distribution $\mathcal{D}_{\mathbf{x}}$ over covariates were, say, isotropic sub-Gaussian, then any univariate projection of the uncorrupted x's in the dataset would have bounded higher moments. Indeed, this idea is the basis for essentially all recent works on robust regression, e.g. [KKM18, BP20, ZJS20, CAT+20]. The main issue of course is that in our setting we make no such assumptions about $\mathcal{D}_{\mathbf{x}}$.

Instead, our choice of $\ell(a, \theta)$ will arise from the following toy calculation. Recall from Definition 1.2.9 that the figure of merit in Huber-contaminated regression is the clean mean-squared error $\mathbb{E}[\langle \theta^* - \theta, x\rangle^2]$, which is well-approximated by

$$\frac{1}{N}\sum_{i=1}^{N} \langle \theta^* - \theta, x_i\rangle^2$$

For convenience, define the weighting $a^* \in \Delta_{N,\eta}$ by $a^*_i = 1$ if i is uncorrupted and $a^*_i = 0$ otherwise.

First observe that because the uncorrupted points are a *random* $(1-\eta)$*-fraction of the dataset*, the total squared error over the uncorrupted points *subsamples* the total squared error over all points, so

$$\frac{1}{N}\sum_{i=1}^{N} a_i^*\langle\theta^*-\theta, x_i\rangle^2 \approx \frac{1-\eta}{N}\sum_{i=1}^{N}\langle\theta^*-\theta, x_i\rangle^2. \tag{1.7}$$

Now given any other weighting $a \in \Delta_{N,\eta}$, we can write $a_i^* = a_i^* a_i + a_i^*(1-a_i) \leq a_i^* a_i + (1-a_i)$. Substituting this into (1.7), we conclude that

$$\frac{1}{N}\sum_{i=1}^{N} a_i^* a_i\langle\theta^*-\theta, x_i\rangle^2 + \frac{1}{N}\sum_{i=1}^{N}(1-a_i)\langle\theta^*-\theta, x_i\rangle^2 \gtrsim \frac{1-\eta}{N}\sum_{i=1}^{N}\langle\theta^*-\theta, x_i\rangle^2. \tag{1.8}$$

The first term corresponds to the contribution from uncorrupted points that the weighting a correctly identifies as uncorrupted, and the second comes from points that the weighting a identifies as corrupted. Here is the main idea: even though a need not behave like the indicator of a random $(1-\eta)$-fraction of the data, let us pretend that it does. If so, then using the same subsampling idea from above, we would have that

$$\frac{1}{N}\sum_{i=1}^{N}(1-a_i)\langle\theta^*-\theta, x_i\rangle^2 \approx \frac{\eta}{N}\sum_{i=1}^{N}\langle\theta^*-\theta, x_i\rangle^2. \tag{1.9}$$

The punch line is that if we substitute (1.9) into (1.8) and rearranged, we would obtain

$$\frac{1}{(1-2\eta)N}\sum_{i=1}^{N} a_i^* a_i\langle\theta^*-\theta, x_i\rangle^2 \gtrsim \frac{1}{N}\sum_{i=1}^{N}\langle\theta^*-\theta, x_i\rangle^2.$$

In other words, as long as η is bounded away from $1/2$,[5] we can upper bound the clean mean squared error (right-hand side) in terms of the error of θ on uncorrupted points correctly identified by a (left-hand side). At this point we are done: it turns out that if we incorporate into the penalty function the term $\sum_{i=1}^{N} a_i(y_i - \langle\theta, x_i\rangle)^2$, one can control this term without too much difficulty (see Section 6.5 for details).

The key step in this argument was the ansatz (1.9). How do we ensure this holds for the (a, θ) that our algorithm finds? We would like to enforce that (1.9) holds, but it de-

[5] If $\eta \geq 1/2$, note that the problem is information-theoretically impossible: the corrupted labels might come from another linear model, in which case it is impossible to discern which linear model is the ground truth

pends on θ^*. The workaround is simply to enforce that $\frac{1}{N}\sum_{i=1}^N (1-a_i)\langle v, x_i\rangle^2$ approximates $\frac{\eta}{N}\sum_{i=1}^N \langle v, x_i\rangle^2$ for *all* $v \in \mathbb{R}^d$. This is simply a spectral constraint, namely that *the covariance of the dataset under the weighting a subsamples that of the full dataset.* We could either impose this as a hard constraint or, in the framework of AltMin, define $\ell(a,\theta)$ to include a regularizer corresponding to this subsampling constraint:

$$\ell(a,\theta) = \sum_{i=1}^N a_i(y_i - \langle \theta, x_i\rangle)^2 + \lambda \left\| \frac{1}{N}\sum_{i=1}^N (1 - a_i - \eta) x_i x_i^\top \right\|_2 .$$

Incorporating this into AltMin, we obtain the following near-optimal guarantee for Huber-contaminated regression:

Theorem 1.2.13 (Informal, see Theorem 6.1.1)**.** *Suppose that $\eta < 0.4999$. There is a polynomial-time algorithm for Huber-contaminated regression achieving clean mean-squared error $O(\eta^2\sigma^2\log(1/\eta))$ with high probability.*

We refer to Chapter 6 for explicit sample complexity bounds, noting that they specialize to minimax rates for ordinary least squares when $\eta = 0$. We also emphasize that the clean mean squared error that our algorithm achieves matches the aforementioned information-theoretic lower bound of $\Omega(\eta^2\sigma^2)$ up to a log factor. And as discussed in the footnote above, our algorithm works for η all the way up to the information-theoretic limit of 1/2; in robust statistics jargon, our estimator achieves *optimal breakdown point.*

As alluded to above, we can also extend Theorem 1.2.13 to handle a variety of other settings where the process generating the samples is non-stochastic (i.e. *fixed-design*) or even adaptive based on the data that has been generated so far. In fact, our techniques can handle more challenging settings where, instead of minimizing clean mean squared error which is intrinsically an *offline* guarantee, we minimize a suitable notion of *regret* for various online problems. Specifically, we show

Theorem 1.2.14 (Informal, see Theorems 6.7.2 and 6.7.4)**.** *Suppose that $\eta < 0.4999$. There is a polynomial-time algorithm for Huber-contaminated online regression (see Definition 6.4.1) achieving clean square loss regret $O\left(\sigma^2\eta^2\log(1/\eta)\right)T + \mathrm{poly}(R,\sigma,\eta)\cdot o(T)$ with high probability.*

Theorem 1.2.15 (Informal, see Theorems 6.8.1 and 6.8.2)**.** *Suppose that $\eta < 0.4999$. There is a polynomial-time algorithm for Huber-contaminated linear contextual bandits with action space of size K (see Definition 6.4.4) that achieves clean square loss regret $O\left(\sigma\eta\sqrt{K\log(1/\eta)}\right)\cdot T + \mathrm{poly}(R, K, \eta, \sigma) \cdot o(T)$.*

We note that the leading-order term in both of these bounds is optimal up to log factors.

It may appear surprising that our algorithms for dealing with Huber contamination do not use an established approach like minimizing the Huber or L_1 loss. This is because there are fundamental reasons that *neither* of these approaches can match our strong guarantees in the distribution-free setting. In fact, we prove a lower bound showing the failure of any M-estimator based on minimizing a convex loss function:

Theorem 1.2.16 (Lower bound against convex M-estimators, informal version of Theorem 6.9.1)**.** *There is an instance of Huber-contaminated linear regression for which no vector θ obtained by minimizing a convex loss with respect to the Huber-contaminated distribution over (x, y)'s can achieve clean mean squared error better than $\Omega(\eta^3 R\sigma)$.*

1.2.3 Heterogeneity, Moments, and the Fourier Transform

In this part of the thesis, we study the related question of learning from heterogeneous data. We begin by describing the three models we focus on; these will all be examples of *mixture models*, a prototypical way to model heterogeneous distributions by expressing them as convex combinations of simpler distributions. After defining the models, we will describe our techniques and results, a common theme being new recipes for exploiting the moments and Fourier transform of the underlying distribution.

Definition 1.2.17 (Mixture models)**.** *Let $\mathcal{C}$ be some known class of distributions over a domain D. Given $k \in \mathbb{N}$,* mixing weights *$\lambda_1, \ldots, \lambda_k \in [0, 1]$ encoding a distribution λ over $\{1, \ldots, k\}$, and distributions $\mathcal{D}_1, \ldots, \mathcal{D}_k \in \mathcal{C}$, the corresponding* mixture of k distributions from $\mathcal{C}$ *is the following distribution over D. To sample once from this distribution, sample $i \in \{1, \ldots, k\}$ from λ and then sample from $\mathcal{D}_i$.*

One of the most famous families of mixture models is *mixtures of Gaussians*, where $\mathcal{C}$ is taken to be the class of all Gaussian distributions in d dimensions. The question of learning mixtures of Gaussians from samples has been the subject of intensive study in the theoretical computer science literature in recent years [Das99, DS00, AK01, VW02, AM05, BV08, KMV10, MV10, BS15, HP15, HK13, GHK15, RV17, HL18, KS17, DKS18b], and while we do not focus on them in this thesis, they will serve as a source of motivation for some of the questions we consider.

Model 1: Mixtures of Product Distributions We begin by considering the following mixture model over the Boolean hypercube.

Definition 1.2.18 (Mixtures of product distributions)**.** *Given $n \in \mathbb{N}$, a* product distribution *over $\{0,1\}^n$ is any product of n Bernoulli random variables. We refer to the vector in $[0,1]^n$ consisting of the means of these Bernoullis as the* center *of the product distribution. A* mixture of product distributions *is then defined according to Definition 1.2.17.*

This is an incredibly rich family of discrete distributions and has been studied in a number of contexts where one would like to learn from heterogeneous data. It captures the Dawid-Skene model in crowdsourcing [DS79] where one would like to estimate the competencies of a population of workers from noisy observations, population stratification in genetics [SRH07, CHRZ07] where one would like to control for the fact that there are many genetically differentiated subpopulations represented in the data, and user profiling in recommendation systems [TM$^+$14] where one would like to use user ratings to produce vectors encoding the tastes of every user. Outside of learning contexts, mixtures of product distributions also abound in the theory of nonlinear large deviations, where there are a number of structural results showing that certain "low-complexity" Gibbs measures on product spaces are well-approximated in Wasserstein distance by mixtures of product distributions [EG18, Aus19, Aus20].

We will be especially interested in the following class of mixtures of product distributions:

Definition 1.2.19 (Mixtures of subcubes)**.** *Given $n \in \mathbb{N}$ and a string $s \in \{0,1,\star\}^n$, the subcube associated to s is the set of all strings in $\{0,1\}^n$ that agree with s in all $\{0,1\}$-valued coordinates. We refer to any mixture of uniform distributions over such subcubes as a* mixture of subcubes*. Equivalently, a mixture of subcubes is a mixture of product distributions whose centers all lie in $\{0,1/2,1\}^n$.*

Mixtures of subcubes capture a variety of natural distributions that arise in the context of PAC learning Boolean functions. For instance, given a junta, a decision tree, or an instance of sparse parity with noise, the uniform distribution over positively labeled bitstrings is a mixture of subcubes. A more interesting example is the following class of probabilistic functions, which should be thought of as decision trees with some number of nodes given by stochastic transitions.

Definition 1.2.20 (Stochastic decision trees)**.** *A* stochastic decision tree *T on n bits is a tree with leaves labeled by 0 or 1 and with internal nodes of two types:* decision nodes *and* stochastic nodes*. Each decision node is labeled with some $i \in [n]$ and has two outgoing edges, one labeled with 0 and the other with 1. Each stochastic node u has some number of outgoing edges uv each labeled with a probability p_{uv} such that $\sum_v p_{uv} = 1$.*

T defines a joint probability distribution D_T on $\{0,1\}^n \times \{0,1\}$ as follows. The $x \in \{0,1\}^n$ is sampled uniformly at random. Then given x, the conditional distribution can be sampled from by walking down the tree as follows. At a decision node labeled with i, traverse along the edge labeled by x_i. At a stochastic node u with outgoing edges labeled p_{uv}, pick edge uv with probability p_{uv} and traverse along that edge. When we reach a leaf node, output its value b. In this case we say that x evaluates to b *along this path.*

It turns out that the distribution over x randomly sampled from $\{0,1\}^n$ conditioned on x evaluating to 1 under a k-leaf stochastic decision tree is a mixture of k subcubes. As a result, algorithms for learning mixtures of subcubes from samples automatically give rise to algorithms for learning stochastic decision trees (see Section 7.9).

Lastly, we clarify what we mean by "learning" a mixture of product distributions or subcubes. Note that in general, the parameters of the mixture may not be "identifiable," that is, it may not be information-theoretically possible to recover the centers of such a mixture. As an extreme example, note that the uniform distribution over $\{0,1\}^n$ can be realized both as a mixture of a single subcube (the entire hypercube) and as a mixture of 2^n subcubes (one for every vertex of the hypercube). Instead, our learning goal will be the following:

Definition 1.2.21 (Density Estimation)**.** *Given independent samples from a distribution*

$\mathcal{D}$*, we say that* $\mathcal{A}$ *is an algorithm for* density estimation *achieving error* ε *in total variation distance if it outputs the description of a distribution* $\mathcal{D}'$ *for which* $d_{TV}(\mathcal{D}, \mathcal{D}') \le \varepsilon$ *with high probability over the randomness of the samples and the algorithm.*

Model 2: Mixtures of Linear Regressions Next, we study the following heterogeneous variant of linear regression.

Definition 1.2.22 (Mixtures of Linear Regressions)**.** *Fix* $d \in \mathbb{N}$*, noise parameter* $\sigma > 0$*, and a distribution* $\mathcal{D}_{\mathbf{x}}$ *over covariates. Given a vector* $v \in \mathbb{R}^d$*, consider the distribution over* (x, y) *where* x *is sampled from* $\mathcal{D}_{\mathbf{x}}$ *and* $y = \langle v, x\rangle + \xi$ *where* $\xi \sim \mathcal{N}(0, \sigma^2)$*. A* mixture of linear regressions *(MLR) is a mixture of such distributions in the sense of Definition 1.2.17. If* $\{v_1, \ldots, v_k\}$ *are the vectors for the components of the mixture, we say that the mixture is* Δ-separated *if* $\|v_i - v_j\| \ge \Delta$ *for all* $i \neq j$*.*

This model has applications to problems ranging from trajectory clustering [GS99] to phase retrieval [BCE06, CSV13, NJS13] and has received significant attention in the learning theory literature as a natural non-linear generative model for supervised data [FS10, CYC13, CL13, YCS14, YCS16, ZJD16, SJA16, KYB17, BWY17, KQC+18, LL18, KC19].

Note that our Theorem 1.2.13 for Huber-contaminated regression already implies that we can learn MLRs where one of the components comprises more than half the data, i.e. has mixing weight exceeding 1/2. The reason is that in this case, we can view the samples coming from the remaining components as "corruptions."

In general if no component strictly comprises the majority however, the difficulty of the problem increases dramatically. In fact, without further assumptions on $\mathcal{D}_{\mathbf{x}}$, it becomes NP-hard: even mixtures of two linear regressions with equal mixing weights can encode SubsetSum [YCS14]. As in Section 1.2.1, a natural starting point would be to understand how hard this problem becomes after making the following assumption:

Assumption 2 (Learning MLRs under Gaussian inputs– Chapter 8)**.** *In Definition 1.2.22, we will take* $\mathcal{D}_{\mathbf{x}}$ *to be the standard Gaussian distribution in* d *dimensions.*

We note that all algorithmic results on learning mixtures of more than two linear regressions work with Assumption 2.[6] Indeed, another motivation for studying MLRs is that under

[6] The sole exception is the work of [LL18] which works in a more challenging setting where every component

Assumption 2, MLRs are a special case of mixtures of Gaussians. As we will see however, whereas there is mounting evidence that general mixtures of Gaussians are computationally hard to learn, MLRs represent a particularly natural sub-class where one can hope to circumvent these hardness results.

Finally, we must specify what the learning goal for MLRs is. In this work, we will focus on the following objective:

Definition 1.2.23 (Parameter Estimation for MLRs)**.** *Given independent samples from an unknown mixture of linear regressions with vectors $\{v_1, \ldots, v_k\}$, we say that $\mathcal{A}$ is an algorithm for* parameter estimation *achieving error ε if it outputs a list of vectors $\{\widehat{v}_1, \ldots, \widehat{v}_k\}$ for which there is some permutation π under which $\|v_i - \widehat{v}_{\pi(i)}\| \le \varepsilon$ for all i.*

Definition 1.2.23 justifies our definition of Δ-separation: if two of the underlying vectors v_i, v_j were infinitesimally close to each other, it would be impossible for an algorithm for parameter estimation to discern whether their components comprise two distinct components or a single one.

Model 3: Superpositions of Airy Disks The last mixture model we study will also be our first foray into applications of ideas from data science to the scienes. As we described in the first half of Section 1.1.3, a basic question in optics is to quantify the extent to which diffraction imposes limits on how well we can resolve point sources of light through a classical imaging system like a telescope.

The Airy disk pattern that a point source of light makes when light waves emanating from it pass through the (circular) aperture of a telescope and hit the imaging plane is easy enough to describe. The pattern is radially symmetric, and the intensity of the pattern at distance r from its center is proportional to

$$\left(\frac{J_1(r/\sigma)}{r/\sigma}\right)^2, \tag{1.10}$$

where J_1 is the Bessel function of the first kind of order 1 and σ is a "spread parameter" that is proportional to the wavelength of the light being imaged and inversely proportional to the

of the mixture is distributed according to a different Gaussian.

radius of the aperture. The quantity inside the square in (1.10) comes from the fact that the amplitude of the electric field on the imaging plane is essentially given by the Fourier transform of the indicator function of the disk corresponding to the aperture.

In a sense that can be made rigorous with Feynman's path integral formalism, we can regard the intensity (1.10), suitably normalized, as specifying the infinitesimal probability that a photon is detected at a particular point r away from the center of the point source projected onto the imaging plane. As a result, we can interpret an image arising from several point sources as merely a collection of samples from the following mixture model:

Definition 1.2.24 (Superpositions of Airy disks)**.** *Fix a parameter $\sigma > 0$. Given vector $\mu \in \mathbb{R}^2$, the* Airy disk *centered at μ with spread parameter σ is the distribution over $\mathbb{R}^2$ with density function given by*

$$A_\sigma(z) \propto \left(\frac{J_1(\|z-\mu\|/\sigma)}{\|z-\mu\|/\sigma}\right)^2,$$

A superposition of Airy disks *is then a mixture of Airy disks defined according to Definition 1.2.17. Similar to Definition 1.2.22, we say that a superposition of Airy disks centered at $\mu_1, \ldots, \mu_k$ is* Δ-separated *if $\|\mu_i - \mu_j\| \geq \Delta$ for all $i \neq j$.*

At first blush, this just looks like a two-dimensional mixture of Gaussians but where Gaussians have been replaced by the pdf for an Airy disk. The first important difference is that whereas the method of moments is a key ingredient for algorithms for learning mixtures of Gaussians, superpositions of Airy disks lack even finite *second* moments, necessitating the use of fundamentally different algorithmic techniques. Another key difference is that because we are focusing on a low-dimensional problem, the kinds of phase transitions we are after will be quite different. We explain all of this in greater detail in Chapter 9.

In any case, under Definition 1.2.24 it is clear what it means to be able to resolve nearby point sources: that we are able to estimate the centers of the Airy disks in the mixture from samples. In other words, *resolution in optical systems is equivalent to parameter estimation*, defined analogously to Definition 1.2.23. While our work is not the first to take this viewpoint (see Section 9.7 for detailed comparison to prior work), we are the first to leverage modern tools for learning mixture models to give nontrivial upper and lower bounds for the location of the diffraction limit, which in the language of Definition 1.2.24 can be formulated as follows:

Definition 1.2.25 (Diffraction Limit)**.** *Without loss of generality, suppose the spread parameter $\sigma = 1$. The* diffraction limit *is the smallest Δ for which there exists an algorithm that for any k, ε runs in time* $\mathrm{poly}(k, 1/\varepsilon)$ *and, given* $\mathrm{poly}(k, 1/\varepsilon)$ *samples from a mixture of k Δ-separated Airy disks, estimates the centers of the mixture to within L_2 norm ε.*

Fourier Analysis and Method of Moments Our techniques for handling the three mixture models above will involve Fourier-analytic and moment-based approaches and combinations thereof. To convey how our uses of these tools differ from conventional approaches, we begin by briefly sketching two common algorithmic recipes.

The first is the paradigm of *Fourier approximation*, sometimes called the *low-degree algorithm*, which was introduced in the seminal work of [LMN93] that brought Fourier analysis of Boolean functions into the realm of learning theory. The low-degree algorithm gives a general framework for learning certain kinds of Boolean functions F from random examples (x, y), where x is a uniformly random bitstring from $\{\pm 1\}^n$ and $y = F(x)$. The starting point is to consider the Fourier expansion of F, given by $F(x) = \sum_S \widehat{F}[S] x_S$ where S ranges over subsets of $\{1, \ldots, n\}$ and $x_S \triangleq \prod_{i \in S} x_i$. For certain interesting function classes like bounded-depth circuits, F exhibits *Fourier decay* in the sense that the magnitude of $\widehat{F}[S]$ is rapidly decaying in the size of S, in which case F is well-approximated by the low-degree truncation $\sum_{S:|S| \le \tau} \widehat{F}[S] x_S$ for some $\tau < n$. This suggests the following simple algorithm for learning such F from samples: empirically estimate every low-degree Fourier coefficient $\widehat{F}[S] = \mathbb{E}[F(x) \cdot x_S]$ from samples, and output the resulting estimate for the low-degree truncation of F.

This approach naturally extends to learning stochastic functions like stochastic decision trees, for which [AM91] gave an $n^{O(\log(k/\varepsilon))}$-time algorithm to learn to error ε by showing that stochastic decision trees are well-approximated by their degree-$O(\log(k/\varepsilon))$ truncation. We emphasize that the ε dependence in the exponent is a common artifact of this approach, simply because unless the function to be learned is *exactly* a low-degree polynomial, the degree τ to which one truncates its Fourier expansion must depend on the level of approximation one hopes to achieve.

As we will show however, this kind of dependence can sometimes be avoided if one uses estimates for the Fourier coefficients in a more sophisticated way.

The second recipe is the *method of moments*, originally introduced in a prescient paper of Karl Pearson [Pea94] to study populations of crabs. In its modern incarnation the technique goes as follows. Given a family of distributions $\mathcal{D}_\theta$ over $\mathbb{R}^d$, each indexed by some parameter $\theta \in \Theta$, and given independent samples from some $\mathcal{D}_\theta$, one can form estimates of the low-degree moments $M^\theta_\alpha \triangleq \mathbb{E}_{\mathcal{D}}[x_1^{\alpha_1} \cdots x_d^{\alpha_d}]$ for any tuple $\alpha = (\alpha_1, \ldots, \alpha_d)$. Now if one is able to argue that there is some degree $\tau \in \mathbb{N}$ such that for any distinct $\theta_1, \theta_2 \in \Theta$, there exists a moment α of total degree at most τ for which $M^{\theta_1}_\alpha$ and $M^{\theta_2}_\alpha$ differ noticably, then at least information-theoretically, this gives a way to recover θ from samples of $\mathcal{D}_\theta$.

Computationally efficient versions of this general recipe have led to significant progress in recent years on a number of basic statistical tasks, one notable example being learning mixtures of Gaussians [MV10, HL18, KSS18, DKS18b, LM21b, Kan20, BDJ+20]. The method of moments is so pervasive that in some contexts, the existence of a collection of instances of a learning problem whose low-degree moments all match is considered strong evidence for the computational hardness of that problem [Hop18, KWB19, DKS17].

As we will see below though, there are intriguing settings where moment-matching need not give rise to computational hardness, and in fact one way to get algorithms without just exploiting moments is to access the (continuous) Fourier transform of the underlying distribution!

Result 1: Beating Fourier Approximation with Method of Moments At a high level, our first result on learning mixture models gives a way of circumventing the shortcomings of Fourier approximation using a delicate implementation of the method of moments, instantiated in the setting of mixtures of subcubes and product distributions.

We first describe some of the subtleties to tailoring the method of moments to this context. The first is the issue discussed in the leadup to Definition 1.2.21: it is generally impossible to recover the actual centers of a mixture of product distributions because many different collections of product distributions may not only give rise to the same low-order moments, but even realize the same distribution. This is in stark contrast with other applications of the method of moments to problems like learning mixtures of Gaussians, where it is at least information-theoretically possible to uniquely identify the underlying parameters. A second, more technical subtlety is that because the distribution is over the Boolean hypercube, the only kinds of moments we can exploit are *multilinear* ones, i.e. statistics of the form $\mathbb{E}[x_S]$.

We overcome these issues by showing that low-degree multilinear moments of the underlying distribution $\mathcal{D}$ can be used to *identify the distribution*, even if identifying the parameters is possible. For instance, we show that any two mixtures of subcubes that are far in total variation distance must differ noticably on a moment of degree $O(\log k)$. We also give ways of using multilinear moments to back out properties that *any* collection of product distributions realizing $\mathcal{D}$ would satisfy, e.g. the minimum number of product distributions that could realize $\mathcal{D}$. Ultimately, we obtain the following algorithmic guarantee:

Theorem 1.2.26 (Informal, see Theorem 7.1.1)**.** *Given $\varepsilon > 0$, there is an algorithm that, given independent samples from an unknown mixture of k subcubes $\mathcal{D}$ over $\{0,1\}^n$, runs in time $O_k(n^{O(\log k)}\text{poly}(1/\varepsilon))$ and outputs a distribution $\mathcal{D}'$ for which $d_{TV}(\mathcal{D},\mathcal{D}') \le \varepsilon$ with high probability.*

Note that our $n^{O(\log k)}$ dependence is unavoidable: an $n^{o(\log k)}$-time algorithm for learning mixtures of subcubes would imply an $n^{o(\log k)}$-time algorithm for PAC learning k-leaf (deterministic) decision trees, an $n^{o(s)}$-time algorithm for learning s-sparse parities from noisy examples, and an $n^{o(s)}$-time algorithm for PAC learning s-juntas, all of which are widely conjectured to be impossible.

As a consequence of the connection between stochastic decision trees and mixtures of subcubes, we immediately obtain the following application:

Corollary 1.2.27 (Informal, see Theorem 7.1.3)**.** *Given $\varepsilon > 0$, there is an algorithm that, given random length-n bitstrings labeled according to an unknown decision tree T with k leaves, runs in time $O_k(n^{O(\log k)}\text{poly}(1/\varepsilon))$ and outputs a classifier whose error is within ε of the Bayes optimal classifier.*

It is instructive to contrast this with the $n^{O(\log(k/\varepsilon))}$-time algorithm of [AM91] for learning stochastic decision trees using the low-degree algorithm. Note that because the Fourier basis functions over the hypercube are given by monomials, one can regard the Fourier coefficients

$\widehat{F}[S]$ of a stochastic decision tree as low-degree moments of the joint distribution over (x, y) pairs. As a result, the Fourier-analytic approach of the low-degree algorithm is nothing more than a particular instantiation of the method of moments. Recall however that the main drawback of the low-degree algorithm is that one needs to look at Fourier coefficients of higher and higher degree if one wants to obtain better and better error guarantees. The upshot of our results is that in some cases, like for stochastic decision trees and mixtures of subcubes, there are more clever ways of exploiting just the degree-$O(\log k)$ moments to get arbitrarily small error!

Our techniques can also be extended to more general mixtures of product distributions:

Theorem 1.2.28 (Informal, see Theorem 7.1.6)**.** *Given $\varepsilon > 0$, there is an algorithm that, given independent samples from an unknown mixture of k product distributions $\mathcal{D}$ over $\{0,1\}^n$, runs in time $O_k((n/\varepsilon)^{O(k^2)})$ and outputs a distribution $\mathcal{D}'$ for which $d_{TV}(\mathcal{D}, \mathcal{D}') \leq \varepsilon$ with high probability.*

This is the first algorithmic improvement since the work of [FOS05], which obtained a runtime of $O((n/\varepsilon)^{O(k^3)})$. That said, the theorem above does appear to be significantly weaker than our result for mixtures of subcubes, which only incurred a dependence of $O(\log k)$ in the exponent. As we show, however, exponential dependence on $k^{\Theta(1)}$ is computationally necessary for general mixtures of product distributions:

Theorem 1.2.29 (Informal, see Theorem 7.4.1)**.** *Any algorithm given $\Omega(n^{-\sqrt{k}/3})$-accurate statistical query access (see Definition 7.4.2) to a mixture $\mathcal{D}$ of k product distributions over $\{0,1\}^n$ that outputs a distribution $\mathcal{D}'$ satisfying $d_{TV}(\mathcal{D}, \mathcal{D}') \leq \varepsilon$ for $\varepsilon \leq k^{-c\sqrt{k}}$ must make at least $n^{c'\sqrt{k}}$ queries.*

All known algorithms for learning mixtures of product distributions, and the overwhelming majority of provable algorithms in learning theory, can be implemented as statistical query algorithms, and informally, the above theorem suggests that a runtime of $n^{\mathrm{poly}(k)}$ is unavoidable for learning mixtures of k product distributions.

Result 2: Blending Method of Moments with Fourier Analysis Our second result on learning mixture models is in some sense a rejoinder to the first: for learning problems over *continuous* domains, the Fourier transform of the underlying distribution can sometimes give us access to information that straightforward applications of the method of moments cannnot. We explore this viewpoint in the context of learning MLRs.

Prior to this work, the best-known algorithms for learning mixtures of k linear regressions under Assumption 2 ran in time roughly $O(k^{O(k)}\mathrm{poly}(d))$ [LL18, ZJD16], and it was conjectured [LL18] that perhaps this kind of dependence on k was necessary given that there exist computational hardness results for learning general mixtures of Gaussians that scale exponentially in k [DKS17, BRST21].

Another reason to believe this is that one can in fact show that degree-k moments are not sufficient to distinguish between different mixtures of linear regressions. In particular, one can show the following:

Lemma 1.2.30 (Informal, see Appendix 8.10)**.** *For any $k \in \mathbb{N}$, there exist infinitely many pairs of mixtures of k linear regressions whose parameters differ noticably but whose moments up to degree k match.*

This is consistent with previous approaches for learning MLRs. For instance, in the $\sigma = 0$ case, [LL18] implemented a guided random walk to learn components of the mixture one at a time, and as we will elaborate upon in Section 8.3.1, a key subroutine in their algorithm was, given a candidate vector $v \in \mathbb{R}^d$, to estimate the distance between v and the closest vector v_i in the mixture. To get this subroutine, they observed that for (x, y) sampled from the mixture, the distribution over the residual $y - \langle v, x\rangle$ is distributed as a mixture of mean-zero Gaussians with variances $\{\|v - v_1\|^2, \ldots, \|v - v_k\|^2\}$. In particular the squared distance between v and the closest vector v_i is simply the minimum variance of any component in this mixture of 1D Gaussians. They then used the algorithm of [MV10] for learning mixtures of 1D Gaussians to obtain an estimate of $\min_i \|v - v_i\|$.

The $\exp(\Omega(k))$ dependence in the runtime of [LL18] comes from the fact that [MV10] is a classic application of method of moments: [MV10] prove that no two distinct mixtures of k one-dimensional Gaussians can match on more than $\Omega(k)$ moments, so then one can simply search over an $\exp(\Omega(k))$-sized grid of possible parameters for the mixture and output the one in the grid whose moments are closest to those of the underlying mixture. In light of Lemma 1.2.30, an approach like that of [LL18] is doomed to incur $\exp(\Omega(k))$ dependence for learning MLR's.

To circumvent this issue, we note that there is a way to estimate the minimum variance of any component in a mixture of 1D Gaussians without invoking the method of moments: because the *Fourier transform* of the mixture is a mixture of 1D Gaussians whose variances are reciprocals of the original variances, it suffices to estimate the *maximum* variance of the Fourier transform, and the latter can be done by computing a high enough moment of the Fourier transform. This insight is the workhorse behind our main results on MLRs, which give the first sub-exponential time algorithms for this problem. Here we informally state one result representative of the guarantees we show in Chapter 8:

Theorem 1.2.31 (Informal, see Theorem 8.6.2)**.** *Suppose $\sigma = 0$ in Definition 1.2.22. Given an unknown mixture $\mathcal{D}$ of k Δ-separated linear regressions in $\mathbb{R}^d$, there is an algorithm that takes $N = \widetilde{O}(d \log(1/\varepsilon)) \cdot \exp(\widetilde{O}(\sqrt{k}))$ samples from $\mathcal{D}$, runs in time $\widetilde{O}(N \cdot d)$, and estimates the parameters of $\mathcal{D}$ to error ε.*

Our techniques also extend to give a sub-exponential time algorithm for larger σ (see Theorem 8.7.1) as well as for a related mixture model relevant to *subspace clustering* (see Theorem 8.8.1).

We note that following the publication of this result, a follow-up work [DK20] gave a different algorithm that achieved a runtime scaling only *quasipolynomially* in k. Interestingly, their algorithm is based purely on method of moments, but the reason it does not contradict Lemma 1.2.30 is that they argue that there cannot be *too many* different mixtures which all simultaneously match each other on low-degree moments. This allows them to implement a sophisticated covering construction over which they can brute-force search for the parameters of the mixture.

Result 3: Fourier Analysis for the Diffraction Limit Finally, we describe how Fourier-analytic approaches to learning over continuous domains can be applied to pin down the diffraction limit in classical optical systems.

Recall from the discussion proceeding Definition 1.2.24 that one cannot hope to use the method of moments to learn superpositions of Airy disks because even the second moments of such a distribution are unbounded. Given that the density function (1.10) of an Airy disk is based on the Fourier transform of the indicator function of a disk, it is natural to try to exploit the Fourier transform of the distribution instead. In particular, one can check that the Fourier transform of the density of a superposition of Airy disks $\mathcal{D}$ with spread parameter σ and centers $\mu_1, \ldots, \mu_k$ at frequency $\omega \in \mathbb{R}^2$ is given by the pointwise product of

$$\sum_{j=1}^{k} \lambda_i e^{-2\pi i \langle \mu_j, \omega \rangle} \tag{1.11}$$

with the (two-dimensional) Fourier transform of (1.10). The latter is a known function $\widehat{A}_\sigma$ supported on a disk of radius $\frac{1}{\pi\sigma}$, so for any frequency ω in its support, we can estimate (1.11) by estimating $\mathbb{E}_{x\sim\mathcal{D}}[e^{-2\pi i\langle \omega, x\rangle}]$ from samples and dividing by the value of $\widehat{A}_\sigma$ at ω.

Thus, we can reduce the question of learning superpositions of Airy disks from samples to the question of estimating the locations of $\mu_1, \ldots, \mu_k$ given noisy, band-limited access to the function (1.11). This problem is known as *super-resolution* [Don92], and from a technical perspective, our primary contributions can be interpreted as giving new upper and lower bounds for this problem in the two-dimensional case.

Our first result is to show that if k is bounded by a constant, then there is no notion of a diffraction limit and one can in fact resolve superpositions of Airy disks in polynomial time/samples regardless of their level of separation:

Theorem 1.2.32 (Informal, see Theorem 9.4.1)**.** *If $\mathcal{D}$ is an unknown Δ-separated superposition of k Airy disks, then there is an algorithm that draws $N = \mathrm{poly}\left((k\sigma/\Delta)^{k^2}, 1/\varepsilon\right)$ independent samples from $\mathcal{D}$, runs in time $O(N)$, and estimates the centers of $\mathcal{D}$ to error ε with high probability.*

Our algorithm is based on projecting the data onto different lines, estimating the projections of the centers along these lines using the matrix pencil method [Moi15], and then piecing these estimates together by solving an appropriate linear system. While this technique is fairly standard in the mixture model learning literature [KMV10, MV10], the result demonstrates that even off-the-shelf tools in learning theory can have useful implications in other domains, in this case clarifying why in some domains like astronomy where there are only ever a few tightly spaced point sources, there is evidently no diffraction limit.

We now turn to our main results on the diffraction limit. For unbounded k, we first show that above a certain level of separation, there is a learning algorithm whose runtime and sample complexity scale polynomially in k:

Theorem 1.2.33 (Informal, see Theorem 9.4.2)**.** *Define the absolute constant* $\overline{\gamma} = \frac{2j_{0,1}}{\pi} = 1.530\ldots$, *where* $j_{0,1}$ *is the first positive zero of the Bessel function* J_0*. If* $\mathcal{D}$ *is an unknown* Δ*-separated superposition of* k *Airy disks, where* $\Delta \geq \overline{\gamma}\pi\sigma$*, then there is an algorithm with time and sample complexity* $\mathrm{poly}\,(k, 1/\Delta, 1/\varepsilon)$ *that estimates the centers of* $\mathcal{D}$ *to error* ε *with high probability.*

The algorithm is based on a tensor decomposition approach introduced by [HK15]. Whereas the analysis in that work was tailored to high-dimensional settings, we refine their analysis to handle the two-dimensional case by using certain extremal functions [Gon18, HV$^+$96, CCLM17] arising in the study of de Branges spaces. To our knowledge, this is the first use of such functions in a learning theory setting. We defer the details to Section 9.4.3.

As our main result in this part of the thesis, we give a surprising lower bound:

Theorem 1.2.34 (Informal, see Theorem 9.5.1)**.** *Let* $\underline{\gamma} \triangleq \sqrt{4/3} \approx 1.155$*. For* $\Delta < \underline{\gamma}\pi\sigma$*, any algorithm for learning general* Δ*-separated superpositions of Airy disks requires sample complexity exponential in* $\sqrt{k}$ *in the worst case.*

We emphasize that the most striking aspect of this result is that it contradicts the conventional wisdom that the diffraction limit for classical optical systems occurs at the *Abbe limit*, which in our language is given by $\Delta = \pi\sigma$. We prove our lower bound by exhibiting a pair of superpositions whose parameters are noticably different but which are close in total variation distance. At the heart of this construction as well as the upper bound above is a certain fundamental question about cancellations of exponential sums which we believe to be of independent interest (see Questions 6 and 7).

1.2.4 Quantum State Certification and the Chain Rule

The last set of results that this thesis will cover is a bit further afield from the preceding results from a technical standpoint. From a conceptual standpoint however, we regard them as very much in the spirit not only of asking what ideas from learning theory can say about problems in the sciences, but also of understanding the algorithmic landscape for learning from data that arrives in a dynamic fashion.

The latter point requires a bit of unpacking, so we begin by formulating the algorithmic questions we consider.

Models and Existing Results We will be interested in quantum learning, specifically *learning from quantum data*. We consider a setting where the learner gets access to copies of an unknown quantum state ρ (see Section 1.3.9 for an overview of quantum basics) and would like to learn something about ρ by measuring these copies. For instance, the most basic question one could ask is to learn a *single bit* of information about the underlying state:

Definition 1.2.35 (Quantum distinguishing task)**.** *Let $\mathcal{C}_0$ and $\mathcal{C}_1$ be two known sets of quantum states, and let ρ be an unknown state that is promised to be in either $\mathcal{C}_0$ or $\mathcal{C}_1$. Given a collection of copies of ρ and the ability to measure them, the goal of the learner is to distinguish with high probability between whether $\rho \in \mathcal{C}_0$ or $\rho \in \mathcal{C}_1$. We refer to such a task as a* distinguishing task*. The minimum number of copies of ρ needed by any given algorithm for this task is called the* copy complexity *of the task.*

This is the natural quantum analogue of *distribution testing* (see e.g. the survey of [Can20]). In this thesis, we will focus on the following distinguishing task, which can be thought of as the quantum version of the classic question of testing *goodness-of-fit*: given a known distribution p and samples from an unknown distribution q, determine whether $p = q$ or $d_{\mathrm{TV}}(p, q) > \varepsilon$.

Definition 1.2.36 (Quantum state certification)**.** *Fix error parameter $\varepsilon > 0$ and let σ be a known quantum state. Given a collection of copies of an unknown state ρ and the ability to measure them, the task of* quantum state certification *is to distinguish with high probability between whether $\sigma = \rho$ or $\|\rho - \sigma\|_{\mathsf{tr}} > \varepsilon$. In the notation of Definition 1.2.35, this task is given by $\mathcal{C}_0 = \{\sigma\}$ and $\mathcal{C}_1 = \{\rho : \|\rho - \sigma\|_{\mathsf{tr}} > \varepsilon\}$.*

Just as goodness-of-fit is a foundational question in statistical hypothesis testing that has been studied ever since Pearson developed his eponymous chi-squared test [Pea00], quantum state certification addresses a basic need to verify that the states prepared in a laboratory setup are what we intended them to be.

A particularly illuminating example of Definition 1.2.36 is the following quantum analogue of the well-studied question of *uniformity testing* (given samples from an unknown distribution, determine whether it is the uniform distribution or far from being the uniform distribution):

Definition 1.2.37 (Mixedness testing)**.** Mixedness testing *is the special case of state certification where, in the notation of Definition 1.2.36, σ is the maximally mixed state $\frac{1}{d}$ Id.*

Thus far, we have been intentionally vague about the meaning of an algorithm that makes measurements on a collection of copies of a quantum state. The reason is that there are a number of different ways of formalizing the precise model for this, and this is the key technical distinction between proving copy complexity bounds for quantum distinguishing tasks and proving sample complexity bounds for classical distribution testing tasks.

Definition 1.2.38 (Quantum measurements)**.** *Suppose the learner gets access to N copies of an unknown state $\rho \in \mathbb{C}^{d\times d}$. The following are three different ways in which the learner could interact with these copies:*

1. Fully entangled measurement*: the learner applies a single POVM over $(\mathbb{C}^d)^{\otimes N}$ (see Definition 1.3.47) to the tensor product $\rho^{\otimes N}$ and decides based on the outcome of this measurement.*

2. Adaptive unentangled/incoherent measurements*: the learner applies a POVM to the first copy of ρ, observes the outcome, chooses a POVM based on this outcome and applies it to next copy of ρ, etc. Afterwards, she decides based on the outcome of all N measurements.*

3. Nonadaptive unentangled/incoherent measurements*: the learner selects N POVMs in advance, applies the i-th POVM to the i-th copy of ρ for $i = 1, \ldots, N$. Afterwards, she decides based on the outcome of all N measurements.*

Note that the list in Definition 1.2.38 is in decreasing order of generality, that is, a fully entangled measurement can implement any strategy based on adaptive unentangled measurements, and obviously adaptive strategies are at least as powerful as nonadaptive ones. [OW15] studied algorithms for quantum distinguishing tasks where the learner can make a fully entangled measurement and proved that the copy complexity of mixedness testing is $\Theta(d/\varepsilon^2)$ in this setting. [BOW19] later extended this result by giving an algorithm for general state certification which makes fully entangled measurements and also achieves copy complexity $O(d/\varepsilon^2)$. In direct analogy with classical distribution testing bounds, this is significantly less than the copy complexity for *learning* ρ using a fully entangled measurement, which is known to be $\Theta(d^2/\varepsilon^2)$ [HHJ+17, OW16].

Unfortunately, fully entangled measurements are well outside the scope of what is practically feasible, as they require the learner to maintain a quantum memory that is exponential in the number of copies of ρ. It is therefore natural to ask whether one can hope to match these bounds with an algorithm that simply makes adaptive unentangled measurements– in the context of mixedness testing, this was explicitly asked by Wright in [Wri16].

It turns out that one can already achieve a nontrivial bound for mixedness testing with a very simple algorithm that uses *nonadaptive* unentangled measurements: pick a random basis, repeatedly measure in this basis, and test whether the resulting outcomes are samples from the uniform distribution over d elements. We analyze this in Section 11.6.1 and show how to extend it to state certification in Section 11.6.3. For mixedness testing, the copy complexity of this algorithm turns out to be $O(d^{3/2}/\varepsilon^2)$. One can then ask: can we do better, or is this algorithm optimal for unentangled measurements?

Taming Adaptivity via Entropy As we will show, one cannot hope to match the performance of fully entangled measurements using adaptive, unentangled measurements. The challenge with showing such a lower bound however is that, unlike distribution testing where we simply interact with the underlying distribution through random samples, for quantum tasks there is the extra dimension of complexity coming from adaptivity.

To handle this, we take inspiration from lower bound techniques in the *adversarial bandits* literature. While the following discussion about bandit lower bounds is not needed to understand our proof technique, we believe it is instructive to highlight the parallels between the adaptivity inherent in quantum measurements and the adaptivity inherent in strategies for bandit problems. We begin with a definition for the latter:

Definition 1.2.39 (Adversarial multi-armed bandits)**.** *Fix parameter $K \in \mathbb{N}$, corresponding* *to the number of* arms *that the player can choose from, and a* time horizon $T \in \mathbb{N}$. *The player interacts with an adversary over T rounds of the following game. In each round t:*

1. *The player chooses an arm $I_t \in \{1, \ldots, k\}$.*

2. *The adversary chooses a vector of rewards $(r_{1,t}, \ldots, r_{K,t})$.*

3. *The player observes only the reward $r_{I_t,t}$.*

We note that the player's and the adversary's moves can be randomized and adaptive, that is, the player can choose the arm in round t based on her previous actions and the rewards she has observed in the previous rounds, and the adversary can choose the vector of rewards in round t based on everything that has happened up to that point.

The goal of the player is to compete with the best fixed action in hindsight, that is, to minimize the regret

$$\max_{1 \le j \le K} \sum_t x_{j,t} - \sum_t x_{I_t,t},$$

either in expectation or with high probability.

The classic paper of [ACBFS02] showed how to achieve $\widetilde{O}(\sqrt{KT})$ expected regret and also gave a matching lower bound (up to log factors):

Theorem 1.2.40 ([ACBFS02], Theorem 5.1)**.** *For any $K \ge 2$ and any time horizon T, there exists a (non-adaptive, randomized) strategy for the adversary such that any player strategy incurs $\Omega(\sqrt{KT} \wedge T)$ expected regret.*

We are primarily interested in the analogy with their lower bound. The basic proof idea for this result is to argue that over a bounded time horizon, the player cannot distinguish between the following two scenarios:

1. **Null hypothesis**: the rewards for all the arms are distributed as unbiased Bernoullis in every round

2. **Mixture of alternatives**: at the outset, an index i is sampled at random from $\{1, \ldots, K\}$, and then subsequently in every round, the rewards for arms not equal to i are distributed as unbiased Bernoullis, and the reward for arm i is Bernoulli with a bias of $\sqrt{K/T}$

To make this rigorous, one must argue that regardless of the player strategy, the total variation distance between the distribution over the transcript of rewards observed by the player under the null hypothesis and the distribution under the mixture of alternatives is small.

Given a player strategy, let $\mathcal{D}_0^{\le t}$ denote the distribution over the first t rewards under the null hypothesis, and let $\mathcal{D}_1^{\le t}$ denote the same under the mixture of alternatives. If $\mathcal{D}_{1,i}^{\le t}$ denotes the distribution $\mathcal{D}_1^{\le t}$ conditioned on arm i having the bias, then we ultimately want to show that

$$d_{\mathrm{TV}}\left(\mathcal{D}_0^{\le T}, \mathop{\mathbb{E}}_i\left[\mathcal{D}_{1,i}^{\le T}\right]\right) = o(1). \tag{1.12}$$

We note that this general setup of bounding the total variation between a null hypothesis and a mixture of alternatives closely follows the usual framework for showing sample complexity lower bounds for distribution testing tasks. The challenge unique to the bandits setup however is that because the player is adaptive, $\mathcal{D}_{1,i}^{\le T}$ is not a product distribution, precluding many of the techniques commonly used in the testing literature, see e.g. [IS12]. The key insight in [ACBFS02] is that one can nevertheless control the left-hand side of (1.12) by passing from total variation to *KL divergence* and then applying the chain rule! We refer the reader to [ACBFS02, BCB12] for the details of this calculation.

Our Results Inspired by the entropic approach of [ACBFS02], which has also found use in other non-bandit contexts like proving phase transitions related to Gaussian matrix ensembles [BG18], we exploit similar ideas to understand the copy complexity of quantum distinguishing tasks under adaptive, unentangled measurements.

Our lower bound construction follows a similar recipe: argue that under any adaptive measurement strategy, the distribution $\mathcal{D}_0$ over measurement outcomes under a null hypothesis is indistinguishable from the distribution $\mathcal{D}_1$ under a mixture of alternatives. For mixedness testing, the null hypothesis is simply that the underlying state is maximally mixed, while the mixture of alternatives is a certain ensemble of states which are ε-far in trace distance from the maximally mixed state. Specifically, we consider the mixture consisting of states

$$\rho_{\mathbf{U}} \triangleq \frac{1}{d}\left(\mathrm{Id} + \varepsilon \cdot \mathbf{U}\mathbf{Z}\mathbf{U}^\dagger\right),$$

where $\mathbf{Z} \triangleq \mathrm{diag}(1, 1, ..., -1, -1, ...)$ and $\mathbf{U}$ is a Haar-random unitary matrix. For readers familiar with Paninski's proof of the $\Omega(\sqrt{d}/\varepsilon^2)$ sample complexity lower bound for uniformity testing [Pan08], this is the natural quantum analogue of his construction and has appeared previously, e.g. in the lower bound of [OW15] for mixedness testing with a fully entangled measurement.

Letting $\mathcal{D}_{1,\mathbf{U}}$ denote the distribution over measurement outcomes under a particular unentangled measurement strategy when the underlying state is $\rho_{\mathbf{U}}$, we would like to show that $d_{\mathrm{TV}}\left(\mathcal{D}_0^{\leq T}, \mathbb{E}_{\mathbf{U}}\left[\mathcal{D}_{1,\mathbf{U}}^{\leq T}\right]\right) = o(1)$, in direct analogy with (1.12). When the strategy is nonadaptive, we can leverage existing techniques in the distribution testing literature, specifically the so-called *Ingster-Suslina method* [IS12], to show a tight lower bound:

Theorem 1.2.41 (Informal, see Theorem 10.1.1)**.** *Fix error parameter $0 < \varepsilon < 1/2$. The copy complexity of mixedness testing to error ε using unentangled, nonadaptive measurements is $\Theta(d^{3/2}/\varepsilon^2)$ copies.*

By using the chain rule for KL divergence, we are able to show the following slightly weaker lower bound for adaptive measurements:

Theorem 1.2.42 (Informal, see Theorem 10.1.2)**.** *Fix error parameter $\varepsilon > 0$. Any algorithm for mixedness testing to error ε that only uses unentangled, adaptive measurements must use at least $\Omega(d^{4/3}/\varepsilon^2)$ copies.*

Contrasting this with the $O(d/\varepsilon^2)$ upper bound using fully entangled measurements [OW15], we see that this gives the first known separation in quantum learning between algorithms that can make fully entangled measurements and algorithms that make general unentangled, adaptive measurements. We defer the technical details to Chapter 10, noting that although the general trick of using the chain rule is inspired by the aforementioned bandit lower bounds, we need to leverage a number of tools specific to the quantum setting, like concentration of measure for Haar-random unitaries, to get our results.

It turns out that the tools we develop are quite flexible and allow us to tackle the more general problem of quantum state certification. While the theorems above immediately imply lower bounds for quantum state certification for *worst-case* choices of reference state σ, it is conceivable that one could get better copy complexity bounds for σ that are more structured than the maximally mixed state. For instance, if σ were maximally mixed over a known $O(1)$-dimensional subspace, it is straightforward to modify the nonadaptive strategy based on measuring in a random basis to get a copy complexity upper bound of $O(1/\varepsilon^2)$.

This motivates the question of obtaining *instance-optimal* copy complexity bounds: is there some simple functional $f(\sigma)$ for which quantum state certification relative to σ has copy complexity $\Theta(f(\sigma)/\varepsilon^2)$? After all, from the perspective of a researcher trying to test whether a state prepared in the lab satisfies certain properties, σ is given to us, and we would like to use as few samples as possible by exploiting our knowledge of σ.

This kind of question has previously been studied in the context of classical distribution estimation and testing [ADJ+11, ADJ+12, VV17, VV16]. In the final part of this thesis, by designing more sophisticated mixtures of alternatives than for mixedness testing, we give the first instance-optimal (up to log factors) bounds on the copy complexity of state certification with unentangled, nonadaptive measurements:

Theorem 1.2.43 (Informal, see Theorems 11.5.1 and 11.6.1)**.** *Given a known quantum state σ, the copy complexity of state certification with unentangled, nonadaptive measurements is up to log factors given by*

$$\widetilde{\Theta}\left(\frac{d \cdot d_{\mathsf{eff}}^{1/2}}{\varepsilon^2} \cdot F_\sigma\right),$$

where d_{eff} is the "effective dimension" of ρ and F_σ is essentially the fidelity between σ and the maximally mixed state (see Theorem 11.5.1 for formal definitions).

By taking $\sigma = \frac{1}{d}\,\mathrm{Id}$ above, we recover Theorem 1.2.41. We can also prove a lower bound in the adaptive setting which simply replaces $d_{\mathsf{eff}}^{1/2}$ with $d_{\mathsf{eff}}^{1/3}$ (see Theorem 11.7.1), recovering Theorem 1.2.42. That said, the latter is not "instance-optimal" as we are not yet able to prove a tight lower bound even for mixedness testing.

We note that Theorem 1.2.43 is qualitatively quite different from its analogue in the classical setting, proven in [VV16]. In that work, it was shown that the sample complexity for testing whether a distribution is equal to a known distribution p or ε-far from it in total variation distance is essentially given by $\Theta(\|p\|_{2/3}/\varepsilon^2)$ (see Theorem 11.3.6). In particular, the instance-optimal sample complexity is *dimension-free* in the sense that it scales as $\Theta(1/\varepsilon^2)$ for p whose 2/3-quasinorm is $O(1)$. In contrast, our bound– in addition to having a more natural interpretation in terms of quantum fidelity– shows that state certification cannot escape the curse of dimensionality. As we discuss in Example 11.1.2, even when $\sigma = \mathrm{diag}(1-1/d, 1/d^2, \ldots, 1/d^2)$, Theorem 1.2.43 tells us that the copy complexity for state certification relative to σ scales as $\sqrt{d}/\varepsilon^2$ for sufficiently small ε, even though σ looks for all intents and purposes like a rank-1 state! Very roughly, the idea is that in the quantum setting, there is more "room" for building mixtures of alternatives than simply perturbing the eigenvalues and conjugating by a Haar-random unitary.

1.3 Preliminaries

Here we record various technical ingredients that are needed throughout this thesis.

1.3.1 Miscellaneous Notation

- Given positive integer n, we use $[n]$ to denote the set $\{1, \ldots, n\}$.

- We will use $\vee$ and $\wedge$ to denote max and min respectively, though when convenient we will also use $\min(\cdot)$ and $\max(\cdot)$.

- Let $\mathbf{1}_n \in \mathbb{R}^n$ denote the all-ones vector. We omit the subscript when the context is clear. For a vector $u \in \mathbb{R}^d$ and index $\ell \in [d]$, u_ℓ denotes the ℓ-th entry of u. For indices $a, b \in [d]$, $u_{a:b} \in \mathbb{R}^{b-a+1}$ denotes the a-th through b-th entries of u.

- Given a metric space Ω equipped with a metric d, and given a function $f : \Omega \to \mathbb{R}$, we say that f is Λ-*Lipschitz with respect to* d if $|f(x) - f(y)| \le d(x, y)$. When Ω is Euclidean space, we will simply say that f is Λ-Lipschitz unless otherwise specified.

- Given a matrix $M \in \mathbb{C}^{d\times d}$, we use $M^\top$ to denote its transpose and $M^\dagger$ to denote its conjugate transpose. We denote by M_i^j the entry of M in row i and column j.

- We will occasionally use notation like $x = [c_1, c_2]\cdot y$ and $x = 1\pm\delta$ to mean $c_1 y \le x \le c_2 y$ and $1 - \delta \le x \le 1 + \delta$ respectively.

- Let $\mathcal{S}_\ell$ denote the symmetric group on ℓ elements.

- Given two strings s and t, let $s \circ t$ denote their concatenation. Given $t > 1$ and a sequence $x_1, ..., x_{t-1}$, define $x_{<t} \triangleq (x_1, ..., x_{t-1})$. We will also sometimes refer to this as $x_{\leq t-1}$. Also, let $x_{<1} \triangleq \emptyset$.

- In addition to big-O notation, we sometimes find it more convenient to use $f \lesssim g$ and $f \gtrsim g$ to denote $f = O(g)$ and $f = \Omega(g)$ respectively.

1.3.2 Linear Algebra Basics

Norms

Given vector $v \in \mathbb{R}^d$, let $\|v\|_p$ denote its L_p norm. When the context is clear, we let $\|\cdot\|$ denote the L_2 norm.

Given matrix $M \in \mathbb{R}^{m \times n}$, let $\|M\|_F$ denote its Frobenius norm, $\|M\|_2$ its operator norm, and $\|M\|_{\mathsf{tr}}$ its trace norm. Let $\|M\|_{\max}$ denote the maximum absolute value of any entry in M. When the context is clear, we let $\|\cdot\|$ denote the operator norm.

For $r > 0$, let $\mathcal{B}_r^d \subset \mathbb{R}^d$ denote the L_2 ball of radius r centered at the origin. When the context is clear, we will suppress the superscript d.

We will often also work in the following function space. Consider the Banach space of functions $f : \mathbb{R} \to \mathbb{R}$ which are p-th power integrable with respect to the standard Gaussian measure γ, that is for which

$$\|f\|_p \triangleq \int_{\mathbb{R}} |f(x)|^p d\gamma(x) < \infty.$$

We refer to this space as $L^p(\mathbb{R}, \gamma)$. When $p = 2$, this is a Hilbert space with the inner product

$$\langle f, g \rangle = \int_{\infty}^{\infty} f(z) g(z) d\gamma(z).$$

Using the standard d-dimensional Gaussian measure, we can analogously define $L^p(\mathbb{R}^d, \gamma)$ analogously for functions $\mathbb{R}^d \to \mathbb{R}$. When $p = 2$, we will use the shorthand $\|\cdot\|$ to refer to $\|\cdot\|_2$.

Orthogonal Projectors

Given vector spaces $U \subset V$, $V \backslash U = V \cap U^\perp$ denotes the orthogonal complement of U in V. Let $\mathbb{S}_V \subset \mathbb{R}^d$ denote the set of vectors in V of unit L_2 norm.

Given a vector space $U \subset \mathbb{R}^d$, let Π_U denote the orthogonal projection operator onto U, and when the ambient space is clear from context, let $U^\perp$ denote the orthogonal complement of U.

We will overload notation for this in various places. For instance, we will often use U to refer to a set of column vectors $\{u_1, \ldots, u_\ell\}$, in which case $\mathrm{span}(U)$ denotes the span of these vectors, Π_U denotes $\Pi_{\mathrm{span}(U)}$, and $\Pi_{U^\perp}$ denotes $\Pi_{\mathrm{span}(U)^\perp}$. Given $v \in \mathbb{S}^{d-1}$, we will use $\Pi_v \triangleq vv^\top$ and $\Pi_v^\perp \triangleq \mathrm{Id} - vv^{top}$ to denote projection to the span of v and its orthogonal complement, respectively. More generally, given $V \in \mathbb{R}^{n \times r}$ whose columns are orthonormal, we will use $\Pi_V \triangleq VV^\top$ and $\Pi_V^\perp \triangleq \mathrm{Id} - VV^\top$ to denote projection to the span of the columns of V and its orthogonal complement, respectively.

Frames

Let St^d_ℓ denote the *Stiefel manifold* of $n \times \ell$ matrices with orthonormal columns, and let $\mathrm{G}(d, \ell)$ denote the *Grassmannian* of ℓ-dimension subspaces of $\mathbb{R}^d$. $\mathrm{G}(d, \ell)$ can be regarded as the quotient of St^d_ℓ under the natural action of the orthogonal group $O(\ell)$, that is, given any subspace $U \in \mathrm{G}(d, \ell)$ and any $V \in \mathrm{St}^d_\ell$ whose columns form a basis for U, we can associate U to the equivalence class $[V] \triangleq \{V \cdot O : O \in O(\ell)\}$.

We will often refer to elements of the Stiefel manifold as *frames*:

Definition 1.3.1 (Frames)**.** *A set of orthonormal vectors* $\widetilde{w}_1, ..., \widetilde{w}_\ell$ *is a* frame*. Given subspace* $V \subset \mathbb{R}^d$*, we say that this frame is* ν-nearly within V *if* $\|\Pi_V \widetilde{w}_i\| \geq 1 - \nu$ *for all* i*. We will sometimes refer to their span* $\widetilde{W}$ *as a frame* ν*-nearly within to* V*, when the choice of orthonormal basis for* $\widetilde{W}$ *is clear from context.*

Subspace Distances

In various places we will need to quantify the distance between different subspaces, in particular in Chapters 2 and 3. There are a number of ways of doing this

Definition 1.3.2. *Given $V, V' \in St^n_r$, the* Procrustes distance $d_P(V, V')$ *is given by*

$$d_P(V, V') \triangleq \min_{O \in O(r)} \|V - V'O\|_F.$$

Let $0 \leq \theta_1 \leq \cdots \leq \theta_r \leq \pi/2$ be the principal angles between V and V'. Then we also have that

$$d_P(V, V') = 2 \left(\sum_{i=1}^{r} \sin^2(\theta_i/2) \right)^{1/2}.$$

Definition 1.3.3. *Given $V, V' \in St^n_r$, the* chordal distance $d_C(V, V')$ *is given by*

$$d_C(V, V') \triangleq (d - \|V^\top V'\|_F^2)^{1/2}$$

Let $0 \leq \theta_1 \leq \cdots \leq \theta_r \leq \pi/2$ be the principal angles between V and V'. Then we also have that

$$d_C(V, V') = \left(\sum_{i=1}^{r} \sin^2 \theta_i \right)^{1/2}.$$

We collect some basic facts about these distances.

Fact 1.3.4 (Triangle inequality for Procrustes)**.** *Given any $V_1, V_2, V_3 \in St^n_r$,*

$$d_P(V_1, V_2) + d_P(V_2, V_3) \geq d_P(V_1, V_3).$$

Lemma 1.3.5. *$d_P(V, V')^2/2 \leq d_C(V, V')^2 \leq d_P(V, V')^2$.*

Proof. This follows immediately from the elementary inequality $2\sin^2(\theta/2) \leq \sin^2(\theta) \leq 4\sin^2(\theta/2)$ for $\theta \in [0, \pi/2]$. □

Power Method and Perturbation Bounds

Our spectral algorithms require the following well-known tool:

Fact 1.3.6 (Power method, see [RST09])**.** *Let $\mathbf{M} \in \mathbb{R}^{d \times d}$, let $k \leq d$ be a non-negative integer, and let $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_d$ be the nonzero singular values of $\mathbf{M}$. For any $k = 1, \ldots, d-1$, let $\mathsf{gap}_k = \sigma_k/\sigma_{k+1}$. Suppose there is a* matrix-vector oracle *which runs in time R, and which, given $v \in \mathbb{R}^d$, outputs $\mathbf{M}v$. Then, for any $\eta, \delta > 0$, there is an algorithm*

ApproxBlockSVD$(\mathbf{M}, \eta, \delta)$ *which runs in time* $\widetilde{O}(kR \log \frac{1}{\eta \cdot \delta \cdot \mathsf{gap}_k})$*, and with probability at least* $1 - \delta$ *outputs a matrix* $\mathbf{U} \in R^{d \times k}$ *with orthonormal columns so that* $\|\mathbf{U} - \mathbf{U}_k\|_2 < \eta$*, where* $\mathbf{U}_k$ *is the matrix whose columns are the top* k *right singular vectors of* $\mathbf{M}$*.*

The following eigenvalue stability result will be important in our analysis of filtered PCA.

Lemma 1.3.7 (Gap-free Wedin theorem, see e.g. Lemma B.3 in [AZL16])**.** *Let* $\varepsilon, \gamma, \mu > 0$*. For psd matrices* $\mathbf{A}, \hat{\mathbf{A}} \in \mathbb{R}^{d \times d}$ *for which* $\|\mathbf{A} - \hat{\mathbf{A}}\|_2 \le \varepsilon$*, if* U *is the matrix whose columns consist of the eigenvectors of* $\mathbf{A}$ *with eigenvalue at least* μ*, and* $\hat{U}$ *is the matrix whose columns consist of the eigenvectors of* $\hat{\mathbf{A}}$ *with eigenvalue at most* $\mu - \gamma$*, then* $\|\mathbf{U}^\top \hat{\mathbf{U}}\|_2 \le \varepsilon / \gamma$*.*

In particular, we get the following straightforward consequence of Lemma 1.3.7:

Corollary 1.3.8. *Let* $\lambda \ge 2\varepsilon > 0$*. For symmetric matrices* $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{R}^{d \times d}$ *for which* $\|\mathbf{A} - \widehat{\mathbf{A}}\|_2 \le \varepsilon$ *and* $\|\widehat{\mathbf{A}}\|_2 \ge \lambda - \varepsilon$*, if* $w \in \mathbb{S}^{d-1}$ *is the top singular vector of* $\widehat{\mathbf{A}}$*, and* $V \subset \mathbb{R}^d$ *is the orthogonal complement of the kernel of* $\mathbf{A}$*, then* $\|\Pi_V w\|_2 \ge 1 - 4\varepsilon^2/\lambda^2$*.*

Proof. If we take $\xi = \mu = \|\widehat{\mathbf{A}}\|$ in Lemma 1.3.7, then the columns of $\mathbf{U}$ (resp. $\widehat{\mathbf{U}}$) in Lemma 1.3.7 consist of an orthonormal basis $B \in \mathbb{R}^{d \times k}$ for the kernel of $\mathbf{A}$ (resp. w and other singular vectors of $\mathbf{A}$, if any, with the same singular value), where k is the dimension of $\ker(\mathbf{A})$. We have that

$$\|\Pi_{V^\perp} w\| \le \|\widehat{\mathbf{U}}^\top \mathbf{U}\|_2 \le \varepsilon / \|\widehat{\mathbf{A}}\|_2 \le \frac{\varepsilon}{\lambda - \varepsilon},$$

from which we conclude that

$$\|\Pi_V w\| \ge \left(1 - \left(\frac{\varepsilon}{\lambda - \varepsilon}\right)^2\right)^{1/2} \ge 1 - 4\varepsilon^2/\lambda^2$$

as claimed. □

The following says that if a set of r orthogonal unit vectors all have large component in U^*, then their span is close to the true subspace in the sense of either of the distances above.

Lemma 1.3.9. *Let* Π *denote orthogonal projection to a subspace* $U_1 \in G(n, \ell)$*. Let* $v_1, ..., v_\ell \in \mathbb{S}^{n-1}$ *be orthogonal and satisfy* $\|\Pi v_i\|_2 \ge 1 - \varepsilon$ *for all* $i \in [r]$*. Let* $U_2 \triangleq \mathrm{span}(\{v_i\})$*. Then* $d_C(U_1, U_2) \le \sqrt{2\varepsilon \cdot \ell}$ *and* $d_P(U_1, U_2) \le 2\sqrt{\varepsilon \ell}$*.*

Proof. Let $V_1 \in \mathrm{St}^n_\ell$ be any frame with columns forming a basis for U_1, and let $V_2 \in \mathrm{St}^n_\ell$ be the frame with columns given by $\{v_i\}_{i\in[\ell]}$. Observe that

$$d_C(U_1,U_2)^2 = \ell - \|V_1^\top V_2\|_F^2 = \ell - \mathrm{Tr}\left(V_2^\top \Pi V_2\right) \geq \ell - \sum_{i=1}^{\ell} \|\Pi v_i\|_2^2 = 2\varepsilon \cdot \ell.$$

as claimed. □

We can use Lemma 1.3.9 to obtain the following:

Lemma 1.3.10. *Let $U^* \in G(n,r)$, $V \in St^n_r$, and $\varepsilon > 0$. Suppose the columns v_i of V satisfy $\|\Pi_U v_i\|_2 \geq 1-\varepsilon$ for every $i \in [\ell]$. Then there exist orthogonal vectors $v_1^*, ..., v_\ell^* \in U$ for which $\langle v_i, v_i^*\rangle \geq 1 - 2\varepsilon\ell$ for every $i \in [\ell]$.*

Proof. Let $U \triangleq \mathrm{span}(\{v_i\})$. By Lemma 1.3.9, $d_P(U,U^*) \leq 2\sqrt{\varepsilon\cdot\ell}$, so there exists a frame $V^* \in \mathrm{St}^n_r$ for U^* such that $\|V - V^*\|_F \leq 2\sqrt{\varepsilon\cdot\ell}$. Note that $\|V-V^*\|_F^2 = 2\ell - 2\,\mathrm{Tr}(V^\top V^*) = 2\sum_{i=1}^{\ell}(1 - \langle v_i, v_i^*\rangle)$. As v_i, v_i^* are unit vectors $1 - \langle v_i, v_i^*\rangle \geq 0$ for every $i \in [\ell]$, so we conclude that $\langle v_i, v_i^*\rangle \geq 1 - 2\varepsilon\ell$ for each $i \in [\ell]$. □

The following claim says that swapping out orthogonal projectors to a subspace with orthogonal projectors to a nearby subspace incurs small error.

Claim 1.3.11. *For any $M \in \mathbb{R}^{n\times n}$ and projectors $\Pi_1, \Pi_2 \in \mathbb{R}^{n\times n}$ to subspaces $U_1, U_2 \in G(n,\ell)$, $\|\Pi_1^\top M \Pi_1 - \Pi_2^\top M \Pi_2\|_2 \leq \sqrt{2}\cdot\|M\|_2 \cdot d_C(U_1,U_2))$.*

Proof. We bound $\|(\Pi_1 - \Pi_2)^\top M \Pi_1\|_2$ and $\|\Pi_2^\top M(\Pi_1 - \Pi_2)\|_2$ and apply triangle inequality.

By sub-multiplicativity of the operator norm and the fact that projections have spectral norm 1, $\|(\Pi_1 - \Pi_2)^\top M\Pi_1\|_2 \leq \|\Pi_1 - \Pi_2\|_2 \cdot \|M\|_2$. Finally, note that

$$\|\Pi_1 - \Pi_2\|_2 \leq \|\Pi_1 - \Pi_2\|_F = \sqrt{2}\cdot d_C(U_1,U_2),$$

from which the claim follows. □

1.3.3 Probability Basics

Given a probability distribution $\mathcal{D}$, we use $x \sim \mathcal{D}$ to denote an independent sample from $\mathcal{D}$. Given a finite set S, we will use $x \sim_u S$ to denote x sampled uniformly at random from S.

Given distributions P, Q, the *total variation distance* between P and Q is $d_{\mathrm{TV}}(P, Q) \triangleq \frac{1}{2}\|P - Q\|_1$.

If P is absolutely continuous with respect to Q, let $\frac{\mathrm{d}P}{\mathrm{d}Q}(\cdot)$ denote the Radon-Nikodym derivative. The KL-divergence between P and Q is $\mathrm{KL}\left(P\|Q\right) \triangleq \mathbb{E}_{x\sim Q}[\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\log\frac{\mathrm{d}P}{\mathrm{d}Q}(x)]$. The chi-squared divergence between P and Q is $\chi^2\left(P\|Q\right) \triangleq \mathbb{E}_{x\sim Q}[\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x) - 1\right)^2]$.

Let $\Delta^n \subset \mathbb{R}^n$ be the simplex of nonnegative vectors whose coordinates sum to 1. Any $p \in \Delta^n$ naturally corresponds to a probability distribution over $[n]$.

We note that throughout this thesis, we will freely abuse notation and use the same symbols to denote probability distributions, their laws, and their density functions.

1.3.4 Fourier Transform

We will use the following convention in defining the continuous Fourier transform. Given square-integrable $f : \mathbb{R}^d \to \mathbb{R}$, define

$$\widehat{f}(\omega) \triangleq \int_{\mathbb{R}^d} f(x) \cdot e^{-2\pi i\langle\omega,x\rangle}\,dx. \tag{1.13}$$

The following fact about Fourier transforms of Gaussian pdfs is standard.

Fact 1.3.12.

$$\widehat{\mathcal{N}(0,\sigma^2)}[\omega] = e^{-2\pi^2\omega^2\sigma^2} = \frac{1}{\sqrt{2\pi}\sigma}\mathcal{N}\left(0, \frac{1}{4\pi^2\sigma^2}; \omega\right)$$

1.3.5 Concentration

Gaussians

Given $x \in \mathbb{R}$, let $\mathcal{N}(0, 1, x)$ denote the standard Gaussian density's value at x. Let χ^2_m denote the chi-squared distribution with m degrees of freedom.

We will need the following elementary estimates for Gaussian tails and correlated Gaussians. Define $\mathrm{erf}(\beta) \triangleq \mathrm{Pr}_{h\sim\mathbb{N}(0,1)}[|h| \le \beta]$ and $\mathrm{erfc}(\beta) \triangleq 1 - \mathrm{erf}(\beta)$ (note we eschew the usual normalization). It is an elementary fact that under this normalization, for all $z > 0$ we have that $\mathrm{erfc}(z) \le e^{-z^2/2}$. We sometimes use the following estimates:

Fact 1.3.13 (See e.g. Proposition 2.1.2 in [Ver18])**.**

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \le \mathrm{erfc}(t) \le \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Fact 1.3.14. *The function $f : \mathbb{R}_{\ge 0} \to \mathbb{R}$ given by $f(z) = erfc(1/\sqrt{z}) \cdot z$ is convex over $\mathbb{R}_{\ge 0}$.*

Proof. We can explicitly compute

$$f''(z) = \frac{e^{-1/2z}(1+z)}{2z^{5/2}\sqrt{2\pi}},$$

which is clearly nonnegative for any $z \ge 0$. □

We will use the following fundamental fact about moments of Gaussian polynomials repeatedly:

Fact 1.3.15 (Gaussian hypercontractivity, see e.g. Theorem 11.23 of [O'D14])**.** *For a polynomial $f : \mathbb{R}^r \to \mathbb{R}$ of degree d, and integer $q \ge 2$,*

$$\mathbb{E}[f(g)^q]^{1/q} \le (q-1)^{d/2} \mathbb{E}[f(g)^2]^{1/2},$$

where the expectation is over $g \sim \mathbb{N}(0,1)$.

An immediate consequence of this is the following tail bound for Gaussian polynomials:

Lemma 1.3.16. *Let $Z_1, ..., Z_T$ be iid scalar random variables which are each given by polynomials of degree d in Gaussian variables $\zeta_1, ..., \zeta_T \sim \mathcal{N}(0,1)$ respectively. If $\mathbb{V}[Z] \le \sigma^2$ for each $i \in [T]$, then then for any $t > 0$,*

$$Pr\left[\left|\frac{1}{T}\sum_{i=1}^{T}(Z_i - \mathbb{E}[Z_i])\right| \ge \frac{1}{\sqrt{T}} \cdot O(\log(1/\delta))^{d/2} \cdot \sigma\right] \le \delta.$$

We will also occasionally use the following consequence of hypercontractivity:

Corollary 1.3.17. *For any integer $q \geq 2$, $\mathbb{E}_{g\sim\mathcal{N}(0,Id_m)}[\|g\|_2^{2q}]^{1/q} \leq (q-1)\cdot(m+1)$.*

Proof. By Fact 1.3.15 applied to $f(g) \triangleq \|g\|_2^2$ and $d = 2$, we have that $\mathbb{E}[\|g\|_2^{2q}]^{1/q} \leq \mathbb{E}[\|g\|_2^4]^{1/2}$. But it is straightforward to compute $\mathbb{E}[\|g\|_2^4] = m^2 + 2m$, from which the claim follows. □

We also have the following concentration inequality for the norm of a Gaussian vector:

Fact 1.3.18. *Let $g \sim \mathcal{N}(0, Id_d)$. There is a universal constant $c_{\text{shell}} > 0$ such that for all $t > 0$,*

$$\Pr_g\left[\left|\|g\|_2 - \sqrt{d}\right| \geq t\right] \leq 2e^{-c_{\text{shell}}t^2}$$

Facts 1.3.13 and 1.3.18 imply the following pair of inequalities about the correlation between a Gaussian vector and a given unit vector.

Corollary 1.3.19. *Let $w \in \mathbb{S}^{d-1}$, $g \sim \mathcal{N}(0, Id_d)$, and $v = g/\|g\|_2$. Then for any constant $0 < \gamma < 1/2$, the following holds.*

There exist increasing functions $\underline{f}_\gamma, \overline{f}_\gamma, D : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ such that for any absolute constants $0 < \underline{\alpha} \leq \overline{\alpha}$, we have that for $d \geq D(\overline{\alpha})$,

$$Pr[\langle v, w\rangle \geq \underline{\alpha}\cdot d^{-\gamma}] \geq e^{-\underline{\beta}d^{1-2\gamma}}$$

and

$$Pr\left[\langle v, w\rangle \leq \overline{\alpha}d^{-\gamma}\right] \geq 1 - e^{-\overline{\beta}d^{1-2\gamma}}$$

for $\underline{\beta} = f(\underline{\alpha})$ and $\overline{\beta} = f(\overline{\alpha})$.

We defer the proof of this to Section 8.12.3

It will also be useful to obtain a similar bounds for the probability that the inner products of a random vector with two orthogonal directions are *simultaneously* in a particular range.

Corollary 1.3.20. *Let $w_1, w_2 \in \mathbb{S}^{d-1}$ be orthogonal, $g \sim \mathcal{N}(0, Id_d)$, and $v = g/\|g\|_2$.*

For any $\alpha_1, \alpha_2 > 0$, we have that for sufficiently large d,

$$Pr\left[\left(\langle v, w_1\rangle \geq \alpha_1\cdot d^{-1/4}\right) \wedge \left(\langle v, w_2\rangle \leq \alpha_2\cdot d^{-1/2}\right)\right] \geq \frac{1}{\text{poly}(d)}Pr[\langle v, w_1\rangle \geq \alpha_1\cdot d^{-1/4}]. \tag{1.14}$$

We defer the proof of this to Section 8.12.4

Standard Inequalities

Here we use standard martingale terminology, see e.g. [Dur19]; in particular, we say that a sequence of random variables $X_1, \dots, X_t$ adapted to a filtration $\mathcal{F}_t$ form a *martingale difference sequence* if $\mathbb{E}[X_t | \mathcal{F}_{t-1}] = 0$ for all t.

We say a mean-zero random variable X is σ^2-subgaussian if $\log \mathbb{E}[e^{\lambda X}] \le \lambda^2 \sigma^2/2$ for all $\lambda \in \mathbb{R}$; recall that if $|X| \le K$ then X is $O(K^2)$-subgaussian [Ver18].

Fact 1.3.21 (Azuma-Hoeffding inequality)**.** *Suppose that $X_1, \dots, X_n$ is a martingale difference sequence and $|X_i| \le M_i$ almost surely. Then*

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \ge t\right] \le \exp\left(-\Omega\left(\frac{nt^2}{\frac{1}{n}\sum_i M_i^2}\right)\right)$$

Fact 1.3.22 (Bernstein's inequality)**.** *For $X_1, ..., X_n$ independent and mean-zero, if $|X_i| \le M$ for all i, then for all $t > 0$,*

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \ge t\right] \le \exp\left(-\Omega\left(\frac{nt^2}{\frac{1}{n}\sum \mathbb{E}[X_i^2] + Mt}\right)\right)$$

We will use the following general version of the Azuma-Hoeffding inequality, which applies to martingales in Euclidean space of arbitrary dimension with subgaussian step sizes. (Note: this result is false if we consider martingales with steps that are general subgaussian vectors, which can make steps of size $\sqrt{d}$ in dimension d.) This result follows from the same proof as Equation 5.18 in [KS91], with some small differences: there they consider bounded variation processes instead of discrete-time martingales. In the bounded step size case, optimal constants were obtained in [Pin94]. For completeness, we prove Theorem 1.3.23 in the Appendix.

Theorem 1.3.23 (Subgaussian-step vector Azuma-Hoeffding, cf. Equation 5.18 in [KS91])**.** *Suppose that $X_1, \dots, X_n$ are random vectors in Euclidean space with $\|X_t\| \le 1$ almost surely for all t, and $\xi_1, \dots, \xi_n$ are random variables such that almost surely, the law of ξ_t conditional* *on $X_1, \dots, X_t, \xi_1, \dots, \xi_{t-1}$ is mean-zero and σ^2-subgaussian. Then*

$$\Pr\left[\left\|\frac{1}{n}\sum_{i=1}^n \xi_i X_i\right\| \geq u\right] \leq 2\exp\left(-\Omega\left(\frac{nu^2}{\sigma^2}\right)\right).$$

Sub-Exponential Random Variables

Here we define the sub-exponential norm of a random variable X to be $\sup_{p\geq 1}\frac{1}{p}\mathbb{E}[|X|^p]^{1/p}$. *Sub-exponential* random variables, that is, ones with bounded sub-exponential norm, enjoy the following tail bound:

Fact 1.3.24 (Sub-exponential tail bounds, see e.g. [Ver10], Proposition 5.16)**.** *If $X_1, ..., X_N$ are i.i.d. random variables with mean zero and sub-exponential norm K, then*

$$\Pr\left[\left|\frac{1}{N}\sum_{i=1}^N X_i\right| \geq t\right] \leq 2\exp\left(-\Omega\left(\frac{Nt^2}{K^2} \wedge \frac{Nt}{K}\right)\right). \tag{1.15}$$

In particular, for any $\delta > 0$, if we take $N = \Theta\left(\frac{K^2}{t^2} \vee \frac{K}{t}\right) \cdot \log 1/\delta$, then (1.15) is at most δ.

We also use the following fact about moment bounds for *sums* of sub-exponential random variables.

Lemma 1.3.25. *Fix any $t \in \mathbb{N}$. Given a collection of independent mean-zero random variables $Z_1, \dots, Z_m$ whose odd moments vanish and such that for every $i \in [m]$ and $1 \leq \ell \leq t$, $\mathbb{E}[|Z_i|^\ell]^{1/\ell} \leq \ell \cdot \sigma_i$, we have that for every integer $1 \leq \ell \leq t$*

$$\mathbb{E}[(Z_1 + \cdots + Z_m)^\ell]^{1/\ell} \leq \ell(\sigma_1^2 + \cdots + \sigma_m^2)^{1/2}$$

Proof. Using the sub-exponential moment bound, we can expand $\mathbb{E}[(Z_1 + \cdots + Z_m)^\ell]$ and use the fact that the Z_i's are independent to get

$$\mathbb{E}[(Z_1+\cdots+Z_m)^\ell] = \sum_\alpha \prod_i \mathbb{E}[Z_i^{\alpha_i}] \leq \sum_\alpha \prod_i \alpha_i^{\alpha_i}\sigma_i^{\alpha_i} \leq \ell^\ell \sum_\alpha \prod_i (\sigma_i^2)^{\alpha_i/2} = \ell^\ell(\sigma_1^2+\cdots+\sigma_m^2)^{\ell/2}$$

where α ranges over even monomials of total degree ℓ. □

Matrix Concentration

A key ingredient in our arguments in Chapter 6 is concentration for matrix martingales. See [Tro12, Tro11] for background on matrix concentration; for infinite dimensional settings we use a version of matrix concentration which depends on effective dimension [HKZ+12, Min17]. To briefly recall, a matrix martingale $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ adapted to a filtration $\mathcal{F}_t$ with difference sequence $\mathbf{X}_t$ is an $\mathcal{F}_t$-adapted process satisfying $\mathbf{Y}_t = \sum_{s=1}^{t} \mathbf{X}_s$ and $\mathbb{E}[\mathbf{X}_t|\mathcal{F}_{t-1}] = 0$. We also recall that for a function $f : \mathbb{R} \to \mathbb{R}$ and symmetric matrix M with eigendecomposition $M = \sum_i \lambda_i \rho_i \rho_i^T$, the notation $f(M)$ corresponds to applying f to the spectrum, i.e. $f(M) = \sum_i f(\lambda_i)\rho_i\rho_i^T$.

Theorem 1.3.26 (Matrix Freedman Inequality, [Min17])**.** *Suppose $\mathbf{Y}_1, ..., \mathbf{Y}_n \in \mathbb{R}^{d\times d}$ is a symmetric matrix martingale adapted to filtration $\mathcal{F}_t$, whose associated difference sequence $\{\mathbf{X}_t\}$ satisfies $\|\mathbf{X}_t\| \leq 1$ almost surely for all t. Let $\mathbf{W} = \sum_t \mathbb{E}[\mathbf{X}_t^2|\mathcal{F}_{t-1}]$, then for any $t \geq \frac{1}{6}(1 + \sqrt{1 + 36\sigma_n^2})$*

$$\Pr[\|\mathbf{Y}_n\| \geq t \textit{ and } \|\mathbf{W}\| \leq \sigma_n^2] \leq 50 d_1(t) \cdot \exp\left(\frac{-t^2/2}{\sigma_n^2 + t/3}\right)$$

where

$$d_1(t) = \mathit{Tr}\, f(t\, \mathbb{E}[\mathbf{W}]/\sigma_n^2)$$

and $f(x) = \min(1, x)$.

Corollary 1.3.27. *In the same setting as Theorem 1.3.26, suppose that for some $\sigma \leq 1$, $\mathbb{E}[\mathbf{X}_t^2|\mathcal{F}_{t-1}] \preceq \sigma^2$ almost surely. Then for any $u \geq 1/18n + \sigma\sqrt{1/n}$*

$$\Pr[\|(1/n) \cdot \mathbf{Y}_n\| \geq u] \leq 50 d_2(u) \cdot \exp\left(\frac{-nu^2/2}{\sigma^2 + u/3}\right)$$

where

$$d_2(u) = \mathit{Tr}\, f(u\, \mathbb{E}[\mathbf{W}]/\sigma^2)$$

and $f(x) = \min(1, x)$ as in Theorem 1.3.26.

Proof. Apply Theorem 1.3.26 with $t = nu$ and $\sigma_n^2 = n\sigma^2$, noting that $d_2(u) = d_1(nu)$; in this statement, we only strengthened the assumed lower bound on t. □

Nets

Fact 1.3.28 (e.g. [Ver18], Corollary 4.2.13)**.** *For any $\varepsilon > 0$, there is an ε-net (in L_2 norm) of size $(1 + 2/\varepsilon)^m$ for the unit L_2 ball in m dimensions.*

Corollary 1.3.29. *For any $\varepsilon, \beta > 0$, there is an ε-net (in operator norm) for the set of $m_1 \times m_2$ matrices of operator norm at most β of size at most $(1 + 2\beta/\varepsilon)^{m_1 m_2}$.*

Proof. As operator norm is upper bounded by Frobenius norm, an ε-net in Frobenius norm for the set of $m_1 \times m_2$ matrices of Frobenius norm at most β would contain the claimed ε-net. The former can be obtained from scaling an ε/β-net in Frobenius norm for the set of $m_1 \times m_2$ matrices of unit Frobenius norm, and such a net with size $(1+2\beta/\varepsilon)^{m_1 m_2}$ exists by Fact 1.3.28. □

Integration by Parts

The following fact can be used to translate tail bounds into bounds on expectation values of functionals.

Fact 1.3.30 (Integration by parts)**.** *Let $a, b \in \mathbb{R}$. Let Z be a nonnegative random variable satisfying $Z \le b$ and such that for all $x \ge a$, $\Pr[Z > x] \le \tau(x)$. Let $f : [0, b] \to \mathbb{R}_{\ge 0}$ be nondecreasing and differentiable. Then*

$$\mathbb{E}[f(Z)] \le f(a)(1 + \tau(a)) + \int_a^b \tau(x) f'(x) \, \mathrm{d}\, x.$$

Proof. Let $g : [0, b] \to [0, 1]$ denote the CDF of Z, so that for $x \ge a$, $1 - g(x) \le \tau(x)$. Then

$$\begin{aligned}
\mathbb{E}[Z^n] &= \int_0^b f(Z) \, \mathrm{d}\, g \le f(a) + \int_a^b f(Z) \, \mathrm{d}\, g \\
&= f(a) + f(b)g(b) - f(a)g(a) - \int_a^b g(x) f'(x) \, \mathrm{d}\, x \\
&= f(a)(1 - g(a)) + f(b) - (f(b) - f(a)) + \int_a^b (1 - g(x)) f'(x) \, \mathrm{d}\, x \\
&\le f(a)(1 + \tau(a)) + \int_a^b \tau(x) f'(x) \, \mathrm{d}\, x,
\end{aligned}$$

where the first integral is the Riemann-Stieltjes integral, the third step is integration by parts, the fourth step follows because $g(b) = 1$, and the last follows because $1 - g(x) \leq \tau(x) \leq 1$ for $x \geq a$. □

Anti-Concentration

We will sometimes need to show that a certain random variable is not too small too often. The following basic estimate will suffice for our purposes.

Fact 1.3.31 (Elementary anticoncentration)**.** *If Z is a random variable for which $|Z| \leq M$ almost surely, and $\mathbb{E}[Z^2] \geq \sigma^2$, then $\Pr[|Z| \geq t] \geq \frac{1}{M^2}(\sigma^2 - t^2)$.*

Proof. We have

$$\begin{aligned}\sigma^2 \leq \mathbb{E}\big[Z^2\big] &= \mathbb{E}\big[Z^2 \mid |Z| \geq t\big] \cdot \Pr[|Z| \geq t] + \mathbb{E}\big[Z^2 \mid |Z| < t\big] \cdot \Pr[|Z| < t] \\ &\leq M^2 \cdot \Pr[|Z| \geq t] + t^2,\end{aligned}$$

from which the claimed bound follows upon rearranging. □

Other Tail Bounds

We will need the following standard concentration inequality in our analysis of filtered PCA.

Lemma 1.3.32 ([Ver10])**.** *Let $\phi : \mathbb{R} \to [0,1]$ be any function. Let $M = \mathbb{E}_{x \sim \mathcal{N}(0, Id_n)}[\phi(x) \cdot (xx^\top - Id)]$. If $x_1, ..., x_N \sim \mathcal{N}(0, Id_n)$ for $N = \Omega(\{n \vee \log(1/\delta)\}/\varepsilon^2)$, then*

$$Pr\left[\left\|\mathbf{M} - \frac{1}{N}\sum_{i=1}^{N} \phi(x_i) \cdot (x_i x_i^\top - Id)\right\|_2 \geq \varepsilon\right] \leq \delta.$$

Proof. This follows from standard sub-Gaussian concentration; see e.g. Remark 5.40 in [Ver10]. □

We will occasionally also use the following concentration inequality for sums of random variables which only satisfy one-sided bounds. This is a specialization of the martingale concentration result of [Ben03] to the iid case, though we also need that result in its full generality for Lemma 2.3.4 below.

Lemma 1.3.33 (Special case of [Ben03])**.** *Let $Z_1, ..., Z_T$ be iid, mean-zero random variables. Let $c, s > 0$ be deterministic constants for which $Z_i \le c$ with probability one and $\mathbb{V}[Z_i] \le s^2$ for all $i \in [T]$. Let $\sigma = c \vee s$. Then for any $\delta > 0$,*

$$Pr\left[\frac{1}{T}\sum_{i=1}^{T} Z_i \ge \frac{1}{\sqrt{T}} \cdot \sqrt{2}\log(1/\delta) \cdot \sigma\right] \le \delta.$$

1.3.6 Hermite Polynomials

Let He_ℓ denote the degree-ℓ probabilist's Hermite polynomial. It will be convenient to scale these to form the *normalized Hermite polynomials*

$$\phi_\ell(z) = \frac{1}{(\ell!)^{1/2}} \underset{\ell}{\mathrm{He}}(z)$$

for $\ell \in \mathbf{Z}_{\ge 0}$. This scaling is chosen so that $\{\phi_\ell\}$ forms an orthonormal basis for $L^2(\mathbb{R}, \gamma)$.

The following identity will be useful:

Fact 1.3.34 (See e.g. Proposition 11.31 in [O'D14])**.** *For any $v, v' \in \mathbb{S}^{n-1}$ and $\ell, \ell' \in \mathbf{Z}_{\ge 0}$,*

$$\mathbb{E}_{g\sim\mathbb{N}(0,Id_n)}\left[\phi_\ell(\langle v, g\rangle \cdot \phi_{\ell'}(\langle v', g\rangle)\right] = \mathbb{1}[\ell = \ell'] \cdot \langle v, v'\rangle^\ell.$$

From the normalized Hermite polynomials one can construct an orthonormal basis for $L^2(\mathbb{R}^d, \gamma)$ by defining for any multiset I consisting of elements of $[d]$ the *tensored normalized Hermite polynomial*

$$\phi_I(z) \triangleq \prod_{i=1}^{d} \phi_{\ell_i}(z_i),$$

where ℓ_i denotes the number of occurrences of i in the multiset I.

1.3.7 Stability of Linear Threshold Functions

The following lemma is a crucial ingredient in our analysis of filtered PCA in Chapters 2 and 3.

Lemma 1.3.35. *For $\tau > 0$ and vectors $v, v' \in \mathbb{R}^d$,*

$$\Pr_{x\sim\mathcal{N}(0,Id)}[\langle v, x\rangle > \tau \wedge \langle v', x\rangle \leq \tau] \leq O\left(\frac{\|v - v'\|}{\tau}\right) \tag{1.16}$$

While the result is fairly elementary, we are not aware of such a result appearing in previous works and therefore give a self-contained proof.

Proof. First note that without loss of generality, we may assume that $\|v\| \geq \|v'\|$; if not, then the random variable $\mathbb{1}[\langle v, x\rangle > \tau \wedge \langle v', x\rangle \leq \tau]$ is stochastically dominated by $\mathbb{1}[\langle v, x\rangle > \tau \wedge \langle \zeta v', x\rangle \leq \tau]$ for $\zeta = \|v\|/\|v'\|$, and furthermore $\|v - \zeta v'\| \leq \|v - v'\|$ by the Pythagorean theorem.

Also note that we may assume $\|v'\| > \|v - v'\|$. Otherwise, we would have $\|v\| \leq 2\|v - v'\|$. But then we could upper bound the left-hand side of (1.16) by

$$\Pr[\langle v, x\rangle > \tau] \leq e^{-\tau^2/2\|v\|^2} \leq e^{-\frac{\tau^2}{8\|v-v'\|^2}} \leq 2\|v - v'\|/\tau.$$

Now define $\widehat{v} = v/\|v\|$ and $\widehat{v'} = v'/\|v'\|$ so that (1.16) equals $\Pr[\langle \widehat{v}, x\rangle > \widehat{\tau} \wedge \langle \widehat{v'}, x\rangle \leq \widehat{\tau}']$ for $\widehat{\tau} \triangleq \tau/\|v\|$ and $\widehat{\tau}' \triangleq \tau/\|v'\|$. Write $\widehat{v}' = \alpha\widehat{v} + \sqrt{1-\alpha^2}v^\perp$ for $v^\perp$ orthogonal to $\widehat{v}$, and denote the random variables $\langle \widehat{v}, x\rangle$ and $\langle \widehat{v}', x\rangle$ by γ and γ' respectively (these are α-correlated standard Gaussians).

Note that by the assumption that $\|v\| \geq \|v'\| \geq \|v - v'\|$, the angle between v and v' is at most $\pi/3$, so $\alpha \geq 1/2$.

We are now ready to upper bound (1.16). We will split into two cases, either $\gamma > \widehat{\tau}'/\alpha$ or $\widehat{\tau} \leq \gamma \leq \widehat{\tau}'$, and upper bound the contribution of either case to the probability in (1.16) by $O(\|v - v'\|/\tau)$, from which the lemma will follow.

Case 1: $\gamma > \widehat{\tau}'/\alpha$.

The density of γ' relative to γ is given by

$$\int_{-\infty}^{\frac{\widehat{\tau}'-\alpha\gamma}{\sqrt{1-\alpha^2}}} \mathcal{N}(0,1,x)dx = \frac{1}{2}\operatorname{erfc}\left(\frac{\alpha\gamma - \widehat{\tau}'}{\sqrt{1-\alpha^2}}\right) \leq \frac{1}{2}\exp\left(-\frac{(\alpha\gamma - \widehat{\tau}')^2}{2(1-\alpha^2)}\right).$$

We have that

$$\begin{aligned}\mathop{\mathbb{E}}_{\gamma}\left[\frac{1}{2}\exp\left(-\frac{(\alpha\gamma-\widehat{\tau}')^2}{2(1-\alpha^2)}\right)\cdot \mathbb{1}[\gamma>\widehat{\tau}']\right] &= \frac{1}{4}\sqrt{1-\alpha^2}\cdot\exp(-\widehat{\tau}'^2/2)\cdot\operatorname{erfc}(\widehat{\tau}'\sqrt{1-\alpha^2}/\alpha)\\ &\le \frac{1}{4}\sqrt{1-\alpha^2}\cdot\exp(-\widehat{\tau}'^2/2\alpha^2)\\ &\le \frac{\|v-v'\|}{4\sqrt{2}\|v'\|}\cdot\frac{|\alpha|\sqrt{2}}{\widehat{\tau}'}\le\frac{\|v-v'\|}{4\tau},\end{aligned}$$

where the first step is standard Gaussian integration, the second step uses the inequality $\operatorname{erfc}(z)\le e^{-z^2/2}$ for all $z\ge 0$, and the third step uses the fact that $\exp(-x)\le 1/x$ for all $x>0$ and the fact that $\sqrt{1-\alpha^2}=\frac{1}{\sqrt{2}}\|\widehat{v}-\widehat{v}'\|\le\frac{\|v-v'\|}{\sqrt{2}\|v'\|}$.

Case 2: $\widehat{\tau}<\gamma\le\widehat{\tau}'/\alpha$.

We can naively upper bound the probability $\widehat{\tau}<\gamma\le\widehat{\tau}'/\alpha$ and $\gamma'\le\widehat{\tau}'$ by the probability $\widehat{\tau}<\gamma\le\widehat{\tau}'/\alpha$, which is at most $e^{-\widehat{\tau}^2/2}\cdot(\widehat{\tau}'/\alpha-\widehat{\tau})$. Note that

$$\widehat{\tau}'/\alpha-\widehat{\tau}\le\tau\cdot\left(\frac{1/\alpha}{\|v'\|}-\frac{1}{\|v\|'+\|v-v'\|}\right)\le\frac{\tau}{\alpha}\cdot\frac{(1-\alpha)\|v'\|+\|v-v'\|}{\|v'\|^2}\le\frac{3\tau\|v-v'\|}{2\alpha\|v'\|^2},\quad(1.17)$$

where in the last step we have used that $1-\alpha=\frac{1}{2}\|\widehat{v}-\widehat{v}'\|\le\frac{\|v-v'\|}{2\|v'\|}$.

Suppose to the contrary that $e^{-\widehat{\tau}^2/2}\cdot(\widehat{\tau}'/\alpha-\widehat{\tau})>\frac{9\|v-v'\|}{\tau}$ so that by (1.17),

$$e^{\widehat{\tau}^2/2}<\frac{\tau^2}{6\alpha\|v'\|^2}.\qquad(1.18)$$

Recall that we may assume that $\|v'\|\ge\|v-v'\|$, so $\widehat{\tau}\ge\frac{\tau}{2\|v'\|}$, and that $\alpha\ge 1/2$. From this, (1.18) would imply that $e^{\frac{\tau^2}{8\|v'\|^2}}<\frac{\tau^2}{3\|v'\|^2}$, and such an inequality cannot hold. □

1.3.8 Sum-of-Squares Programming

For a thorough treatment on the SoS proof system, we refer the reader to [OZ13, BS14]. In this section we review essential components that are needed in Chapter 4 and parts of Chapter 6.

Let $x_1,...,x_n$ be formal variables, and let Program $\mathcal{P}$ be a set of polynomial equations and inequalities $\{p_1(x)\ge 0,...,p_m(x)\ge 0,q_1(x)=0,...,q_m(x)=0\}$.

We say that the inequality $p(x) \geq 0$ has a degree-d SoS proof using $\mathcal{P}$ if there exists a polynomial $q(x)$ in the ideal generated by $q_1(x), ..., q_m(x)$ at degree d, together with sum-of-squares polynomials $\{r_S(x)\}_{S \subseteq [m]}$ (where the index S is a multiset), such that

$$p(x) = q(x) + \sum_{S \subseteq [m]} r_S(x) \cdot \prod_{i \in S} p_i(x),$$

and such that $\deg(r_S(x) \cdot \prod_{i \in S} p_i(x)) \leq d$ for each multiset $S \subseteq [m]$. We denote this by the notation

$$\mathcal{P} \vdash_d p(x) \geq 0$$

When $\mathcal{P} = \{1\}$, we will denote this by $\vdash_d p(x) \geq 0$.

A fact we will use throughout without comment is that SoS proofs compose well:

Fact 1.3.36. *If $\mathcal{P} \vdash_d p(x) \geq 0$ and $\mathcal{B} \vdash_{d'} q(x) \geq 0$, then $\mathcal{P} \cup \mathcal{B} \vdash_{\max(d,d')} p(x) + q(x) \geq 0$ and $\mathcal{P} \cup \mathcal{B} \vdash_{dd'} p(x)q(x) \geq 0$.*

It is useful to consider the objects dual to SoS proofs, namely pseudodistributions. A degree-d pseudodistribution is a linear functional $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leq d} \to \mathbb{R}$ satisfying the following properties:

1. Normalization: $\tilde{\mathbb{E}}[1] = 1$

2. Positivity: $\tilde{\mathbb{E}}[p(x)^2]$ for every p of degree at most $d/2$.

We will use the terms "pseudodistribution" and "pseudoexpectation" interchangeably.

A degree-d pseudodistribution $\tilde{\mathbb{E}}$ *satisfies* Program $\mathcal{P} = \{p_1(x) \geq 0, ..., p_m(x) \geq 0, q_1(x) = 0, ..., q_m(x) = 0\}$ if for every multiset $S \subseteq [m]$ and sum-of-squares polynomial $r(x)$ for which $\deg(r(x) \cdot \prod_{i \in S} p_i(x)) \leq d$, we have $\tilde{\mathbb{E}}[r(x) \cdot \prod_{i \in S} p_i(x)] \geq 0$, and for every $q(x)$ in the ideal generated by $q_1, ..., q_m$ at degree d, we have $\tilde{\mathbb{E}}[q(x)] = 0$.

For any fixed $\ell \in \mathbb{N}$, given such a program $\mathcal{P}$, one can efficiently compute a degree ℓ pseudodistribution satisfying $\mathcal{P}$ in polynomial time:

Fact 1.3.37. *([Nes00], [Par00], [Las01], [Sho87]). For any $n, \ell \in \mathbf{Z}^+$, let $\widetilde{\mathbb{E}}_\zeta$ be degree ℓ pseudodistribution satisfying a polynomial system $\mathcal{P}$. Then the following set has a $n^{O(\ell)}$-time* *weak separation oracle (in the sense of [GLS81]):*

$$\{\widetilde{\mathbb{E}}_\zeta(1, x_1, x_2, ..., x_n)^{\otimes \ell} | \text{ degree } \ell \text{ pseudoexpectations } \widetilde{\mathbb{E}}_\zeta \text{ satisfying } \mathcal{P}\}$$

Using this separation oracle, the ellipsoid algorithm finds a degree ℓ pseudoexpectation in time $n^{O(\ell)}$, which we call the degree ℓ sum-of-squares algorithm.

The following fundamental fact is a consequence of SDP duality:

Fact 1.3.38. *If $\mathcal{P} \vdash_d p(x) \geq 0$ and $\tilde{\mathbb{E}}$ is a degree-d pseudodistribution satisfying $\mathcal{P}$, then $\tilde{\mathbb{E}}$ satisfies $\mathcal{P} \cup \{p \geq 0\}$.*

We collect some basic inequalities that are captured by the SoS proof system, the proofs of which can be found, e.g., in Appendix A of [HL18] and [MSS16].

Fact 1.3.39 (SoS Cauchy-Schwarz). *Let $x_1, ..., x_n, y_1, ..., y_n$ be formal variables. Then*

$$\vdash_4 \left(\sum_{i=1}^n x_i y_i\right)^2 \leq \left(\sum_{i=1}^n x_i^2\right) \cdot \left(\sum_{i=1}^n y_i^2\right)$$

Fact 1.3.40 (SoS Holder's). *Let $w_1, ..., w_n, x_1, ..., x_n$ be formal variables. Then for any $t \in \mathbb{N}$ a power of 2, we have*

$$\{w_i^2 = w_i \ \forall i \in [n]\} \vdash_{O(t)} \left(\sum_{i=1}^n w_i x_i\right)^t \leq \left(\sum_{i=1}^n w_i\right)^{t-1} \cdot \sum_{i=1}^n x_i^t$$

and

$$\{w_i^2 = w_i \ \forall i \in [n]\} \vdash_{O(t)} \left(\sum_{i=1}^n w_i x_i\right)^t \leq \left(\sum_{i=1}^n w_i\right)^{t-1} \cdot \sum_{i=1}^n w_i x_i^q.$$

We will also use the following consequence of scalar Holder's inequality.

Fact 1.3.41. *Let $\ell(x)$ be a linear form in the formal variables $x_1..., x_n$. Then if $\tilde{\mathbb{E}}$ is a degree-t pseudodistribution, then*

$$\tilde{\mathbb{E}}[\ell(x)]^t \leq \tilde{\mathbb{E}}[\ell(x)^t].$$

Proof. Because $\tilde{\mathbb{E}}$ is a degree-t pseudodistribution, there exists a pseudo-density $H(\cdot)$ such that $\tilde{\mathbb{E}}[p(x)] = \sum_x H(x) \cdot p(x)$ for any degree-t polynomial p. So by scalar Holder's inequality we get that

$$\tilde{\mathbb{E}}[\ell(x)]^t = \left(\sum_x H(x)\ell(x)\right)^t \le \left(\sum_x H(x)\right)^{t-1} \cdot \left(\sum_x H(x)\ell(x)^t\right) = \tilde{\mathbb{E}}[1]^{t-1} \cdot \tilde{\mathbb{E}}[\ell(x)^t] = \tilde{\mathbb{E}}[\ell(x)^t]$$

as claimed. □

Fact 1.3.42. *(Pseudoexpectation Cauchy Schwarz). Let $f(x)$ and $g(x)$ be degree at most $\ell \le \frac{D}{2}$ polynomial in indeterminate x, then*

$$\widetilde{\mathbb{E}}[f(x)g(x)]^2 \le \widetilde{\mathbb{E}}[f(x)^2]\widetilde{\mathbb{E}}[g(x)^2].$$

We make extensive use of the following to analyze certain roundings of pseudodistributions:

Lemma 1.3.43. *For any psd matrix $\mathbf{\Sigma}$ which induces a norm $\|\cdot\|_{\mathbf{\Sigma}}$, any vector w^*, and any degree-2 pseudoexpectation $\widetilde{\mathbb{E}}[\cdot]$ over $\mathbb{R}^d$-valued variable w, we have that*

$$\|\widetilde{\mathbb{E}}[w] - w^*\|_{\mathbf{\Sigma}}^2 \le \widetilde{\mathbb{E}}[\|w - w^*\|_{\mathbf{\Sigma}}^2]. \tag{1.19}$$

Proof. By the dual definition of L_2 norm, the left-hand side of (1.19) can be written as

$$\sup_{v \in \mathbb{S}^{d-1}} \langle \mathbf{\Sigma} v, \widetilde{\mathbb{E}}[w] - w^* \rangle^2.$$

For any $v \in \mathbb{S}^{d-1}$,

$$\left\langle \mathbf{\Sigma} v, \widetilde{\mathbb{E}}[w] - w^* \right\rangle^2 = \left(\widetilde{\mathbb{E}}[\langle \mathbf{\Sigma} v, w - w^* \rangle]\right)^2 \le \widetilde{\mathbb{E}}[\langle \mathbf{\Sigma} v, w - w^* \rangle^2] \le \widetilde{\mathbb{E}}[\|w - w^*\|_{\mathbf{\Sigma}}^2],$$

where the first inequality follows by the pseudoexpectation version of SoS Cauchy-Schwarz (see e.g. Lemma A.5 of [BKS14]). Therefore, taking the maximum over all $v \in S^{d-1}$ proves the inequality. □

The following elementary inequality will also be useful.

Fact 1.3.44. $\{x^2 = 1\} \vdash_2 -1 \leq x \leq 1$.

Proof. Noting that

$$1 - x = \frac{1}{2}(1 - x^2 + (x-1)^2) \quad \text{and} \quad 1 + x = \frac{1}{2}(1 - x^2 + (x+1)^2), \tag{1.20}$$

the claim follows. □

Finally, we note that SoS can also be used to prove spectral upper and lower bounds:

Fact 1.3.45. *(Spectral Bounds) Let $A \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix with λ_{max} and λ_{min} being the largest and smallest eigenvalues of A respectively. Let $\widetilde{\mathbb{E}}$ be a pseudoexpectation with degree greater than or equal to 2 over indeterminates $v = (v_1, ..., v_d)$. Then we have*

$$\vdash_2 \langle A, vv^T\rangle \leq \lambda_{max}\|v\|^2 \quad \text{and} \quad \vdash_2 \langle A, vv^T\rangle \geq \lambda_{min}\|v\|^2.$$

1.3.9 Quantum Basics

Definition 1.3.46 (States)**.** *A d-dimensional quantum state is specified by a* density matrix *$\rho \in \mathbb{C}^{d\times d}$, i.e. a psd matrix that satisfies $Tr(\rho) = 1$.*

In Chapters 10 and 11, given a density matrix M, we will use $\widehat{M}$ to denote $M/\operatorname{Tr}(M)$.

Let $\rho_{\mathsf{mm}} \triangleq \frac{1}{d}\,\mathrm{Id}$ denote the maximally mixed state.

We now formally define the notion of a POVM with possibly infinite outcome set.

Definition 1.3.47 (POVMs)**.** *Given space Ω with Borel σ-algebra $\mathcal{B}(\Omega)$, let μ be a regular positive real-valued measure μ on $\mathcal{B}(\Omega)$, and let $M : \Omega \to \mathbb{C}^{d\times d}$ be a measurable function taking values in the set of psd Hermitian matrices. We will denote the image of $x \in \Omega$ under M by M_x.*

We say that the pair (μ, M) specifies a POVM *$\mathcal{M}$ if $\int_\Omega M\,\mathrm{d}\mu = Id_{d\times d}$ and, for any $d \times d$ density matrix ρ, the map $B \mapsto \int_B \langle M_x, \rho\rangle\,\mathrm{d}\mu$ for $B \in \mathcal{B}(\Omega)$ specifies a probability measure over Ω. We call the distribution given by this measure the* distribution over outcomes from measuring ρ with $\mathcal{M}$.[7]

[7] This definition looks diferent from standard ones because we are implicitly invoking the Radon-Nikodym theorem for POVMs on finite-dimensional Hilbert spaces, see e.g. Theorem 3 from [MHC13] or Lemma 11 from [CDS10].

Given a POVM $\mathcal{M}$, we will refer to the space of measurement outcomes as $\Omega(\mathcal{M})$.

With no meaningful loss in understanding, the reader may simply imagine that all POVMs mentioned henceforth have finitely many outcomes so that a POVM is simply the data of some finite set of positive semidefinite Hermitian matrices $\{M_x\}_{x\in\Omega}$ for which $\sum_x M_x = \mathrm{Id}_{d\times d}$, though our arguments extend to the full generality of Definition 1.3.47.

Haar-Random Unitary Matrices

In this section we recall some standard facts about Haar-unitary integrals. Given a permutation $\pi \in \mathcal{S}_\ell$, let $\mathrm{Wg}(\pi, d)$ denote the *Weingarten function*. Given a matrix $M \in \mathbb{C}^{d\times d}$ and permutation $\pi \in \mathcal{S}_\ell$, let $\langle M\rangle_\pi \triangleq \prod_{C\in\pi} \mathrm{Tr}(M^{|C|})$, where C ranges over the cycles of π and $|C|$ denotes the length of C. Equivalently, if P_π is the permutation operator associated to π, then $\langle M\rangle_\pi = \mathrm{Tr}(P_\pi M^{\otimes\ell})$.

Fact 1.3.48. *Given matrix $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{d\times d}$, for Haar-random $\mathbf{U} \in U(d)$, $\mathbb{E}_{\mathbf{U}}[\mathbf{U}^\dagger \mathbf{X}\mathbf{U}\mathbf{Y}] = \frac{1}{d}\, Tr(\mathbf{X})\, Tr(\mathbf{Y})$.*

Lemma 1.3.49. *For $d \geq 2$, $\ell \in \mathbb{N}$, and any $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{d\times d}$, we have that*

$$\mathop{\mathbb{E}}_{\mathbf{U}}[Tr(\mathbf{A}\mathbf{U}^\dagger \mathbf{B}\mathbf{U})^\ell] = \sum_{\sigma,\tau\in\mathcal{S}_\ell} \langle \mathbf{A}\rangle_\sigma \langle \mathbf{B}\rangle_\tau \,\mathrm{Wg}(\sigma\tau^{-1}, d)\ .$$

Lemma 1.3.50 ([Mon13]). *For $\ell \leq d^{2/3}$ and $\pi \in \mathcal{S}_\ell$, $\mathrm{Wg}(\pi, d) \leq O(d^{\kappa(\pi)-2\ell})$.*

Lemma 1.3.51. *For any $\ell \leq d^{2/3}$, $\sum_{\pi\in\mathcal{S}_\ell} |\mathrm{Wg}(\pi, d)| \leq \Omega(d)^{-\ell}$.*

Proof. Recalling Lemma 1.3.50, we see that

$$\sum_{\pi\in\mathcal{S}_\ell} |\mathrm{Wg}(\pi, d)| = \frac{1}{d^{2\ell}} \cdot \sum_\pi O(d)^{\kappa(\pi)} = \frac{O(d)(O(d)+1)\cdots(O(d)+\ell-1)}{d^{2\ell}} \leq \Omega(d)^{-\ell}$$

as claimed. □

Concentration of measure for Haar-random unitary matrices will also be crucial to our analysis:

Theorem 1.3.52 ([MM13], Corollary 17, see also [AGZ10], Corollary 4.4.28)**.** *Equip* $M \triangleq U(d)^k$ *with the* L_2*-sum of Frobenius metrics. If* $F : M \to \mathbb{R}$ *is* L*-Lipschitz, then for any* $t > 0$*:*

$$\Pr_{(\mathbf{U}_1,...,\mathbf{U}_k)\in M}[|F(\mathbf{U}_1, ..., \mathbf{U}_k) - \mathbb{E}[F(\mathbf{U}_1, ..., \mathbf{U}_k)]| \geq t] \leq e^{-dt^2/12L^2},$$

where $\mathbf{U}_1, ..., \mathbf{U}_k$ *are independent unitary matrices drawn from the Haar measure.*

Part I

Learning Rich Function Classes

Chapter 2

Low-Rank Polynomials

2.1 Introduction

Consider the classical *polynomial regression* problem in learning and statistics. In its most basic form, we receive samples of the form (x, y) with $x \in \mathbb{R}^n$ coming from some distribution and y is $P(x)$ for a degree at most d polynomial in x. Our goal is to *learn* the polynomial P. Here learning could either mean learning the coefficients of P or even finding some other function that gets small prediction error (as in find Q with $E[(Q(x) - P(x))^2] \ll Var(y)$).

Polynomial regression of course is one of the most basic primitives in statistics and machine learning especially in the more general *non-realizable* case. For example, it is crucial in many kernalization applications, and it gives the best known PAC learning algorithms for various central complexity classes such as constant-depth circuits [LMN93], intersection of halfspaces [KOS04], DNFs [KS04], convex sets [KOS08, Vem10a], the last of which even exploits intrinsic dimension as we do but for a different problem.

The basic bound for polynomial regression is that one can achieve good error with sample complexity and run-time that are $O(n^d)$. This dependence is also necessary (the space of degree d polynomials is of dimension $\approx n^d$) even when $y = P(x)$. But often, such high complexity either in run-time or sample requirements is not feasible for many applications.

This begs the question: can we formulate natural and useful scenarios where one can beat n^d complexity? One such example is the work of [APVZ14] who study *sparse polynomials* and achieve complexity that is $f(d)\text{poly}(n, s)$ where s is sparsity (in a suitable basis).

Motivated by the rich body of work on *phase retrieval* (see, e.g., [CSV13,CLS15,CEHV15, NJS13] and references therein), work on *multi-index models* in learning (see Section 2.1.2 below) and the above broad question, we study the question of learning polynomials that depend on few relevant dimensions. We call such polynomials *low-rank polynomials.* We begin by restating their definition, as introduced in Definition 1.2.3, in slightly different notation:

Definition 2.1.1. *A degree d polynomial $P : \mathbb{R}^n \to \mathbb{R}$ is of rank r if there exists a degree d polynomial $p : \mathbb{R}^r \to \mathbb{R}$ and vectors $u_1^*, \ldots, u_r^* \in \mathbb{R}^n$ such that*

$$P(x) = p(\langle u_1^*, x\rangle, \langle u_2^*, x\rangle, \ldots, \langle u_r^*, x\rangle).$$

We will refer to p as the link polynomial *and $U^* \triangleq \mathrm{span}(u_1^*, \ldots, u_r^*)$ as the* hidden subspace*.*

In other words, even though the ambient dimension of the polynomial P is n, its *intrinsic dimension* is only r. If we knew the subspace spanned by $u_1^*, \ldots, u_r^*$, then we could learn P with sample-complexity that does not depend on n at all and run-time that is linear in n (and not n^d). Here, there are many natural notions of learning P one could consider. Arguably the two most important goals are 1) to recover the hidden subspace U^* spanned by $u_1^*, \ldots, u_r^*$, and 2) to find a polynomial q that is close to P.

Concretely, we are given samples (x, y) where $y = P(x)$ and P is a low-rank polynomial. For most natural distributions y, one can show it is information-theoretically possible to learn P with sample-complexity that is only $O_{d,r}(n)$. That is, the dependence on the ambient dimension is only linear. Can we achieve this goal *efficiently*? Henceforth, by efficient we mean that the sample-complexity and run-time are at most some fixed polynomial in n that is of the form $O(f(r, d)n^c)$ for universal constant c.

As desirable as the above goal is, it might be too good to be true for general distributions. For example, as mentioned in Section 1.2.1, if x is uniform on the hypercube $\{1, -1\}^n$, then the above question can encode the problem of learning *k-juntas.* There, we are given samples $(x, f(x))$ where $x \in_u \{\pm 1\}^n$ and f is a function of at most k variables, and the goal is to recover the indices of the relevant variables. Despite much attention, the best algorithms run in time $n^{\Omega(k)}$, and achieving $f(k)\mathrm{poly}(n)$ sample complexity is an outstanding challenge conjectured to be computationally hard [MOS03]. The connection to rank is that any k-junta is a polynomial of rank and degree at most k.

Nevertheless, it makes sense to ask the question for other natural distributions. The most basic question in this vein (as we will further motivate later) is the case when x is Gaussian:

Q1. *Given samples $(x, y = P(x))$ where $x \sim \mathcal{N}(0, Id_n)$, and P is an unknown degree-d, rank-r polynomial, can one approximately recover the subspace defining P efficiently? Can we efficiently approximate P? Further, what is the dependence on the error ε?*

Note that while we ask the question for isotropic Gaussian covariates, our guarantees immediately carry over to general Gaussians, because the space of low-rank polynomials is affine invariant. Before stating our results, we first briefly discuss different ways of looking at the above question.

Learning Multi-Index Models While we motivated the above problem from the context of polynomial regression, an equally valid way to introduce it is from the perspective of learning *multi-index models* in Gaussian space.

Recall from Definition 1.2.2 that here, we are given samples from a distribution (x, y) where $x \sim \mathcal{N}(0, \mathrm{Id}_n)$ and

$$y = g(\langle u_1^*, x\rangle, \langle u_2^*, x\rangle, \ldots, \langle u_r^*, x\rangle),$$

where $g : \mathbb{R}^r \to \mathbb{R}$ is some unknown *link function* and $u_1^*, u_2^*, \ldots, u_r^*$ are unknown orthonormal vectors, and the goal is to learn the subspace U^* spanned by $u_1^*, \ldots, u_r^*$.

The main question we study is the case where the unknown link function g is a low-degree polynomial. Most relevant to the present work is the recent work of [DH18] which we discuss next. There is a tremendous amount of work on learning multi-index models, and we refer to [DH18] for a detailed overview of previous work. [DH18] address the case where g is *smooth* in a Lipschitz sense quantified by a parameter R. They show:

1. For *single-index models* (i.e. when $r = 1$): an algorithm that takes $\tilde{O}(n^{O(R^2)}) + n/\varepsilon^2)$ samples and computes a direction u that is ε-close to the hidden direction.

2. For *multi-index models*: an algorithm that takes $\tilde{O}(n^{O(rR^2)}) + n/\varepsilon^2)$ samples and com-

putes a direction u that has at least $1-\varepsilon$ of its ℓ_2-mass in the span of the unknown $u_1^*, u_2^*, \ldots, u_r^*$.

Firstly, note that while most works on learning multi-index models assume some sort of Lipschitz-smoothness of the link function, polynomials are a natural class of link functions that do not satisfy such smoothness. More importanty, unlike existing works on multi-index models, our main goal is to achieve near-linear sample complexity, run-time scaling with n^c for c independent of r, d, and polylogarithmic dependence on the error ε.

Generalizing Phase Retrieval Further impetus for the above problem comes from the vast literature on phase retrieval. Here, one is given samples of the form $(x, \langle w, x\rangle^2)$ where x is typically Gaussian for most provable guarantees [CSV13, CLS15, CEHV15, NJS13], and the goal is to learn w. Besides being natural by itself, the problem is extremely important in practice: as is explained in the references above, in certain physical devices one only observes the *amplitudes* of linear measurements (corresponding to $\langle w, x\rangle^2$) and not the phase. In this setting, the signal and the inputs are taken to be complex but the question is often studied over the reals as well.

Note that the low-rank polynomial in question here is rank 1 and degree 2; moreover the link polynomial $p(z) = z^2$ is even known *a priori*. In this sense, the problem we consider in this chapter is a substantial generalization, the study of which could potentially lead to new insights for phase retrieval, especially over more general covariate distributions.

Connections to Tensor Decompositions Our work also broadly fits in the category of *tensor decompositions*. A k-ary tensor in n-dimensions is a multi-dimensional array $T \in \mathbb{R}^{[n]^k}$. More relevant to the present work, one can also view a tensor T as a multi-linear map from $T : (\mathbb{R}^n)^k \to \mathbb{R}$ as $T(x^1, x^2, \ldots, x^k) = \sum_{1 \le i_1 \le i_2 \le \cdots \le i_k \le n} T[i_1, i_2, \ldots, i_k] x^1_{i_1} x^2_{i_2} \ldots x^k_{i_k}$. For tensors, the term "rank" has a different meaning: a rank 1 tensor is a tensor of the form $v^1 \otimes \cdots \otimes v^k$, and in general, the rank of a tensor T is the least number of rank one tensors whose sum is T.

The basic problem in tensor decomposition is to find a low-rank decomposition of a given tensor. Tensor decomposition algorithms have received a lot of attention recently [AGJ14, AGJ15, GM15, HSS15, HSSS16, SS17, MSS16] with various works studying many different aspects. The connection to our polynomial learning problem comes from the fact that a degree d polynomial can be viewed as a d-ary tensor. Moreover, if a polynomial has rank r, then the corresponding d-ary tensor has rank roughly $O(rd)$.

However, our goals and setting are quite different from those studied in the literature. For one, we are not given access to the tensor directly but only implicitly in the form of evaluations of the symmetric multi-linear form of the tensor on random inputs. Secondly, the central goal for us is to exploit the implicit representation to run in time that is much less than the time to even store the corresponding d-ary tensor. As far as we can tell, existing methods for tensor decompositions do not have these properties, at least provably. It is an intriguing question to find further scenarios where one could find tensor decompositions with much better run-time, for instance for constant-rank tensors, when the tensor has a *succinct implicit representation.*

2.1.1 Main Result

Our main result is that we can indeed efficiently learn low-rank polynomials in Gaussian space. To the best of our knowledge, no such results were known even for the rank-1 case. Before stating our result formally, we have to introduce a definition to deal with *degeneracy* in the notion of low-rank.

To understand the issue, consider the example where the link polynomial $p(z_1, z_2) = z_1 + z_2$. Then, if we look at $P(x) = p(\langle w_1^*, x\rangle, \langle w_2^*, x\rangle)$, even though the polynomial is represented as a rank two polynomial, it is really only of rank one and we cannot hope to recover the span of w_1^*, w_2^* but only the span of $w_1^* + w_2^*$. The following is necessary to overcome such non-identifiability issues:

Definition 2.1.2. *(Informal; see Definition 2.3.1) A polynomial P is α-non-degenerate rank r if P is of rank r and for any $(r-1)$-dimensional subspace H, the conditional variance of $P(x)$ given the projection of x onto H is at least $\alpha \cdot Var(p)$.*

Intuitively, there should not be a $(r-1)$-dimensional space that captures all of the variance of P. We give an equivalent analytic definition in Section 2.3. Note that any rank-1 polynomial satisfies the condition with $\alpha = 1$.

Theorem 2.1.3. *There exists a universal constant c_0 and for all r, d, α, there exists $C_0(r, d, \alpha)$ such that the following holds. For all $\delta > 0$ and $\varepsilon \in (0, 1)$, there is an efficient algorithm that takes $N = C_0(r, d, \alpha)(\log(n/\delta))^{c_0 d} \cdot n \log^2(1/\varepsilon)$ samples $(x, P(x))$, where $x \sim \mathcal{N}(0, Id_n)$ and P is an unknown α-non-degenerate rank r, degree-d polynomial defined by hidden subspace U^*, and outputs*

1. *Orthonormal $u_1, \ldots, u_r \in \mathbb{S}^{n-1}$ such that $d_P(\mathrm{span}(u_1, \ldots, u_r), U^*) \leq \varepsilon$*

2. *Degree d, r-variate polynomial g such that $\mathbb{E}[(y - g(\langle u_1, x\rangle, \ldots, \langle u_r, x\rangle))^2] \leq \varepsilon \cdot Var(y)$.*

The run-time of the algorithm is at most $\widetilde{O}(r^{c_0 d} N \cdot n)$.

This will follow from Theorem 2.2.2 and Theorem 2.5.1 later in the paper. Here, $d_P(U, U^*)$ denotes the *Procrustes distance* which is one of the standard measures for quantifying distances between subspaces. See Definition 1.3.2 for the exact definition.

Note that the run-time of the algorithm is essentially $O_{r,d}(n^2(\log n)^{O(d)})$ — a fixed polynomial in n as desired. The sample complexity is also essentially linear in the ambient dimension n and poly-logarithmic in $1/\varepsilon$. No such result was known even for the rank 1 case.

Remark 2.1.4. *A word about the constant $C_0(r, d, \alpha)$ in the theorem. Our proof involves a compactness argument and as a result does not give an explicit upper bound on this quantity. Bounding this comes down to an extremal problem for low-degree polynomials in r variables. For instance for $r = 1$, $C_0(1, d, 1)$ is essentially the inverse of*

$$\sup_{\tau} \inf_{h} (\mathbb{E}[1(|p(g)| > \tau)(g^2 - 1)]),$$

where $g \sim \mathcal{N}(0, 1)$ and the infimum is over degree d polynomials of variance 1. We believe that this quantity is at least 2^{-Cd^2} (as achieved by a suitably scaled degree d Chebyshev polynomial). In general, our arguments can potentially yield a bound of $C(r, d, \alpha) \approx 2^{O(rd)^2)}/\alpha^{\Theta(1)}$.

Also, we study the noiseless case where $Y = P(X)$. It is possible to modify the first part of our argument (Theorem 2.2.2) to get a version tolerant to some noise in Y, but we do not focus on this here. In any case, one of our main technical emphases is on getting *run-time and sample complexity scaling with* $\mathrm{poly}(\log(1/\varepsilon))$*, which would not be possible in the presence of noise.*

2.1.2 Related Work

Filtering Data by Thresholding Our algorithm for obtaining a warm start (see Theorem 2.2.2) relies on filtering the data via some form of thresholding. This general paradigm has been used in other, unrelated contexts like robustness, see [SS19, SS18, DKK+19a, Li18b, DKK+19b, DKK+17] and the references therein, though typically the points which are *bigger* than some threshold are removed, whereas our algorithm, FilteredPCAv1, is an intriguing case where the opposite kind of filter is applied.

Riemannian Optimization It is beyond the scope of this paper to reliably survey the vast literature on Riemannian optimization methods, and we refer the reader to the standard references on the subject [Udr94, AMS09] which mostly provide asymptotic convergence guarantees, as well as the thesis of Boumal [Bou14] and the references therein. Some notable lines of work include optimization with respect to orthogonality constraints [EAS98], applications to low-rank matrix and tensor completion [MMBS13, Van13, IAVHDL11, KSV14], dictionary learning [SQW16], independent component analysis [SJG09], canonical correlation analysis [LWW15], matrix equation solving [VV10], complexity theory and operator scaling [AZGL+18], subspace tracking [BNR10, ZB16], and building a theory of geodesically convex optimization [ZS16, HS15, ZRS16].

We remark that the update rule we use in our boosting algorithm is very similar to that of [BNR10, ZB16], as their and our work are based on geodesics on the Grassmannian manifold. That said, they solve a very different problem from ours, and the analysis is quite different.

Single/Multi-Index Models and Other Link Functions As mentioned above, the problem of learning low-rank polynomial is a special case of that of learning a multi-index model, for which there is also a large literature which we cannot hope to cover here. In addition to [DH18] other works include those based on a connection to Stein's lemma [PVY17, NWL16, Bri12, Li92, PV16, YBL17], sliced inverse regression [BB^{+}18] as introduced in [Li91], and gradient-based estimators [HJS01, HJP^{+}01, DJS08]. Other works consider specific link functions or families of link functions:

- $z \mapsto \mathrm{sgn}(z)$, i.e. one-bit compressed sensing [PV13, ALPV12, GNJN13].
- $z \mapsto |z|^2$, i.e. phase retrieval [CSV13, CLS15, CEHV15, NJS13].
- $z \mapsto F(z)$ where $F : \mathbb{R}^r \to \mathbb{R}$ is computable by a constant-layer neural network [GLM17, BJW18, JSA15, GKLW18, GKKT17, GK19].
- $z \mapsto \mathbb{1}[\varepsilon_i \cdot \mathrm{sgn}(z_i) \ \forall \ i \in [r]]$ for signs $\varepsilon \in \{\pm 1\}^r$, i.e. intersections of halfspaces [Vem10b, KLT09, KOS04, KS08, Vem10a, DKS18a].
- $z \mapsto F(z)$ for some function $F : \mathbb{R}^r \to \{0,1\}$, i.e. subspace juntas [VX11, DMN19].

That said, none of the above seem to imply the guarantees for learning low-rank polynomials that we want, namely a run-time that is a fixed polynomial in n and poly-logarithmic in $1/\varepsilon$.

2.2 Outline of Algorithm and Analysis

A natural first step is to try to adapt the various techniques from the phase retrieval literature or existing works on multi-index models to the problem. But this seems challenging even for rank 1. For example, the phase retrieval problem corresponds to the polynomial $p(z) = z^2$, which is rather special (see below), and if we don't even know the polynomial, then there are further difficulties. The works on multi-index models such as [DH18] also seem to be difficult to apply off the shelf. For one, they require smoothness of the link function. While it may be possible to circumvent the strict smoothness condition, it seems hard to find useful notions where the smoothness would not grow with the degree, leading to inefficient algorithms.

We present a different line of attack, inspired by ideas of [DH18], [CLS15], [BNR10]. Let $P(x) = p(\langle u_1^*, x\rangle, \langle u_2^*, x\rangle, \ldots, \langle u_r^*, x\rangle)$ be the unknown α-non-degenerate rank r polynomial. For the remainder of the paper, let $\mathcal{D}$ denote the distribution (x, y) where $x \sim \mathcal{N}(0, \mathrm{Id}_n)$ and

$y = P(x)$. Let $U^* = \mathrm{span}(u_1^*, \ldots, u_r^*)$ be the hidden subspace. Without loss of generality assume $\mathbb{V}(y) = 1$.[1]

Our approach has two modular steps:

1. **Warm start**: Obtain a "good" approximation to the true subspace U^* by a modified PCA.

2. **Boost accuracy**: Use the subspace computed above as a starting point to boost the accuracy by *Riemannian stochastic gradient descent*.

We next explain the steps at a high-level. The methods to carry out each of the steps could potentially be useful elsewhere especially for problems dealing with subspace recovery.

2.2.1 Getting a Warm Start

The first step is to find a good subspace V of dimension r that ε-close to U^* (i.e., $d_P(V, U^*) \leq \varepsilon$) in $O_{r,d}(n/\varepsilon^2)$ samples. Note that identifying the subspace U^* is the best we can do as the individual directions are not uniquely identifiable.

Rank-One Case: To motivate the algorithm, let us first focus on the rank 1 or single-index case. Here $P(x) = p(\langle u^*, x\rangle)$ where $u^* \in \mathbb{S}^{n-1}$ and our goal is to find some $u \in \mathbb{S}^{n-1}$ close to u^*.

To do so, we propose a modified PCA by estimating a matrix of the form $M^\phi \equiv E[\phi(y)xx^T] - E[\phi(y)]E[xx^T]$ where $\phi : \mathbb{R} \to \mathbb{R}$ is a suitable "filtering" function. The intuition behind looking at M^ϕ is that the matrix has kernel of dimension $n-1$ corresponding to directions orthogonal to u^*. Thus, the non-zero eigenvalue of M^ϕ, if any, could help us approximate or even identify u^*.

But what should the function ϕ be? For example, for phase retrieval where $P(x) = \langle u^*, x\rangle^2$, taking $\phi(z) = z^2$ suffices. The key issue is that this choice of ϕ does not work for general link polynomials. For example, if the link polynomial p is $p(z) = z^2 - 3$, the matrix M^ϕ for this particular choice of ϕ is identically zero.

[1] We can do so as our algorithms only need a good lower and upper bound on the variance y which can be obtained easily.

We propose overcoming this by instead applying a simple *thresholding* filter for ϕ. Specifically, for a parameter $\tau > 0$ to be chosen later, let

$$M^\tau \triangleq \mathbb{E}[1(|y| > \tau)(xx^T - I)].$$

We show that for all d there exists $\tau \equiv \tau(d)$ that only depends on d such that M^τ is a non-zero matrix. Note that this by itself is not enough for our purposes: if the least non-zero eigenvalue of M^τ were extremely small, then this would affect our sample complexity in estimating M^τ. We show there exists τ such that M^τ has an eigenvalue with magnitude at least $\lambda_d > 0$ for some constant depending on d only. As argued before, the corresponding eigenvector is u^*. The intuition behind the proof is that conditioning on $|y| > \tau$ makes x more likely to be large in the relevant direction.

The above structural statement is enough to get a warm start for u^* by looking at the empirical approximation of M^τ: for N samples, let

$$\widehat{M}^\tau \triangleq \frac{1}{N}\sum_{i=1}^{N} 1(|y_i| > \tau)(x_i x_i^T - I).$$

We can now use standard matrix concentration inequalities to argue that for $N = O_d(n/\varepsilon^2)$ samples, the top eigenvector $\hat{u}$ of $\widehat{M}^\tau$ satisfies $\|u^* - \hat{u}\| \leq \varepsilon$.

Remark 2.2.1 (Relation to Sliced Inverse Regression)**.** *The trick of conditioning only on (x, y) for which $|y|$ is sufficiently large is reminiscent of the technique of* slicing *originally introduced by [Li91] in the context of learning multi-index models. The high-level idea of slicing is that for any fixed value of y, the conditional law of $x|F(x) = y$ is likely to be non-Gaussian in most directions $v \in V$, so in particular, $\mathbb{E}[xx^\top - Id|F(x) = y]$ should be nonzero, and its singular vectors will lie in V. This can be thought of as filtered PCA with the choice of function $\psi(z) = \mathbb{1}[z = y]$. The first issue with using such an approach to get an actual learning algorithm is that $\Pr_x[F(x) = y] = 0$ for any y, and the workaround in non-asymptotic analyses of sliced inverse regression [BB+18] is to estimate something like $\mathbb{E}_y[\mathbb{E}[xx^\top - Id|F(x) = y]]$ instead. While finite sample estimators for such objects are known, the conditions under which this approach can provably recover the relevant subspace are quite* *strong and not applicable to our setting.*

Higher-Rank Case: Extending the above to higher ranks seems much more challenging.

A natural attempt would be to look at a matrix M^τ as above for a suitable τ. It is once again easy to argue that M^τ has $n - r$ vectors in its kernel corresponding to the vectors orthogonal to U^*. We would now like to say that for some suitable $\tau \equiv \tau(r, d)$, the top r eigenvalues of M^τ are at least $\lambda_{r,d}$. If so, we can proceed as before to get an approximation to U^* (the non-zero eigenvectors are in U^*). While we can currently show that there is at least one such eigenvalue, we do not know if the matrix M^τ has rank at least r and it seems considerably more challenging to prove. The difficulty is that unlike the rank 1 case, while conditioning on $|y| > \tau$ should intuitively bias x to have large norm in the relevant directions, it is not clear if it does so in every relevant direction.

Instead, we follow an iterative strategy where we identify one direction at a time in U^*. This is similar in spirit to the standard technique of computing the eigenvalues of a matrix by first computing the top eigenvector, projecting it out, and then iterating.

Concretely, suppose we have identified orthonormal vectors $V = \{v_1, v_2, \dots, v_\ell\}$ for $\ell < r$ that individually have most of their mass in U^*. Let $\Pi_{V^\perp}$ be the projection operator onto the space orthogonal to $v_1, \dots, v_\ell$. Then, to compute the next direction we look at the top eigenvector of

$$M^{\ell,\tau} \triangleq \Pi_{V^\perp}\mathbb{E}[1(|y| > \tau)1(|\langle v_i, x\rangle| \leq 1,\ \forall i \leq \ell)(xx^T - I)]\Pi_{V^\perp}.$$

As before, we argue that the top eigenvector of the above matrix will have most of its mass in U^* and this gives us our next vector $v_{\ell+1}$. While the sequence of matrices we look at are a bit more complicated, standard random matrix concentration inequalities still allow us to identify the new directions with sample complexity $O_{r,d}(n)$.

In summary, we get the following:

Theorem 2.2.2. *For all r, d, α, there exists $C(r, d, \alpha)$ such that the following holds. For all $\delta > 0$ and $\varepsilon \in (0, 1)$, there is an efficient algorithm that takes $N = C(r, d, \alpha) n \log(1/\delta)/\varepsilon^2$ samples $(x, P(x))$ for $x \sim \mathcal{N}(0, Id_n)$ and unknown P which is α-non-degenerate of rank r,* *and outputs a subspace U such that with probability at least $1-\delta$, $d_P(U, U^*) < \varepsilon$. The algorithm runs in time $O(r(Nn^2 + n^3))$.*

2.2.2 Boosting via Geodesic-Based Riemannian Gradient Descent

The results from the previous section give us a way to find a subspace U that is ε-close to the true subspace U^* with sample complexity $O_{r,d}(n/\varepsilon^2)$.

However, the dependence on ε above is problematic and quite far from what is achievable, e.g., for the special case of phase retrieval. There, results starting with work of [CLS15] show that one can get *exact* recovery of the unknown direction with sample complexity $\tilde{O}(n)$; in this case, while the sample complexity is $\tilde{O}(n)$, the *run-time* to get within error ε scales with $\log(1/\varepsilon)$). In a similar vein, the result of [NJS13] shows that one can find a vector w that is ε-close to the unknown vector with sample-complexity $\tilde{O}(n\log(1/\varepsilon))$ and a similar run-time. We address this issue next and give an algorithm that achieves error ε with sample-complexity $\tilde{O}_{r,d}(n\log^2(1/\varepsilon))$ and run-time $\tilde{O}_{r,d}(n^2\log^2(1/\varepsilon))$. In the proceeding discussion, we will use some basic terminology from differential geometry in motivating our algorithm, though we emphasize that the algorithm itself is stated solely in terms of matrices, and its proof only involves, e.g., linear algebra and concentration of measure.

First, it is important to understand what fundamentally changes when going from phase retrieval to the more general problem of learning an unknown, low-rank polynomial. At a high level, there are two closely related challenges:

1. **Unknown r-variate polynomial**: Unlike in phase retrieval where we know that the link polynomial is $h(z) = z^2$ *a priori*, in our setting we are not given the coefficients of the true polynomial. The natural workaround is to simply run gradient descent jointly on the space of coefficients and the space of $n \times r$ matrices V. As we will see in Section 2.2.2 next, this poses novel difficulties even in the rank-1 case.

2. **Identifiability only up to rotation**: A more fundamental issue is the number of inherent symmetries in the problem, which explodes as r increases. Indeed, there is an infinitely large orbit of parameters $\Theta^* = (\mathbf{c}^*, V^*)$ which give rise to the same underlying low-rank polynomial P, parametrized by the group of all rotations of the underlying

subspace. Whereas for $r = 1$ it is easy to quotient out most of the symmetries by simply running projected gradient descent on the unit sphere, as we will see in Section 2.2.2, to define the right quotient geometry we will need to run gradient descent on a manifold for which the corresponding optimization landscape is far less straightforward. In addition, as we will see in Section 2.2.2, these symmetries also pose problems for defining and analyzing a suitable progress measure.

In light of 2), it will be good to give a name to the set of parameters $\Theta^* = (\mathbf{c}^*, V^*)$ which correspond to the underlying low-rank polynomial.

Definition 2.2.3. *For a collection of coefficients $\mathbf{c}^*$ of a degree-d r-variate polynomial, and a column-orthonormal matrix $V^* \in \mathbb{R}^{n\times r}$, we say that the parameters $\Theta^* = (\mathbf{c}^*, V^*)$ are a* realization *of $\mathcal{D}$ if the polynomial $p_*(z) \triangleq \sum_I c^*_I \phi_I(z)$ satisfies $P(x) = p_*(V^{*\top} x)$ for all $x \in \mathbb{R}^n$, where $\{\phi_I\}$ are the (normalized) tensor-product Hermite polynomials of degree at most d over r variables (see Section 2.3.3).*

Not Knowing the Polynomial: A Toy Calculation

The issue of not knowing p manifests even in the $r = 1$ case. Below, we examine at a high level where the calculations for analyzing gradient descent for phase retrieval break down for us.

Let us try to imitate the approach of [CLS15]. Let $\Theta^* = (\mathbf{c}^*, v^*)$ be one of the two possible realizations of $\mathcal{D}$ for which $v^* \in \mathbb{S}^{n-1}$, and suppose we already have a warm start of $\Theta = (\mathbf{c}, v)$, where the coefficients $\mathbf{c}$ and $\mathbf{c}^*$ define the univariate degree-d polynomials $p(z) \triangleq \sum_{i=1}^d c_i\phi_i(z)$ and $p_*(z) \triangleq \sum_{i=1}^d c^*_i\phi_i(z)$ respectively. Given samples $(x^1, y^1), ..., (x^N, y^N) \sim \mathcal{D}$, a natural approach would be to analyze vanilla gradient descent over $\mathbb{R}^{d+1} \times \mathbb{R}^n$ for the empirical risk

$$L(\Theta) \triangleq \frac{1}{N}\sum_{i=1}^N (F_{x^i}(\Theta) - y_i)^2 \qquad \text{for} \qquad F_x(\Theta) \triangleq p(V^\top x).$$

To show that this converges linearly from a warm start, the first thing to show would be that the negative gradient at Θ is correlated with the direction in which we would like to move, a property that sometimes goes under the name *local curvature.* Noting that

$\frac{1}{2}\nabla L(\Theta) = \frac{1}{N}\sum_{i=1}^N (F_{x^i}(\Theta) - F_{x^i}(\Theta^*)) \cdot \nabla F(x^i)(\Theta)$, using the fact that we initialize at a warm start in order to linearly approximate $F_x(\Theta) - F_x(\Theta^*)$ by $\nabla F_x(\Theta^*) \cdot \langle \Theta - \Theta^* \rangle$, and explicitly computing the gradient of F_x (see Proposition 2.5.3), one can check that

$$\left\langle \frac{1}{2}\nabla L(\Theta), \Theta - \Theta^* \right\rangle \approx \frac{1}{N}\sum_{i=1}^N \langle \nabla F_{x^i}(\Theta^*), \Theta - \Theta^* \rangle^2$$
$$= \frac{1}{N}\sum_{i=1}^N \left[\langle v - v^*, x^i\rangle \cdot p'_*(\langle v^*, x^i\rangle) + (p - p_*)(\langle v^*, x^i\rangle)\right]^2.$$

The expectation of this quantity is

$$\mu \triangleq \mathbb{E}_g\left[(\langle v - v^*, g\rangle \cdot p'_*(\langle v^*, g\rangle) + (p - p_*)(\langle v^*, g\rangle))^2\right]$$

Write $v - v^* = \alpha \cdot v^* + \beta \cdot v^\perp$ for $v^\perp \in \mathbb{S}^{n-1}$ orthogonal to v^*, where $\alpha = \langle v, v^*\rangle - 1 \approx -\|v - v^*\|_2^2$. By some elementary calculations which we omit here, one can show that

$$\mu = \beta^2 \cdot \mathbb{E}[p'_*(x)^2] + \sum_{\ell=0}^d \left((\alpha\ell + 1) \cdot c_\ell + a\sqrt{(\ell+1)(\ell+2)} \cdot c_{\ell+2} - c^*_\ell\right)^2. \tag{2.1}$$

In the case of phase retrieval, $p(z) = p_*(z) = z^2 = \sqrt{2} + \sqrt{2} \cdot \phi_2(z)$, so $\mathbf{c} = \mathbf{c}^* = (\sqrt{2}, 0, \sqrt{2})$ and we simply get that

$$\mu = 12\alpha^2 + 4\beta^2 \geq 4\|v - v^*\|_2^2.$$

In other words, the correlation between the negative gradient and the residual direction $v^* - v$ in which we would like to go is positive and scales with the squared norm of the residual. This simple calculation lies at the heart of the proof that vanilla gradient descent converges linearly to v^* from a warm start for phrase retrieval.

More generally, if $\mathbf{c}^* = \mathbf{c}$, then the quantity in (2.1) will enjoy this positive scaling with $\|v^* - v\|_2^2$, and one can also show linear convergence of vanilla gradient descent. But it is apparent that when $\mathbf{c}^* \neq \mathbf{c}$, μ can be arbitrarily close to zero, e.g. by taking β to be much smaller than α. So when $\mathbf{c}^* \neq \mathbf{c}$, we may get stuck at spurious infinitesimal-curvature points of the optimization landscape and fail to make sufficient progress in a single step.

The basic underlying issue is simply that vanilla gradient steps can move us in unhelpful directions, e.g. we might end up moving mostly in the direction of v when we should be moving in directions orthogonal to v. And whereas this evidently does not pose an issue when $\mathbf{c} = \mathbf{c}^*$, which corresponds to the case where we know the underlying polynomial and only need to run gradient descent to learn the hidden direction, in the case where $\mathbf{c} \neq \mathbf{c}^*$ and we must run gradient descent jointly on v and $\mathbf{c}$, the usual analysis of vanilla gradient descent fails.

Non-Identifiability: Which Space to Run SGD In?

The workaround for the issue posed in Section 2.2.2 is clear at least in the rank-1 case: to avoid moving in the wasteful directions which are orthogonal to the current iterate v, simply compute the vanilla gradient and project to the orthogonal complement of v. We would also like to ensure that our iterates themselves are unit vectors like v^*, so the following two-step update rule would suffice: 1) walk against the projected gradient and then 2) project back to $\mathbb{S}^{n-1}$. In fact, one can show that this algorithm actually achieves linear convergence for learning arbitrary unknown rank-1 polynomials.

It turns out there is a principled way to extend this approach to higher rank. Indeed, the above mentioned projected gradient scheme is nothing more than (retraction-based) gradient descent on the Riemannian manifold $\mathbb{S}^{n-1}$: the orthogonal complement of v is precisely the tangent space of $\mathbb{S}^{n-1}$ at v, and the projection back to $\mathbb{S}^{n-1}$ is a special instance of a *retraction*, roughly speaking a continuous mapping from the tangent spaces of a manifold back onto the manifold itself. We do not attempt to define these notions formally, referring the reader to, e.g. [AMS09].

The rank-r analogue of $\mathbb{S}^{n-1}$ is the Grassmannian $\mathrm{G}(n, r)$ of r-dimensional subspaces of $\mathbb{R}^n$. However, while various retraction operations, e.g. via QR decomposition, can be constructed, retraction-based Riemannian optimization is somewhat more difficult to analyze in our setting. Instead, we appeal to an alternative formulation of Riemannian gradient descent via geodesics.

Roughly, geodesics are acceleration-free curves on a manifold determined solely by their initial position on the manifold, initial velocity, and length. Gradient descent on a Riemannian manifold $\mathcal{M}$ via geodesics is then very simple to formulate: at an iterate $p \in \mathcal{M}$, 1) compute the gradient ∇ after projecting to the tangent space at p, 2) walk along the geodesic that starts at p and has initial velocity ∇ and length η, where η is the learning rate.

We now see what this would yield in our setting. Let $\Theta = (\mathbf{c}, V)$ be an iterate. For now, we will keep $\mathbf{c}$ fixed and describe how to update V, regarded as a column-orthonormal $n \times r$ matrix of basis vectors for the subspace V, by following the appropriate geodesic on $\mathrm{G}(n, r)$. Given a single sample (x, y), define the single-sample empirical risk $L^{\mathbf{c}}_x(V) = (F_x(\Theta) - y)^2$. Let $\nabla L^{\mathbf{c}}_x(V) \in \mathbb{R}^{n\times r}$ be the vanilla gradient, where $L^{\mathbf{c}}_x(V) \triangleq L_x(\Theta)$. It turns out its projection to the tangent space at V is simply $\nabla \triangleq \Pi^{\perp}_V \cdot \nabla L^{\mathbf{c}}_x(V) \in \mathbb{R}^{n\times r}$, where $\Pi^{\perp}_V$ denotes projection to the orthogonal complement of V (note that this is a natural generalization of the tangent spaces for $\mathbb{S}^{n-1}$).

The geodesic Γ with initial point V and velocity ∇, and length η has a simple closed form in terms of the SVD of ∇, which is made even simpler by the fact that in our setting, ∇ turns out to be rank-1. We defer the details of the exact update, which can be computed in time $O(n)$, to Section 2.5.

Tracking Progress in both c and V

In the previous section we sketched our approach for updating our estimate V for the subspace given an estimate $\mathbf{c}$ for the coefficients of the polynomial, but did not explain how to update $\mathbf{c}$. As $\mathbf{c}$ just lives in Euclidean space, we can simply update $\mathbf{c}$ to some $\mathbf{c}'$ via vanilla gradient descent on L^V, where $L^V(\mathbf{c}) \triangleq L(\Theta)$, and this is the approach we take.

To analyze such an approach, one would want to show that each step $(\mathbf{c}, V) \mapsto (\mathbf{c}', V')$ contracts some suitably defined progress measure. Indeed, the natural progress measure one could try analyzing is

$$\inf_{(\mathbf{c}^*, V^*) \text{ realizing } \mathcal{D}} \|\mathbf{c} - \mathbf{c}^*\|_2^2 + \|V - V^*\|_F^2. \tag{2.2}$$

The key difficulty here is that the minimizing realization $(\mathbf{c}^*, V^*)$ could change with each new iterate, and tracking how this changes is tricky as there is no clean non-variational proxy for (2.2).

Our workaround is to have our boosting algorithm alternate between two phases. For an iterate $V \in \mathbb{R}^{n\times r}$, we run the following algorithm, GeoSGD, which alternates between two phases: 1) recomputing a good $\mathbf{c}$, and 2) updating V using that $\mathbf{c}$. An informal specification of this algorithm is given in Algorithm 3.

```
Algorithm 3: GeoSGD (informal)
Input: Sample access to $\mathcal{D}$, warm start $V^{(0)} \in \mathbb{R}^{n\times d}$, target error $\varepsilon$, failure
       probability $\delta$
Output: Estimate $(\mathbf{c}^{(T)}, V^{(T)})$ which is $\varepsilon$-close to a realization of $\mathcal{D}$
for $0 \le t < T$ do
    Run RealignPolynomial using $V^{(t)}$. That is, draw samples and run vanilla
      gradient descent with respect to empirical risk $L^{V^{(t)}}$ over those samples to
      produce $\mathbf{c}^{(t)}$ which approximates the "best" choice of $\mathbf{c}$ given fixed $V^{(t)}$.
    Run SubspaceDescent initialized to $V^{(t)}$ and using $\mathbf{c}^{(t)}$. That is, draw samples
      and, starting from $V^{(t)}$, run a small step of geodesic gradient descent with
      respect to empirical risk $L^{\mathbf{c}}_x$ for each of those samples $x$. Call the result $V^{(t+1)}$
return $V^{(T)}$.
```

We will defer an exact specification of GeoSGD and the subroutines RealignPolynomial and SubspaceDescent until Section 2.5.

To analyze this scheme, rather than track progress in (2.2) we can simply track progress in $d_P(V, V^*) = \inf_{V^*} \|V - V^*\|_F^2$, where V^* ranges over $n \times r$ matrices whose columns form an orthonormal basis for the true subspace. This progress measure is, up to constants, simply the Procrustes distance between our current subspace V and the true subspace V^*, and can be approximated by the *chordal distance* which has a simple closed-form expression amenable to analysis.

Roughly, we will show the following:

Theorem 2.2.4 (Informal, see Theorem 2.6.1)**.** *If V is sufficiently close to the true subspace in Procustes distance, then running* RealignPolynomial *using V will yield $\mathbf{c}$ such that for the realization $(\mathbf{c}^*, V^*)$ of $\mathcal{D}$ where $d_P(V, V^*) = \|V - V^*\|_F$, $\|\mathbf{c} - \mathbf{c}^*\|_2 \approx d_P(V - V^*)$.*

Theorem 2.2.5 (Informal, see Theorem 2.7.1)**.** *If V is sufficiently close to the true subspace in Procustes distance, If V and $\mathbf{c}$ are such that $\|\mathbf{c} - \mathbf{c}^*\|_2 \approx d_P(V - V^*)$for the realization $(\mathbf{c}^*, V^*)$ of $\mathcal{D}$ where where $d_P(V, V^*) = \|V - V^*\|_F$, then running* SubspaceDescent *initialized to V and using $\mathbf{c}$ will yield V' so that the progress measure $d_P(V, V^*)$ contracts by a factor of $1 - \tilde{O}_{r,d}(1/n)$.*

Having defined the “right” gradient descent subroutines, the proofs of Theorems 2.2.4 and 2.2.5 will be based on showing the same kind of estimates alluded to in Section 2.2.2. That is, for instance we must show that the steps in both subroutines have good correlation with the direction in which we want to go. Showing this holds with high probability will then entail exhibiting the appropriate second moment bounds. In the case of Theorem 2.2.4, we can then invoke standard hypercontractivity-based tail bounds to show concentration. In the case of Theorem 2.2.5, concentration will be more delicate as each small step of SubspaceDescent will be a geodesic gradient step with respect to a *single-sample* empirical risk $L^{\mathbf{c}}_x$. For the analysis to be doable, it is crucial that these risks be single-sample so that the geodesic steps are *rank-one* updates. But then, to show concentration over a sequence of small geodesic steps, we must invoke non-standard martingale concentration inequalities, see Section 2.3.2. Intuitively, if we take the sizes of these small steps to scale with $O(1/T)$, the corresponding martingale does not move away from its starting point by too much, and the sum of the martingale differences ends up behaving more or less like a sum of iid random variables (see the beginning of Section 2.7.2). We refer the reader to Sections 2.6 and 2.7 for the complete proofs of Theorems 2.2.4 and 2.2.5 respectively.

Roadmap In Section 2.3 we introduce notation and miscellaneous technical facts that we will use in our proofs. In Section 2.4, we give our algorithm FilteredPCAv1 for obtaining a warm start. In Section 2.5, we give the formal specification for our boosting algorithm GeoSGD, and in Sections 2.6 and 2.7 we prove guarantees for its key subroutines. We complete the proof of correctness for GeoSGD in Section 2.8. In Appendix 2.9 we give the martingale concentration inequalities we will need, and in Appendix 2.10 and Appendix 2.11 we complete proofs deferred from the body of the paper.

2.3 Technical Preliminaries

Notation Throughout this chapter of the thesis, n will denote the ambient dimension, r the rank of the polynomial, and d the degree.

For polynomial $p : \mathbb{R}^r \to \mathbb{R}$, define $\mathbb{V}[p] = \mathbb{E}[(p - \mathbb{E}[p])^2]$. Given indices $\mathbf{j} \triangleq (j_1, ..., j_\ell) \in$

$[r]^\ell$, and $z \in \mathbb{R}^r$ we will use the shorthand

$$\mathrm{D}_{\mathbf{j}}\, p(z) \triangleq \frac{\partial}{\partial z_{j_1} \cdots \partial z_{j_\ell}} p(z). \tag{2.3}$$

Similarly, for $F : \mathbb{R}^{n\times r} \to \mathbb{R}$, indices $\mathbf{i} \in [n]^\ell$ and $\mathbf{j} \in [r]^\ell$, and $V \in \mathbb{R}^{n\times r}$, we will use the shorthand

$$\mathrm{D}_{\mathbf{i},\mathbf{j}}\, F(V) \triangleq \frac{\partial}{\partial V_{i_1,j_1} \cdots \partial V_{i_\ell,j_\ell}} F(V). \tag{2.4}$$

2.3.1 Non-degeneracy

Recall the notion of α-non-degenrate rank r polynomials introduced in Definition 2.1.2. While that notion is intuitive, it is less amenable to analysis. It turns out that the notion is essentially equivalent (up to scaling α by d) to the following and we will use this going forward.

Definition 2.3.1. *A polynomial $h : \mathbb{R}^r \to \mathbb{R}$ is α non-degenerate if $M = \mathbb{E}_{g\sim\mathcal{N}(0,Id_r)}\left[\nabla h(g) \nabla h(g)^\top\right]$ satisfies $M \succeq \alpha \cdot \|M\|_2\, Id_r$.*

We say a rank r polynomial $P : \mathbb{R}^n \to \mathbb{R}$ is α non-degenerate if P is non-degenerate in the r-dimensional space corresponding to the relevant directions. That is, there exist orthonormal vectors $u_1, \dots, u_r$ such that $P(x) = h(\langle u_1, x\rangle, \dots, \langle u_r, x\rangle)$ and h is α non-degenerate.

While it is not clear immediately from the definition, the notion above does not depend on the specific basis chosen. Henceforth, fix constant $\nu_{\mathsf{cond}} > 0$. we will let $\mathcal{P}^{\nu_{\mathsf{cond}}}_{n,r,d}$ denote the set of all ν_{cond} non-degenerate rank r polynomials P of degree at most d in n variables that satisfy the normalization conditions $\mathbb{E}_{X\sim\mathcal{N}(0,\mathrm{Id}_n)}[P(X)] = 0$ and $\mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{Id}_r)}\left[\nabla h(g) \nabla h(g)^\top\right] \preceq \mathrm{Id}_n$. We write $\mathcal{P}^{\nu_{\mathsf{cond}}}_{r,d}$ for $\mathcal{P}^{\nu_{\mathsf{cond}}}_{r,r,d}$.

Finally, we will use the following elementary property of non-degeneracy.

Fact 2.3.2. *If $P \in \mathcal{P}^{\nu_{\mathsf{cond}}}_{n,r,d}$, then $\nu_{\mathsf{cond}}/d \le \mathbb{V}[P(X)] \le r$.*

Proof. It suffices to consider $n = r$. For the upper bound, we have $\mathbb{V}[P] \le \mathbb{E}_g\left[\|\nabla p_*(g)\|_2^2\right] \le r$ by taking traces in the definition of non-degeneracy and invoking Lemma 2.3.9 below.

For the lower bound, we have $\mathbb{V}[P] \ge \mathbb{E}_g\left[\|\nabla p_*(g)\|_2^2\right]/rd \ge \nu_{\mathsf{cond}}/d$ by taking traces and invoking Lemma 2.3.8 below. □

2.3.2 Other Concentration Inequalities

Martingale Concentration Here we will generalize two concentration inequalities from Section 1.3 to the martingale setting. Let $\zeta_1, ..., \zeta_T$ be independent atom variables which each take values in Euclidean space. Let $Y(\zeta_1, ..., \zeta_T)$ be a real-valued random variable depending on the atom variables $\zeta_1, ..., \zeta_T$ which each take values in Euclidean space. Define the martingale differences $Z_i(\zeta) \triangleq \mathbb{E}[Y|\zeta_1, ..., \zeta_i] - \mathbb{E}[Y|\zeta_1, ..., \zeta_{i-1}]$. When the context is clear, we will suppress the parenthetical ζ. For brevity, we will use the acronym MDS throughout to refer to martingale difference sequences.

The first lemma is the martingale analogue of Lemma 1.3.16, with the slight twist that we only have high-probability moment bounds for the increments. The bounds are slightly weaker than those of Lemma 1.3.16 but will suffice for our applications.

Lemma 2.3.3. *There is a constant $c_1 > 0$ for which the following holds. Let $\sigma > 0$, and suppose the atom variables $\zeta_1, ..., \zeta_T$ are standard n-dimensional Gaussians, and suppose the martingale differences $\{Z_i\}$ are such that for any realization of $\zeta_1, ..., \zeta_{i-1}$, $Z_i(\zeta)$ is a polynomial of degree at most d in ζ_i, and moreover $Pr[\mathbb{E}[Z_i^2|\zeta_1, ..., \zeta_{i-1}] \le \sigma^2] \ge 1 - \beta$ for each $i \in [T]$. Then for any $t > 0$,*

$$Pr\left[\max_{\ell\in[T]}\left|\sum_{i=1}^{\ell} Z_i\right| \ge (2\log(1/\delta)\cdot d)^{c_1 d}\cdot\sqrt{T}\cdot\sigma\right] \le \delta + T\cdot\beta.$$

The second lemma is the martingale analogue of Lemma 1.3.33, again with the twist that the bounds on the differences only hold with high probability.

Lemma 2.3.4. *Let $\{c_i\}_{i\in[T]}$ and $\{s_i\}_{i\in[T]}$ be collections of positive constants, and let $\mathcal{E}_i$ be the event that $Z_i \le c_i$ and $\mathbb{E}[Z_i^2|\zeta_1, ..., \zeta_{i-1}] \le s_i^2$. Let $\sigma_i = c_i \vee s_i$, and define $\sigma^2 = \sum_i \sigma_i^2$. Then if $Pr[\mathcal{E}_i|\zeta_1, ..., \zeta_{i-1}] \ge 1 - \beta$ for each $i \in [T]$, then for any $\delta > 0$,*

$$Pr\left[\sum_{i=1}^{T} Z_i \ge \sqrt{2}\log(1/\delta)\cdot\sigma\right] \le \delta + T\cdot\beta.$$

2.3.3 Hermite Polynomials and Gradients

Recall the definition of the normalized Hermite polynomials in Section 1.3.6. We will need the following identities. Here we record some additional identities that they satisfy.

Fact 2.3.5 (Linearization Coefficients)**.** *For any $a, b, c \in \mathbf{Z}_{\geq 0}$ such that $a + b \geq c$, $a + c \geq b$, $b + c \geq a$, and $a + b + c$ is even.*

$$\mathbb{E}_{g \sim \mathcal{N}(0,1)} \left[\phi_a(g) \phi_b(g) \phi_c(g) \right] = \frac{\sqrt{a! \cdot b! \cdot c!}}{\left(\frac{a+b-c}{2}\right)! \cdot \left(\frac{a-b+c}{2}\right)! \cdot \left(\frac{-a+b+c}{2}\right)!}$$

For all other a, b, c, this quantity is zero.

Corollary 2.3.6. *For any $0 \leq a \leq b$,*

$$\mathbb{E}_{g \sim \mathcal{N}(0,1)}[g \cdot \phi_a(g) \phi_b(g)] = \mathbb{1}[b = a + 1] \cdot \sqrt{a+1}$$

$$\mathbb{E}_{g \sim \mathcal{N}(0,1)}[\phi_2(g) \phi_a(g) \phi_b(g)] = \mathbb{1}[b = a + 2] \cdot \sqrt{\frac{(a+1)(a+2)}{2}} + \mathbb{1}[b = a] \cdot a\sqrt{2}$$

We also record some basic facts about gradients and moments of polynomials in Gaussians, the first of which is a corollary of Gaussian hypercontractivity (Fact 1.3.15):

Corollary 2.3.7. *For any polynomial $p \in \mathbb{R}_d[x_1, ..., x_r]$, $\mathbf{j} = (j_1, ..., j_\ell) \in [r]^\ell$, and integer $q \geq 2$,*

$$\mathbb{E}_g[(\mathrm{D}_{\mathbf{j}}\, p(g))^q]^{1/q} \leq (q-1)^{d/2} \cdot d^{\ell/2} \cdot \mathbb{V}[p]^{1/2}$$

Proof. By Fact 1.3.15,

$$\mathbb{E}_g[(\mathrm{D}_{\mathbf{j}}\, p(g))^q]^{1/q} \leq (q-1)^{d/2} \mathbb{E}_g[(\mathrm{D}_{\mathbf{j}}\, p(g))^2]^{1/2}$$

Write $\mathrm{D}_{\mathbf{j}}\, p$ as $\frac{\partial^\ell}{\partial x_1^{a_1} \cdots \partial x_r^{a_r}} p$, where a_i is the number of entries of $\mathbf{j}$ equal to i, and write p in the tensored Hermite basis $p = \sum_I c_I \phi_I$. By Fact 2.5.3,

$$\mathrm{D}_{\mathbf{j}}\, p(x) = \frac{\partial^\ell}{\partial x_1^{a_1} \cdots \partial x_r^{a_r}} p(x) = \sum_I c_I \left(\prod_{i \in [r]} \phi_{I_i}^{[a_i]}(x_i) \right) = \sum_I c_I \left(\prod_{i \in [r]} \sqrt{\frac{I_i!}{(I_i - a_i)!}} \phi_{I_i - a_i}(x_i) \right),$$

so by orthogonality and the fact that $a_1 + \cdots + a_r = \ell$, we see that

$$\mathbb{E}_g[(\mathrm{D}_{\mathbf{j}}\, p(g))^2] = \sum_I c_I^2 \cdot \prod_{i\in[r]} \frac{I_i!}{(I_i - a_i)!} \leq \sum_{I\neq\emptyset} c_I^2 \cdot \prod_{i\in[r]} d^{a_i} = d^\ell \cdot \mathbb{V}[p],$$

from which the claim follows. □

We can use Corollary 2.3.7 to bound the moments of $\|\nabla p(g)\|_2^2$.

Lemma 2.3.8. *For any polynomial $p \in \mathbb{R}_d[x_1, ..., x_r]$ and any integer $q \geq 2$,*

$$\mathbb{E}[\|\nabla p(g)\|_2^{2q}]^{1/q} \leq rd \cdot (2q-1)^d \cdot \mathbb{V}[p]$$

Proof. We have

$$\mathbb{E}[\|\nabla p(g)\|_2^{2q}] \leq r^{q-1} \cdot \mathbb{E}[\|\nabla p(g)\|_{2q}^{2q}] = r^{q-1} \cdot \sum_{i=1}^r \mathbb{E}\left[\left(\frac{\partial}{\partial x_i} p(g)\right)^{2q}\right] \leq r^q \cdot (2q-1)^{dq} \cdot d^q \cdot \mathbb{V}[p]^q,$$

where the first inequality follows by Holder's, and the last step follows by Corollary 2.3.7. □

It will be useful to give a corresponding lower bound for $\mathbb{E}[\|\nabla p(g)\|_2^2]$:

Lemma 2.3.9. *For any polynomial $p \in \mathbb{R}_d[x_1, ..., x_r]$, $\mathbb{E}_g[\|\nabla p(g)\|_2^2] \geq \mathbb{V}[p]$.*

Proof. Again, write p in the tensored Hermite basis $p = \sum_I c_I \phi_I$. We know that

$$\sum_i \mathbb{E}\left[\left(\frac{\partial}{\partial x_i} p(g)\right)^2\right] = \sum_I c_I^2 \cdot \sum_i I_i \geq \sum_{I\neq\emptyset} c_I^2 = \mathbb{V}[p],$$

from which the claim follows. □

The following more careful estimate gives something better than what Cauchy-Schwarz, Corollary 1.3.17, and Lemma 2.3.8 imply.

Lemma 2.3.10. *For any $p \in \mathbb{R}_d[x_1, ..., x_r]$, $\mathbb{E}_g\left[\|g\|^2 \cdot \|\nabla p(g)\|_2^2\right]^{1/2} \leq O(rd) \cdot \mathbb{V}[p]^{1/2}$.*

Proof. Take any $i, j \in [r]$. Let $q_I^{i,j}$ denote the polynomial $\prod_{\ell\in[|I|]:\ell\neq i,j} \phi_{I_\ell}(x_\ell)$. If $i = j$, then

$$\mathbb{E}\left[g_i^2 \cdot \left(\frac{\partial}{\partial x_j} p(g)\right)^2\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\sum_I c_I \cdot q_I^{i,i}(g) \cdot \sqrt{I_i} \cdot g_i \cdot \phi_{I_i-1}(x_i)\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_I c_I \cdot q_I^{i,i}(g) \cdot \sqrt{I_i} \cdot \left(\sqrt{I_i} \cdot \phi_{I_i}(g_i) + \sqrt{I_i - 1} \cdot \phi_{I_i-2}(g_i)\right)\right)^2\right] \\
&\leq 2\sum_I c_I^2 \cdot I_i^2 + 2\sum_I c_I^2 \cdot I_i(I_i - 1) \leq 4d^2\,\mathbb{V}[p],
\end{aligned}
$$

where the second step follows by Corollary 2.3.5, and the third step follows by the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$. Likewise, if $i \neq j$, then we have that

$$
\begin{aligned}
&\mathbb{E}\left[g_i^2 \cdot \left(\frac{\partial}{\partial x_j} p(g)\right)^2\right] \\
&\quad= \mathbb{E}\left[\left(\sum_I c_I \cdot q_I^{i,j}(g) \cdot g_i \cdot \phi_{I_i}(g_i) \cdot \sqrt{I_j} \cdot \phi_{I_j-1}(g_j)\right)^2\right] \\
&\quad= \mathbb{E}\left[\left(\sum_I c_I \cdot q_I^{i,j}(g) \cdot \left(\sqrt{I_i + 1} \cdot \phi_{I_i+1}(g_i) + \sqrt{I_i} \cdot \phi_{I_i-1}(g_i)\right) \cdot \sqrt{I_j} \cdot \phi_{I_j-1}(g_j)\right)^2\right] \\
&\quad\leq 2\left(\sum_I c_I^2 \cdot (I_i + 1)I_j + \sum_I c_I^2 \cdot I_i I_j\right) \leq 4d(d+1)\,\mathbb{V}[p] \leq 5d^2\,\mathbb{V}[p].
\end{aligned}
$$

The lemma follows upon summing over $i, j \in [r]$. □

The following basic inequality will also be useful.

Lemma 2.3.11. *Let $\mathcal{S}$ denote the collection of all multisets I of size at most d consisting of elements of $[r]$. Then $\mathbb{E}\left[\left(\sum_I \phi_I(g)^2\right)^2\right] \leq O(r)^{2d}$.*

Proof. We have that

$$
\mathbb{E}\left[\left(\sum_I \phi_I(g)^2\right)^2\right] \leq |\mathcal{S}| \cdot \mathbb{E}\left[\sum_I \phi_I(g)^4\right] = |\mathcal{S}| \cdot 9^d \sum_I \mathbb{E}\left[\phi_I(g)^2\right] = |\mathcal{S}|^2 \cdot 9^d = O(r)^{2d},
$$

where the first step follows by Cauchy-Schwarz, the second by Fact 1.3.15, the third by orthonormality of $\{\phi_I\}$, and the last by the fact that $|\mathcal{S}| = O(r)^d$. □

2.3.4 More Subspace Distance Inequalities

We first give a more refined estimate for $d_P(V, V')^2 - d_C(V, V')^2$ than what Lemma 1.3.5 tells us:

Lemma 2.3.12. $d_P(V, V')^2 - d_C(V, V')^2 \leq d_P(V, V')^4$.

Proof. From the elementary inequality $4\sin^2(\theta/2) - \sin^2(\theta) \leq \sin^4(\theta)$ for $\theta \in [0, \pi/2]$, we see that

$$d_P(V, V')^2 - d_C(V, V')^2 = \left(\sum_{i=1}^{r} \sin^4 \theta_i\right)^2 \leq \left(\sum_{i=1}^{r} \sin^2 \theta_i\right)^2 = d_C(V, V')^4 \leq d_P(V, V')^4$$

as claimed. □

The following consequence of Lemma 2.3.12 will be useful in our analysis of GeoSGD.

Lemma 2.3.13. *For $V, V^* \in St^n_r$, we have that $\|Id - V^\top V^*\|_2 \leq \|V - V^*_F\|$. If V, V^* additionally satisfy that $\|V - V^*\|_F = d_P(V, V^*)$, then we have that $\|Id - V^\top V^*\|_2 \leq d_P(V, V^*)^2$.*

Proof. It suffices to upper bound $\|\mathrm{Id} - V^\top V^*\|_F$. Note that

$$\begin{aligned}\|\mathrm{Id} - V^\top V^*\|_F^2 &= d - 2\,\mathrm{Tr}(V^\top V^*) + \|V^\top V^*\|_F^2 \\ &= \|V - V^*\|_F^2 - d_C(V, V^*)^2 \leq \|V, V^*\|_F^2,\end{aligned}$$

from which the first part of the lemma follows.

For the second bound, note that

$$\|V - V^*\|_F^2 - d_C(V, V^*)^2 = d_P(V, V^*)^2 - d_C(V, V^*)^2 \leq d_P(V, V^*)^4,$$

where the final step follows by Lemma 2.3.12. □

2.4 Warm Start via Filtered PCA

The main result of this section is the proof of Theorem 2.2.2. Let $\mathcal{D}$ denote the distribution (X, Y) where $Y = P(X)$ is a α non-degerate polynomial of rank r and degree at most d

as in the hypothesis of the theorem. Let U^* be the true hidden subspace defining P. The proof follows the outline described in the introduction closely. To this end, for a *threshold parameter* $\tau > 0$ and a collection of unit vectors $V = \{v_1, \dots, v_\ell\}$, define the matrix

$$\mathbf{M}_V^\tau \triangleq \Pi_{V^\perp} \cdot \left(\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}[\{|y| > \tau\} \ \wedge \ \{|\langle v_i, x\rangle| \le 1, \ \forall\, i \in [\ell]\}] \cdot (xx^\top - \mathrm{Id})\right]\right) \cdot \Pi_{V^\perp}.$$

```
Algorithm 4: FilteredPCAv1($\mathcal{D}, \varepsilon, \delta$)
Input: Sample access to $\mathcal{D}$, target error $\varepsilon$, failure probability $\delta$
Output: Frame for a subspace $U$ with $d_P(U, U^*) \le \varepsilon$, with probability at least $1-\delta$
$V_0 \leftarrow \emptyset$.
$\tau \leftarrow \tau(r, d, \alpha)$                                   // Lemma 2.4.1
for $0 \le \ell \le r-1$ do
    Draw $N = O_{r,d,\varepsilon}(n)$ samples $(x_1, y_1), ..., (x_N, y_N)$        // Theorem 2.2.2
    Compute an empirical approximation $\widehat{\mathbf{M}}^\ell$ to $\mathbf{M}_{V_\ell}^\tau$ by drawing $N = O_{r,d,\varepsilon}(n)$
      samples from the distribution $\mathcal{D}$.
    Let $v^{\ell+1}$ be the eigenvector with the largest eigenvalue of $\widehat{\mathbf{M}}^\ell$.
    $V_{\ell+1} \leftarrow V_\ell \cup \{v^{\ell+1}\}$.
return $V_r$.
```

We will show that the above algorithm satisfies the guarantees of Theorem 2.2.2. The core of its analysis will be the following main inductive lemma.

Lemma 2.4.1. *There exists $\tau = \tau(r, d, \alpha)$, a constant $C = C(r, d, \alpha)$ such that the following holds. Let $V = \{v_1, \dots, v_\ell\}$ for $\ell < r$ be orthonormal vectors such that $\|\Pi_{U^*} v_i\| \ge 1 - \rho$, and $\mathbf{M}$ a matrix such that $\|\mathbf{M} - \mathbf{M}_V^\tau\| \le \rho$. Then, the largest eigenvector v of $\mathbf{M}$ satisfies $\|\Pi_{U^*} v\| \ge 1 - C\rho^{1-1/r}$.*

Before proving the lemma, we first show how the main theorem follows from the above.

Proof of Theorem 2.2.2. Let C, τ be as in the above lemma. For a ρ_0 to be chosen later, let $\rho_{\ell+1} = C\rho_\ell^{1-1/r}$ for $\ell \ge 0$. Let $N = O(n \log(r/\delta)/\rho_0^2)$.

We will show by induction that $\|\Pi_{U^*} v_\ell\| \ge 1 - \rho_\ell$. Suppose we have the statement for $v_1, \dots, v_\ell$ computed by the algorithm. Then, in the next iteration, by Lemma 1.3.32, with probability at least $1 - \delta/r$, we will have $\|\widehat{\mathbf{M}}^\ell - \mathbf{M}_{V_\ell}^\tau\| \le \rho_\ell$. In this case, by Lemma 2.4.1, the top eigenvector $v_{\ell+1}$ of $\widehat{\mathbf{M}}^\ell$ satisfies $\|\Pi_{U^*} v_{\ell+1}\| \ge 1 - C\rho_\ell^{1-1/r} = 1 - \rho_{\ell+1}$.

By a union bound over the r events, we get that with probability at least $1-\delta$, we would have computed orthonormal vectors $v_1, \ldots, v_r$ such that $\|\Pi_{U^*} v_i\| \geq 1-\rho_r$. Now, by Lemma 1.3.9, $d_P(\mathrm{span}(v_1, \ldots, v_r), U^*) \leq O(\sqrt{\rho_r r})$.

As $\rho_r \leq C^r \rho_0^{(1-1/r)^r} \leq C^r \rho_0^{\Theta(1)}$, the lemma follows by setting $\rho_0 = \mathrm{poly}(\varepsilon)/C^r$. The overall sample complexity will be $N = O(r \cdot n \log(r/\delta)/\rho_0^2 = C(r,d,\alpha) n \log(r/\delta)/\varepsilon^2$ as stated in the theorem. The runtime then follows by applying Fact 1.3.6. □

2.4.1 Proof of Lemma 2.4.1

We next prove the Lemma 2.4.1 which allows us to identify one direction at a time. The proof proceeds as follows:

1. We first show a lower bound on the largest eigenvalue of the matrix $\mathbf{M}_V^\tau$ when the vectors $v_1, \ldots, v_\ell$ lie in the subspace U^*. This is the heart of the proof and follows from a compactness argument. This essentially gives a proof of the lemma when $V \subseteq U^*$ (and $\mathbf{M}$ approximates $\mathbf{M}_V^\tau$). See Lemmas 2.4.2, 1.3.8.

2. The second step is to reduce to the above case. Given V as in the lemma, we find orthonormal vectors $V^* = \{v_1^*, \ldots, v_\ell^*\} \in U^*$ such that $\|v_i - v_i^*\| \leq O(\ell\rho)$. We then do a perturbation analysis (using elementary linear algebra) to argue that perturbing the vectors V slightly will only incur a small error in the matrix $\mathbf{M}_V^\tau$. Specifically, we will show that $\|\mathbf{M}_V^\tau - \mathbf{M}_{V^*}^\tau\| \leq O(\mathrm{poly}(r)\rho^{1/2-1/2r})$. See Lemma 2.4.3.

For brevity, in the remainder of this section let Π^* denote orthogonal projection to the true subspace $U^* \subset \mathbb{R}^n$.

First, we show that if the vectors in $V^* = \{v_1^*, ..., v_\ell^*\}$ were vectors in the true subspace, then the top eigenvector of $\mathbf{M}_{V^*}^\tau$ will be a new vector in the subspace orthogonal to the preceding ones.

Lemma 2.4.2. *There are absolute constants $\tau = \tau_{r,d,\nu_{\mathsf{cond}}} > 0$ and $\lambda = \lambda_{r,d,\nu_{\mathsf{cond}}} > 0$ for which the following holds. Suppose $V^* = \{v_1^*, ..., v_\ell^*\} \subset \mathbb{S}^{n-1}$ are orthogonal and is in U^*. Then*

1. *The kernel of $\mathbf{M}_{V^*}^\tau$ contains $\mathrm{span}(v_1^*, ..., v_\ell^*)$ as well as the orthogonal complement of U^*.*

2. The top eigenvalue of $\mathbf{M}^\tau_{V^*}$ *is at least* λ *and corresponds to a vector in* $U^*\backslash\mathrm{span}(V^*)$.

Note that Lemma 2.4.2 already gives a nontrivial algorithmic guarantee for $\ell = 0$: given exact access to $\mathbf{M}^\tau_\emptyset$, we can recover a vector inside the true subspace by taking its top eigenvector.

Proof. Extend $\{v_i^*\}_{i\in[\ell]}$ to an orthonormal basis $\{v_i^*\}_{i\in[r]}$ of U^*, and let $p^*((V^*)^\top x)$ be a realization of the true low-rank polynomial, where the frame $V^* \in \mathrm{St}^n_r$ consists of these basis elements.

(Proof of 1) Certainly $\mathrm{span}(\{v_i^*\}_{i\in[\ell]})$ lies in the kernel of $\mathbf{M}^\tau_{V^*}$ by definition. Moreover for any $v \in \mathbb{S}^{n-1}$ orthogonal to U^*, because $\langle v_i^*, x\rangle, ..., \langle v_r^*, x\rangle, \langle v, x\rangle$ are independent Gaussians, call them $g_1, ..., g_r, g_\perp \sim \mathcal{N}(0,1)$, we have that

$$\begin{aligned} v^\top \mathbf{M}^\tau_{V^*} v &= \mathbb{E}\left[\mathbb{1}[\{|p^*(g_1,...,g_r)| > \tau\} \ \wedge\ \{|g_i| \le 1\ \forall\ i \in [\ell]\}] \cdot (g_\perp^2 - 1)\right] \\ &= \mathbb{E}\left[\mathbb{1}[\{|p^*(g_1,...,g_r)| > \tau\} \ \wedge\ \{|g_i| \le 1\ \forall\ i \in [\ell]\}]\right] \cdot \mathbb{E}\left[(g_\perp^2 - 1)\right] \\ &= 0. \end{aligned}$$

(Proof of 2) The fact that the top eigenvector lies in $U^*\backslash\mathrm{span}(\{v_i^*\}_{i\in[\ell]})$ follows immediately from the fact that it must be orthogonal to both $\mathrm{span}(\{v_i^*\}_{i\in[\ell]}$ and the orthogonal complement of U^*.

To get a bound on the top eigenvalue, define the quantities $Z_i \triangleq {v^*}_i^\top \mathbf{M}^\tau_{V^*} v_i^*$ for $\ell < i \le r$. Again using the fact that $\langle v_1^*, x\rangle, ..., \langle v_r^*, x\rangle$ are independent Gaussians $g_1, ..., g_r$, we have

$$\sum_{i=\ell+1}^r Z_i = \mathbb{E}\left[\mathbb{1}[\{|p^*(g_1,...,g_r)| > \tau\} \ \wedge\ \{|g_i| \le 1\ \forall\ i \in [\ell]\}] \cdot \left(\sum_{i>\ell} g_i^2 - (r-\ell)\right)\right].$$

We would like to lower bound this quantity, at which point by averaging over i we conclude the proof of the lemma.

Let $K \subset \mathbb{R}^r$ denote the set of all points x for which $|x_i| \le 1$ for all $1 \le i \le \ell$ and for which $\sum_{i=\ell+1}^r x_i^2 \le 2(r-\ell)$. For any $p \in \mathcal{P}^{\nu_{\mathsf{cond}}}_{r,d}$, define $\|p\|_K \triangleq \sup_{x\in K} |p(x)|$. By compactness of K, $\|p\|_K < \infty$ for all p, and furthermore $\|p\|_K$ is a continuous function of p. If we take $\tau = \tau(\nu_{\mathsf{cond}}, r, d, \ell) \triangleq \sup_{p\in\mathcal{P}^{\nu_{\mathsf{cond}}}_{r,d}} \|p\|_K$, then by compactness of $\mathcal{P}^{\nu_{\mathsf{cond}}}_{r,d}$, is some finite quantity depending only on ν_{cond}, r, d, and ℓ. For this choice of τ, we conclude that if a point $(g_1, ..., g_r) \in \mathbb{R}^r$ satisfies $|p^*(g_1, ..., g_r)| > \tau$ and $|g_i| \leq 1$ for all $i \in [\ell]$, then it must lie outside K. We conclude that

$$\sum_{i=\ell+1}^{r} Z_i \geq (r-\ell) \cdot \Pr\left[\{|p^*(g_1, ..., g_r)| > \tau\} \ \wedge \ \{g \notin K\}\right].$$

In particular, there exists some $i > \ell$ for which $Z_i \geq \Pr\left[\{|p^*(g_1, ..., g_r)| > \tau\} \ \wedge \ \{g \notin K\}\right]$. The right-hand side is a continuous function in p, call it A_p. For any p, there must exist some point $x \notin K$ for which $p^*(x) > \tau$, so again by compactness of $\mathcal{P}^{\nu_{\mathsf{cond}}}_{r,d}$, we see that $Z_i \geq \lambda$ for some strictly positive constant λ depending only on $\nu_{\mathsf{cond}}, r, d, \ell$. □

Henceforth, for brevity, we will denote the constants $\tau_{r,d,\nu_{\mathsf{cond}}}$ and $\lambda_{r,d,\nu_{\mathsf{cond}}}$ from Lemma 2.4.2 by τ and λ respectively.

Note that by Corollary 1.3.8, the above lemma implies Lemma 2.4.1 for the case when $V \subseteq U^*$.

Finally, we show that for orthonormal vectors $V = \{v_1, ..., v_\ell\}$ which all have large component in U^*, the matrix $\mathbf{M}^\tau_V$ is spectrally close to some $\mathbf{M}^\tau_{V^*}$ for $V^* = \{v^*_1, \ldots v^*_\ell\}$ in U^*.

Lemma 2.4.3. *There is an absolute constant $c_2 > 0$ for which the following holds. Let $1 \leq \ell \leq r$. Given orthonormal vectors $V = \{v_1, ..., v_\ell\}$ for which $\|\Pi^* v_i\|_2 \geq 1 - \varepsilon$ for some $0 \leq \varepsilon < 1$ for all $i \in [\ell]$, there exist orthonormal vectors $V^* = \{v^*_1, \ldots, v^*_\ell\} \subset U^*$ such that $\|\mathbf{M}^\tau_V - \mathbf{M}^\tau_{V^*}\|_2 \leq c_2(\varepsilon\ell)^{1/2-1/2r}\ell r$.*

Proof. Let $V^* = \{v^*_1, \ldots, v^*_\ell\}$ be orthonormal vectors in U^* guaranteed by Lemma 1.3.10 such that $\langle v_i, v^*_i\rangle \geq 1 - 2\varepsilon\ell$.

For each $0 \leq a \leq \ell$, define the hybrid collections of vectors $V^{(a)} \triangleq \{v^*_1, ..., v^*_{a-1}, v_a, ..., v_\ell\}$, and also define the hybrid matrices

$$\mathbf{M}^{(a)} \triangleq \left(\Pi^\perp_{\{v_i\}}\right)^\top \cdot \left(\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}[\{|y| > \tau\} \ \wedge \ \{|\langle v^{(a)}_i, x\rangle| \leq 1 \ \forall \, i \in [\ell]\}] \cdot (xx^\top - \mathrm{Id})\right]\right) \cdot \Pi^\perp_{\{v_i\}}.$$

Note that $V^{(0)} = V$ and $V^{(\ell)} = V^*$, and similarly $\mathbf{M}^{(0)} = \mathbf{M}^\tau_V$.

We will bound $\|\mathbf{M}^{(a+1)} - \mathbf{M}^{(a)}\|_2$ for every $0 \leq a < \ell$, and then bound $\|\mathbf{M}^{(\ell)} - \mathbf{M}^\tau_{V^*}\|_2$.

The lemma will then follow by triangle inequality.

Claim 2.4.4. *For any* $0 \le a < \ell$, $\|\mathbf{M}^{(a+1)} - \mathbf{M}^{(a)}\|_2 \le O\left((\varepsilon\ell)^{1/2-1/2r} \cdot r\right)$.

Proof. We will bound $v^\top(\mathbf{M}^{(a+1)} - \mathbf{M}^{(a)})v$ for any $v \in \mathbb{S}^{n-1}$; without loss of generality, we may assume v is orthogonal to $v_1, ..., v_\ell$.

Let $\mathcal{E}$ denote the event that $|\langle v_a, x\rangle| > 1$ and $|\langle v_a^*, x\rangle| \le 1$ or vice-versa, noting that the indicator events in the definitions of $\mathbf{M}^{(a)}$ and $\mathbf{M}^{(a+1)}$ only differ when $\mathcal{E}$ occurs. Therefore,

$$\begin{aligned}\left|v^\top(\mathbf{M}^{(a+1)} - \mathbf{M}^{(a)})v\right| &\le \Pr[\mathcal{E}]^{1-1/r} \cdot \mathbb{E}[(\langle v, x\rangle^2 - 1)^r]^{1/r} \\ &= \Pr[\mathcal{E}]^{1-1/r} \cdot O(r)\end{aligned}$$

where the last inequality follows by Holder's and the fact that $\mathbb{E}_{g\sim\mathcal{N}(0,1)}[(g^2-1)^r]^{1/r} = O(r)$.

Finally note that by Lemma 1.3.35,

$$Pr[\mathcal{E}] \le O\left(\sqrt{1 - \langle v_i, v_i^*\rangle^2}\right) = O(\sqrt{\varepsilon\ell}).$$

The claim now follows. □

To bound $\|\mathbf{M}^{(\ell)} - \mathbf{M}^\tau_{V^*}\|_2$, we will use Claim 1.3.11. We note that the matrix

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}[\{|y| > \tau\} \ \wedge \ \{|\langle v_i^*, x\rangle| \le 1 \ \forall \ i \in [\ell]\}] \cdot (xx^\top - \mathrm{Id})\right]$$

has spectral norm at most $\|\mathbb{E}[xx^\top]\|_2 + 1 = 2$. So if $U \triangleq \mathrm{span}(v_1, ..., v_\ell)$ and $U' \triangleq \mathrm{span}(v_1^*, ..., v_\ell^*)$, then by Claim 1.3.11,

$$\|\mathbf{M}^{(\ell)} - \mathbf{M}^\tau_{V^*}\|_2 \le O(d_C(U, U')) \le O(\sqrt{\varepsilon \cdot \ell}),$$

where the last step follows by Lemma 1.3.9.

Lemma 2.4.3 follows by applying the above inequality, Claim 2.4.4 for all $0 \le a < \ell$, and triangle inequality. □

We now put Corollary 1.3.8, 2.4.3 together to prove Lemma 2.4.1.

Proof of Lemma 2.4.1. Choose τ to be as in Lemma 2.4.2. We will choose $C = c\ell^{3-1/r}r^2/\lambda^2$ for $\lambda = \lambda_{r,d,\nu_{\mathsf{cond}}}$ from Lemma 2.4.2 and $c > 0$ a universal constant.

Let V^* be the set of ℓ orthonormal vectors in U^* as in Lemma 2.4.3 so that

$$\|\mathbf{M}^\tau_V - \mathbf{M}^\tau_{V^*}\| \le O((\rho\ell)^{1/2-1/2r}\ell r).$$

By triangle inequality, $\|\mathbf{M} - \mathbf{M}^\tau_{V^*}\| \le O((\rho\ell)^{1/2-1/2r}\ell r)$. The lemma now follows by applying Corollary 1.3.8. $\square$

2.5 Boosting via Stochastic Riemannian Optimization

In this section we describe our algorithm for boosting a warm start to arbitrary accuracy and defer the details of its analysis to Sections 2.7 and 2.6.

Theorem 2.5.1 (Error Guarantee for GeoSGD)**.** *There is an absolute constant $c_3 > 0$ such that the following holds. Let U^* be the true subspace of $\mathcal{D}$. Given $V^{(0)} \in St^n_r$ spanning a subspace U for which $d_P(U, U^*) \le (c_3 \cdot dr^3)^{-d-2}$, if in the specification of* GeoSGD *we take*

$$T = \frac{n}{\nu_{\mathsf{cond}}} \cdot \log(1/\varepsilon) \cdot \mathrm{poly}(\ln(1/\nu_{\mathsf{cond}}), r, d, \ln(1/\delta), \ln(n))^d, \tag{2.5}$$

then GeoSGD*$(\mathcal{D}, V^{(0)}, \varepsilon, \delta)$ returns $(\mathbf{c}^{(T)}, V^{(T)})$ for which there exists a realization $(\mathbf{c}^*, V^*)$ of $\mathcal{D}$ such that $d_P(V^{(T)}, V^*) \le \varepsilon$ and $\|\mathbf{c}^{(T)} - \mathbf{c}^*\|_2 \le \varepsilon$.*

Theorem 2.5.2 (Complexity of GeoSGD)**.** *Let $T_1 \triangleq O(rd^4)^{d+1} \cdot \log(1/\varepsilon)$, $B \triangleq O(\log(T_1 \cdot T/\delta))^{2d}$, and $T_2 \triangleq (r/\nu_{\mathsf{cond}})^2 \cdot O(d \cdot \log(T/\delta))^{2c_1 d}$. Then* GeoSGD *draws*

$$N \triangleq T \cdot (B \cdot T_1 + T_2) = \tilde{O}\left(\frac{n \log^2(1/\varepsilon)}{\nu^3_{\mathsf{cond}}} \cdot \mathrm{poly}(\ln(1/\nu_{\mathsf{cond}}), r, d, \ln(1/\delta), \ln(n))^d\right)$$

samples and runs in time $n \cdot r^{O(d)} \cdot N$ time.

2.5.1 Preliminaries

Let $M = r^{O(d)}$ be the dimension of the linear space of polynomials of polynomials of degree d over r variables. For $\mathbf{c} = \{c_I\} \in \mathbb{R}^M$, where I ranges over multisets of size at most d consisting of elements of $[r]$, and $V \in \mathrm{St}^n_r$, let parameters $\Theta = (\mathbf{c}, V)$ correspond to a rank-r polynomial $F_x(\Theta) \triangleq \sum_I c_I \phi_I(V^\top x)$ in the variable x. Given a sample $(x, y) \sim \mathcal{D}$, let $L_x(\Theta) \triangleq (F_x(\Theta) - y)^2$ denote the empirical risk of a single sample.

We will often regard F_x and L_x as functions solely in $\mathbf{c}$ (resp. V) for a fixed choice of V (resp. $\mathbf{c}$): given a fixed V (resp. a fixed $\mathbf{c}$), define $F^V_x(\mathbf{c})$ and $L^V_x(\mathbf{c})$ (resp. $F^{\mathbf{c}}_x(V)$ and $L^{\mathbf{c}}_x(V)$) in the obvious way.

Let $\nabla F_x(\Theta)$ denote the gradient of F_x as a function on Euclidean space, and let $\nabla^{\mathsf{vec}} F_x(\Theta) \triangleq \nabla F^c_x(V)$ and $\nabla^{\mathsf{coef}} F_x(\Theta) \triangleq \nabla F^V_x(\mathbf{c})$ denote its components corresponding to V and $\mathbf{c}$ respectively. We can compute their gradients, indeed all of their higher derivative tensors, explicitly:

Proposition 2.5.3. *For any $x \in \mathbb{R}^n$, $a, b \in \mathbf{Z}_{\geq 0}$, and $\Theta = (c, V)$,*

$$\frac{\partial^{a+b}}{\partial c_{I^{(1)}} \cdots \partial c_{I^{(a)}} \partial V_{i_1,j_1} \cdots \partial V_{i_b,j_b}} F_x(\Theta) = \begin{cases} \left(\prod_{\nu=1}^b x_{i_\nu}\right) \cdot p^{[b]}(V^\top x) & \textit{if } a = 0 \\ \left(\prod_{\nu=1}^b x_{i_\nu}\right) \cdot \phi^{[b]}_I(V^\top x) & \textit{if } a = 1 \\ 0 & \textit{otherwise} \end{cases}$$

From Proposition 2.5.3 we conclude that

$$\nabla^{\mathsf{vec}} F_x(\Theta) = x \cdot (\nabla p(V^\top x))^\top \quad \text{and} \quad \nabla^{\mathsf{coef}} F_x(\Theta) = \{\phi_I(V^\top x)\}_I.$$

It will be important to consider $\overline{\nabla}^{\mathsf{vec}} F_x(\Theta) \triangleq \Pi^\perp_V \nabla^{\mathsf{vec}} F_x(\Theta)$ the projection of $\nabla^{\mathsf{vec}} F_x(\Theta)$, to the tangent space of $\mathrm{G}(n, r)$ at the point $[V]$.

Lastly, we record here an elementary estimate which will be used repeatedly in the proceeding sections and defer its proof to Appendix 2.11.1.

Lemma 2.5.4. *For any integer $m \geq 1$ and $\ell = (\ell_1, ..., \ell_m) \in [d+1]^m$,*

$$\left|\mathbb{E}\left[\prod_{\nu=1}^{m} \left\langle \nabla^{[\ell_\nu]} F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell_\nu} \right\rangle\right]\right| \leq 2^m \cdot \left(2mdr^2\right)^{m(d+1)/2} \cdot \|V^* - V\|_F^{\sum_\nu \ell_\nu} \cdot \left(1 + \frac{\|\mathbf{c} - \mathbf{c}^*\|_2}{\|V^* - V\|_F}\right)^m$$

2.5.2 Gradient Updates: Vanilla and Geodesic

GeoSGD alternates between one of two phases: updating $\mathbf{c}$ or updating V. Our updates for $\mathbf{c}$ are straightforward: at iterate $\Theta = (\mathbf{c}, V)$ and given a batch of samples $(x_0, y_0), ..., (x_{B-1}, y_{B-1}) \sim \mathcal{D}$, we fix V and take a vanilla gradient descent step using $\frac{1}{B}\sum_{i=0}^{B-1} L^V_{x_i}(\mathbf{c})$. For learning rate η_{coef}, this leads to the update

$$c'_I = c_I - 2\eta_{\mathsf{coef}} \cdot \frac{1}{B}\sum_{i=0}^{B-1}(F_{x_t}(\Theta) - F_{x_i}(\Theta^*)) \cdot \phi_I(V^\top x_i) \triangleq c_I - \frac{1}{B}\sum_{i=0}^{B-1}\left(\Delta^{\Theta, x_i}_{\mathsf{coef}}\right)_I \ \forall\, I. \tag{2.6}$$

The updates for V will be less standard. At iterate $\Theta = (\mathbf{c}, V)$, and given a sample $(x, y) \sim \mathcal{D}$, consider the geodesic Γ on $\mathrm{G}(n, r)$ with initial point $[V] \in \mathrm{G}(n, r)$ and initial velocity $\dot{\Gamma}(0) \triangleq \Pi^\perp_V \nabla L^{\mathbf{c}}_x(V)$, where $L^{\mathbf{c}}_x(V) \triangleq L_x(\Theta)$.[2]

Define the vectors $h^{\Theta,x} \in \mathbb{R}^n, \nabla^{\Theta,x} \in \mathbb{R}^r$ by

$$h^{\Theta,x} \triangleq 2(F_x(\Theta) - F_x(\Theta^*)) \cdot \Pi^\perp_V \cdot x \quad \text{and} \quad \nabla^{\Theta,x} \triangleq \nabla p(V^\top x) \tag{2.7}$$

so that $\dot{\Gamma}(0) = h^{\Theta,x} \cdot (\nabla^{\Theta,x})^\top$. Geodesics on $\mathrm{G}(n, r)$ are determined by the SVD of the initial velocity $\dot{\Gamma}(0)$, which is simply given by

$$\dot{\Gamma}(0) = \sigma \cdot \widehat{h}^{\Theta,x} \cdot (\widehat{\nabla}^{\Theta,x})^\top,$$

where

$$\widehat{h}^{\Theta,x} \triangleq \frac{h^{\Theta,x}}{\|h^{\Theta,x}\|} \qquad \widehat{\nabla}^{\Theta,x} \triangleq \frac{\nabla^{\Theta,x}}{\|\nabla^{\Theta,x}\|} \qquad \sigma^{\Theta,x} \triangleq \|h^{\Theta,x}\| \cdot \|\nabla^{\Theta,x}\|.$$

[2] We emphasize that technically this is not well-defined as this velocity depends on the choice of representative V; indeed, $F^{\mathbf{c}}_x(V)$ cannot be regarded as a function on $\mathrm{G}(n, r)$, as $\mathbf{c}$ is fixed so that different rotations of V will actually yield different values. But as our goal is simply to produce an update rule, we can freely ignore this point and see where this line of reasoning leads.

Walking along the geodesic with initial velocity $\dot{\Gamma}(0)$ for time η_{vec} then yields the following update rule (for the details, see the derivation of equation (2.65) in [EAS98]),

$$V' \triangleq V - \left(\cos\left(\sigma^{\Theta,x}\eta_{\mathsf{vec}}\right) - 1\right) \cdot V \cdot \widehat{\nabla}^{\Theta,x}(\widehat{\nabla}^{\Theta,x})^\top - \sin\left(\sigma\eta_{\mathsf{vec}}\right) \cdot \widehat{h}^{\Theta,x}\left(\widehat{\nabla}^{\Theta,x}\right)^\top \triangleq V - \Delta^{\Theta,x}_{\mathsf{vec}}. \quad (2.8)$$

One readily checks that the columns of V' are orthonormal.

In Algorithm 7, we state our boosting algorithm GeoSGD, which is composed of two alternating phases, SubspaceDescent and RealignPolynomial which execute the updates (2.6) and (2.8) respectively. In the next two sections, we will analyze these two phases.

```
Algorithm 5: SubspaceDescent($\mathcal{D}, V^{(0)}, \mathbf{c}\delta$)
  Input: Sample access to $\mathcal{D}$; frame $V^{(0)} \in \mathrm{St}^n_r$; coefficients $\mathbf{c} \in \mathbb{R}^M$, failure
         probability $\delta$
  Output: $V^{(T)} \in \mathrm{St}^n_r$ which is slightly closer to the true subspace than $V$, provided
          $(\mathbf{c}, V^{(0)})$ satisfies certain conditions (see Theorem 2.7.1 for formal
          guarantees)
  Define iteration count $T$ according to (2.22).
  Define learning rate $\eta_{\mathsf{vec}}$ according to (2.21).
  $\Theta^{(0)} \leftarrow (\mathbf{c}, V^{(0)})$
  for $0 \le t < T$ do
      Sample $(x^t, y^t) \sim \mathcal{D}$ $\widehat{h} \leftarrow \frac{h^{\Theta^{(t)},x^t}}{\|h^{\Theta^{(t)},x^t}\|}$ and $\widehat{\nabla} \leftarrow \frac{\nabla^{\Theta^{(t)},x^t}}{\|\nabla^{\Theta^{(t)},x^t}\|}$    // equation (2.7)
      $\sigma \leftarrow \|h^{\Theta^{(t)},x^t}\| \cdot \|\nabla^{\Theta^{(t)},x^t}\|.$; $V^{(t+1)} \leftarrow V^{(t)} - \Delta^{\Theta^{(t)},x^t}_{\mathsf{vec}}$    // equation (2.8)
      $\Theta^{(t+1)} \leftarrow (\mathbf{c}, V^{(t+1)})$
  return $V^{(T)}$.
```

2.6 Guarantees for RealignPolynomial

Before we can describe our main result of this section, we require some setup.

Henceforth, fix a frame $V \in \mathrm{St}^n_r$. The aim of RealignPolynomial is to approximately find the r-variate, degree-d polynomial p for which $p(V^\top x)$ is closest to the true low-rank polynomial. Suppose V was β-far in subspace distance from the true subspace for some β, or equivalently, that there was some frame $V^* \in \mathrm{St}^n_r$ for the true subspace for which $\|V - V^*\|_F = \beta$. By working with V instead of V^*, we obviously cannot hope to produce

Algorithm 6: $\textsc{RealignPolynomial}(\mathcal{D}, V, \underline{\varepsilon}, \delta)$

```
Input: Sample access to $\mathcal{D}$; $V \in \mathrm{St}^n_r$; target error $\underline{\varepsilon}$; failure probability $\delta$
Output: $\mathbf{c} \in \mathbb{R}^M$ for which $(\mathbf{c}^{(T)}, V)$ is close to a realization of $\mathcal{D}$ (see Section 2.6
        for details)
Define batch size $B$ according to (2.11).
Define iteration count $T$ according to (2.10).
Define learning rate $\eta_{\mathsf{coef}}$ according to (2.9).
$\mathbf{c}^{(0)} \leftarrow \mathbf{0}$.
$\Theta^{(0)} \leftarrow (\mathbf{c}^{(0)}, V)$.
for $0 \le t < T$ do
    Sample $(x^t_1, y^t_1), ..., (x^t_B, y^t_B) \sim \mathcal{D}$.
    For every $I$, $c^{(t+1)}_I \leftarrow c^{(t)}_I - \frac{1}{B}\sum_{i=0}^{B-1}\left(\Delta^{\Theta, x^t_i}_{\mathsf{coef}}\right)_I$        // equation (2.6)
    $\mathbf{c}^{(t+1)} \leftarrow \left\{c^{(t+1)}_I\right\}_I$ and $\Theta^{(t)} \leftarrow (\mathbf{c}^{(t+1)}, V)$
return $\mathbf{c}^{(T)}$.
```

Algorithm 7: $\textsc{GeoSGD}(\mathcal{D}, V^{(0)}, \varepsilon, \delta)$

```
Input: Sample access to $\mathcal{D}$, $V^{(0)} \in \mathrm{St}^n_r$, target error $\varepsilon$, failure probability $\delta$
Output: $\Theta = (\mathbf{c}^{(T)}, V^{(T)}) \in \mathcal{M}$ for which $d_P(V^{(T)}, V^*) \le \varepsilon$ and $\|\mathbf{c} - \mathbf{c}^*\|_2 \le \varepsilon$ for
        some realization $(\mathbf{c}^*, V^*)$ of $\mathcal{D}$
Define iteration count $T$ according to (2.5)
$\delta' \leftarrow \delta/(2T+1)$
for $0 \le t < T$ do
    $\mathbf{c}^{(t)} \leftarrow \textsc{RealignPolynomial}(\mathcal{D}, V^{(t)}, \varepsilon/2, \delta')$
    $V^{(t+1)} \leftarrow \textsc{SubspaceDescent}(\mathcal{D}, V^{(t)}, \mathbf{c}^{(t)}, \delta')$
$\mathbf{c}^{(T)} \leftarrow \textsc{RealignPolynomial}(\mathcal{D}, V^{(T)}, \varepsilon/2, \delta')$
return $\Theta \triangleq (\mathbf{c}^{(T)}, V^{(T)})$.
```

p for which $p(V^\top x)$ is exactly equal to the true low-rank polynomial $p_*(V^{*\top} x)$. But it is reasonable to hope for a p for which the error incurred by p is comparable to the inherent error β contributed by the misspecified frame V. The main result of this section is to show that RealignPolynomial can find such a p given V:

Theorem 2.6.1. *There are absolute constants $c_4, c_5, c_6, c_7, c_8 > 0$ such that the following holds for any $\underline{\varepsilon}, \delta > 0$. Let $V \in St^n_r$, and let $(\mathbf{c}^*, V^*)$ be the realization of $\mathcal{D}$ for which $d_P(V, V^*) = \|V - V^*\|_F$. Suppose $d_P(V, V^*) \le (c_9 \cdot dr^3)^{-(d+1)/2}$.*

Define $\mathbf{c}^{(T)} = $ RealignPolynomial$(\mathcal{D}, V, \underline{\varepsilon}, \delta)$, where in the specification of RealignPolynomial *we take*

$$\eta_{\mathsf{coef}} \triangleq \left(c_6 r d^4\right)^{d+1} \tag{2.9}$$

$$T \triangleq c_5 \cdot \left(c_6 r d^4\right)^{d+1} \cdot \log(1/\underline{\varepsilon}). \tag{2.10}$$

$$B \triangleq (c_8 \cdot \log(T/\delta))^{2d}. \tag{2.11}$$

Then with probability at least $1 - \delta$, we have that

$$\|\mathbf{c}^{(T)} - \mathbf{c}^*\|_2 \le \left(1 + c_7 \cdot (c_6 d r^4)^{-(d+1)/2}\right) \cdot \{\underline{\varepsilon} \vee d_P(V, V^*)\}. \tag{2.12}$$

Furthermore, RealignPolynomial *requires sample complexity*

$$N \triangleq O(B \cdot T) = \mathrm{poly}\left(\log(1/\delta), r, d, \log\log(1/\underline{\varepsilon})\right)^d \cdot \log(1/\underline{\varepsilon})$$

and runs in time $n \cdot r^{O(d)} \cdot N$.

Before turning to the proof, we set some conventions. Henceforth, fix any V, V^* satisfying the hypotheses of Theorem 2.6.1. Given coefficients $\mathbf{c}$ corresponding to the r-variate polynomial p, define $\boldsymbol{\delta}_{\mathbf{c}} \triangleq p_* - p$. In light of (2.12), it will be convenient in our analysis to quantify, for an iterate $\mathbf{c}^{(t)}$, the extent to which $\|\mathbf{c}^{(t)} - \mathbf{c}^*\|_2$ differs from $d_P(V, V^*)$ via the (unknown) parameter

$$\rho_{\mathbf{c}^{(t)}} \triangleq \frac{d_P(V, V^*)}{\|\mathbf{c}^{(t)} - \mathbf{c}^*\|_2}.$$

For both $\boldsymbol{\delta}_{\mathbf{c}}$ and $\rho_{\mathbf{c}}$, we will sometimes omit the subscript when the context is clear.

Note that we would like the eventual output $\mathbf{c}^{(T)}$ of RealignPolynomial to have large

ρ. The proof of Theorem 2.6.1 thus comes in two parts: 1) when $\rho_{\mathbf{c}^{(t)}}$ is small, the next $\rho_{\mathbf{c}^{(t+1)}}$ is larger by some margin, 2) when $\rho_{\mathbf{c}^{(t)}}$ is large, $\rho_{\mathbf{c}^{(t+1)}}$ may be smaller but will still be no smaller than the bound we are targeting in (2.12). Formally:

Theorem 2.6.2. *Suppose $d_P(V, V^*) \le O(dr^3)^{-(d+1)/2}$. For any $\delta > 0$, let $\mathbf{c}$ be an iterate in the execution of* RealignPolynomial, *and let $\mathbf{c}'$ be the next iterate, given by*

$$c' \triangleq c - \frac{1}{B}\sum_{i=0}^{B-1} \Delta^{\Theta, x_i}_{\mathsf{coef}}$$

as defined in (2.6) for iid samples $(x^0, y^0), ..., (x^{B-1}, y^{B-1}) \sim \mathcal{D}$. If $\eta_{\mathsf{coef}} \triangleq \Theta(dr^4)^{-d-1}$, then with probability at least $1-\delta$ over the samples $\{(x^i, y^i)\}_{i\in[B]}$,

1. *If $\rho_{\mathbf{c}} \le 1$, then $\rho_{\mathbf{c}'} \ge (1 + \Omega(dr^4)^{-d-1}) \cdot \rho_{\mathbf{c}}$.*

2. *If $\rho_{\mathbf{c}} \ge 1$ then $\rho_{\mathbf{c}'} \ge 1 - O(dr^4)^{-(d+1)/2}$.*

We quickly verify that Theorem 2.6.2 implies Theorem 2.6.1.

Proof of Theorem 2.6.1. Take any iterate $\mathbf{c}^{(t)}$ in the execution of RealignPolynomial. Taking δ to be $1/T$ times the error probability in Theorem 2.6.2, we have by a union bound over all T iterations of RealignPolynomial that with probability at least $1-\delta$,

$$\rho_{\mathbf{c}^{(t+1)}} \ge \left\{1 - O(dr^4)^{-(d+1)/2}\right\} \wedge \left\{\rho_{\mathbf{c}^{(t)}} \cdot (1 + \Omega(dr^4)^{-d-1})\right\},$$

for every $0 \le t < T$, which can be unrolled to give

$$\rho_{\mathbf{c}^{(T)}} \ge \left\{1 - O(dr^4)^{-(d+1)/2}\right\} \wedge \left\{\rho_{\mathbf{c}^{(0)}} \cdot (1 + \Omega(dr^4)^{-d-1})^T\right\}.$$

We can rewrite this inequality as

$$\|\mathbf{c}^{(t)} - \mathbf{c}^*\|_2 \le \left\{\frac{d_P(V, V^*)}{1 - O(dr^4)^{-(d+1)/2}}\right\} \vee \left\{\|\mathbf{c}^{(0)} - \mathbf{c}^*\|_2 \cdot (1 + \Omega(dr^4)^{-d-1})^{-T}\right\}.$$

As we are initializing $\mathbf{c}^{(0)} = \mathbf{0}$, we have that $\|\mathbf{c}^{(0)} - \mathbf{c}^*\|_2 = \|\mathbf{c}^*\|_2 \le r$. The theorem follows from taking $T = \Theta(dr^4)^{d+1} \cdot \log(r/\underline{\varepsilon}) = \Theta(dr^4)^{d+1} \cdot \log(1/\underline{\varepsilon})$. □

As Theorem 2.6.2 suggests, we just need to analyze RealignPolynomial on a per-iterate basis. Henceforth, fix an iterate $\mathbf{c}$; we will sometimes refer to the pair $(\mathbf{c}, V)$ as Θ. Let $(x^0, y^0), ..., (x^{B-1}, y^{B-1}) \sim \mathcal{D}$ be the batch of samples drawn for the next iteration of RealignPolynomial.

We first show that it suffices to prove that with high probability, the step $-\frac{1}{B}\sum_{i=0}^{B-1}\Delta^{x^i}_{\mathsf{coef}}$ is both 1) correlated with the direction $\mathbf{c}-\mathbf{c}^*$ in which we want to move, and 2) not too large. 1) and 2) can be interpreted respectively as curvature and smoothness of the gradient of the empirical risk in a neighborhood of our current iterate. Quantitatively, we claim that it suffices to show

Lemma 2.6.3 (Local Curvature with High Probability). *For any $\delta > 0$ and $\gamma > 0$, if $B = \Omega(\log(1/\delta))^{2d} \cdot \gamma^{-2}$, then we have that*

$$\frac{1}{B}\sum_{i=0}^{B-1}\left\langle \Delta^{x^t}_{\mathsf{coef}}, \mathbf{c}-\mathbf{c}^*\right\rangle \geq \upsilon^{\mathsf{cu}}_{\mathbf{c}} \cdot \eta_{\mathsf{coef}} \cdot \|\mathbf{c}-\mathbf{c}^*\|_2^2 \tag{2.13}$$

for

$$\upsilon^{\mathsf{cu}}_{\mathbf{c}} \triangleq 1 - \gamma\rho_{\mathbf{c}} - \|\mathbf{c}-\mathbf{c}^*\|_2 \cdot \left(O(r^{3/2}d)\cdot\rho_{\mathbf{c}}^2 + O(dr^3)^{(d+1)/2}\cdot\rho_{\mathbf{c}}(1+\rho_{\mathbf{c}})\right)$$

with probability at least $1-\delta$.

Lemma 2.6.4 (Local Smoothness With High Probability). *For any $\delta > 0$, if $B = \Omega(\log(1/\delta))^{2d}$, then we have that*

$$\left\|\frac{1}{B}\sum_{i=0}^{B-1}\Delta^{x_i}_{\mathsf{coef}}\right\|_2^2 \leq \upsilon^{\mathsf{sm}}_{\mathbf{c}}\cdot\eta^2_{\mathsf{coef}}\|\mathbf{c}-\mathbf{c}^*\|_2^2 \quad \textit{for} \quad \upsilon^{\mathsf{sm}}_{\mathbf{c}} \triangleq O(dr^4)^{d+1}\cdot(1\vee\rho^2_{\mathbf{c}}).$$

with probability at least $1-\delta$.

We verify that Lemmas 2.6.3 and 2.6.4 are enough to prove Theorem 2.6.2.

Proof of Theorem 2.6.2. (Part 1) By (2.6) we have

$$\|\mathbf{c}'-\mathbf{c}^*\|_2^2 - \|\mathbf{c}-\mathbf{c}^*\|_2^2 = \left\|\frac{1}{B}\sum_{i=0}^{B-1}\Delta^{x^i}_{\mathsf{coef}}\right\|_2^2 - 2\left\langle\frac{1}{B}\sum_{i=0}^{B-1}\Delta^{x^i}_{\mathsf{coef}}, \mathbf{c}-\mathbf{c}^*\right\rangle.$$

If the events of Lemmas 2.6.3 and 2.6.4 occur, then we get that

$$\|\mathbf{c}' - \mathbf{c}^*\|_2^2 - \|\mathbf{c} - \mathbf{c}^*\|_2^2 \leq \|\mathbf{c} - \mathbf{c}^*\|_2^2 \cdot \left(\eta_{\mathsf{coef}} \upsilon_{\mathbf{c}}^{\mathsf{cu}} - \eta_{\mathsf{coef}}^2 \upsilon_{\mathbf{c}}^{\mathsf{sm}}\right),$$

If $\rho_{\mathbf{c}} \leq 1$, then we have that

$$\upsilon_{\mathbf{c}}^{\mathsf{cu}} \geq 1 - \gamma - \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot O(\rho_{\mathbf{c}}) \cdot O(dr^3)^{(d+1)/2} = 1 - \gamma - O(d_P(V, V^*)) \cdot O(dr^3)^{(d+1)/2},$$

so if we take $\gamma = 1/4$ and $d_P(V, V^*)_2 \leq O(dr^3)^{-(d+1)/2}$, then we ensure that $\upsilon_{\mathbf{c}}^{\mathsf{cu}} \geq 1/2$. Additionally, $\rho_{\mathbf{c}} \leq 1$ implies that $\upsilon_{\mathbf{c}}^{\mathsf{sm}} = O(dr^4)^{d+1}$. So if we take $\eta_{\mathsf{coef}} = \Theta(dr^4)^{-d-1}$, we conclude that

$$\|\mathbf{c}' - \mathbf{c}^*\|_2^2 \leq (1 - \eta_{\mathsf{coef}}/3) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^2 \iff \rho_{\mathbf{c}'} \geq \rho_{\mathbf{c}} \cdot (1 - \eta_{\mathsf{coef}}/3)^{-1/2}$$

(Part 2) By triangle inequality,

$$\|\mathbf{c}' - \mathbf{c}^*\|_2 \leq \|\mathbf{c} - \mathbf{c}^*\|_2 + \left\| \frac{1}{B} \sum_{i=0}^{B-1} \Delta_{\mathsf{coef}}^{x^i} \right\|_2.$$

If Lemma 2.6.4 occurs, then we get that

$$\|\mathbf{c}' - \mathbf{c}^*\|_2 \leq \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot (1 + \eta_{\mathsf{coef}} \cdot \sqrt{\upsilon_{\mathbf{c}}^{\mathsf{sm}}}) = \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot \left(1 + \eta_{\mathsf{coef}} \cdot O(dr^4)^{(d+1)/2} \cdot \rho_{\mathbf{c}}\right),$$

or equivalently,

$$\rho_{\mathbf{c}'} \geq \rho_{\mathbf{c}} \cdot \left(1 + \eta_{\mathsf{coef}} \cdot O(dr^4)^{(d+1)/2} \cdot \rho_{\mathbf{c}}\right)^{-1}. \tag{2.14}$$

For our choice of $\eta_{\mathsf{coef}} = \Theta(dr^4)^{-d-1}$, note that the quantity on the right-hand side of (2.14), as a function of $\rho_{\mathbf{c}}$, has minimum value $\left(1 + O(dr^4)^{-(d+1)/2}\right)^{-1}$ over $\rho_{\mathbf{c}} \in [1, \infty)$, attained by $\rho_{\mathbf{c}} = 1$, from which Part 2 of the theorem follows. □

We now proceed to show local curvature and smoothness.

2.6.1 Local Smoothness

In this section we show Lemma 2.6.4.

First, by Jensen's,

$$\left\|\frac{1}{B}\sum_{i=0}^{B-1}\Delta^{x_i}_{\mathsf{coef}}\right\|_2^2 \le \frac{1}{B}\sum_{i=0}^{B-1}\|\Delta^{x^i}_{\mathsf{coef}}\|_2^2,$$

so to show Lemma 2.6.4 it suffices to bound the expectation and variance of the random variable $\|\Delta^x_{\mathsf{coef}}\|_2^2$ with respect to $x \sim \mathcal{N}(0, \mathrm{Id}_n)$ and invoke Lemma 1.3.16.

We will need the following helper lemma which is a straightforward consequence of Lemma 2.5.4 and whose proof we defer to Appendix 2.10.1.

Lemma 2.6.5. *For any* $\Theta = (\mathbf{c}, V)$ *and* $\Theta^* = (\mathbf{c}^*, V^*)$, $\mathbb{E}[(F_x(\Theta)-F_x(\Theta^*))^4]^{1/2} \le O(dr^3)^{d+1}\cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2$.

We now use this to bound the expectation and variance of $\|\Delta^x_{\mathsf{coef}}\|_2^2$.

Lemma 2.6.6. $\mathbb{E}[\|\Delta^x_{\mathsf{coef}}\|_2^2] \le \eta^2_{\mathsf{coef}} \cdot O(dr^4)^{d+1} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2$.

Proof. By Cauchy-Schwarz,

$$\begin{aligned}\frac{1}{4\eta^2_{\mathsf{coef}}}\mathbb{E}\left[\|\Delta^x_{\mathsf{coef}}\|_2^2\right] &\le \mathbb{E}\left[(F_x(\Theta) - F_x(\Theta^*))^4\right]^{1/2} \cdot \mathbb{E}\left[\left(\sum_I \phi_I(V^\top x)^2\right)^2\right]^{1/2} \\ &\le O(dr^4)^{d+1} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2,\end{aligned}$$

where the second step follows by Lemma 2.6.5 and Lemma 2.3.11. □

Lemma 2.6.7. $\mathbb{E}[\|\Delta^x_{\mathsf{coef}}\|_2^4] \le \eta^4_{\mathsf{coef}} \cdot O(dr^4)^{2d+2} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^4$.

Proof. Note that $(F_x(\Theta) - F_x(\Theta^*))^2$ and $\sum_I \phi_I(V^\top x)^2$ are degree-$2d$ polynomials in x. So by Cauchy-Schwarz,

$$\begin{aligned}\frac{1}{16\eta^4_{\mathsf{coef}}}\mathbb{E}\left[\|\Delta^x_{\mathsf{coef}}\|_2^4\right] &\le \mathbb{E}\left[(F_x(\Theta) - F_x(\Theta^*))^8\right]^{1/2} \cdot \mathbb{E}\left[\left(\sum_I \phi_I(V^\top x)^2\right)^4\right]^{1/2} \\ &\le 3^{4d} \cdot \mathbb{E}\left[(F_x(\Theta) - F_x(\Theta^*))^4\right] \cdot \mathbb{E}\left[\left(\sum_I \phi_I(V^\top x)^2\right)^2\right]\end{aligned}$$

$$\leq O(dr^4)^{2d+2} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^4 ,$$

where the second step follows by Fact 1.3.15, and the third step follows by Lemmas 2.6.5 and 2.3.11. □

We are now ready to prove Lemma 2.6.4.

Proof of Lemma 2.6.4. Note that $\|\Delta^x_{\mathsf{coef}}\|_2^2$ is a polynomial of degree $2d$ in x. So by Lemma 1.3.16, Lemma 2.6.6, and Lemma 2.6.7, we see that

$$\frac{1}{B}\sum_{i=0}^{B-1}\|\Delta^{x^i}\|_2^2 \leq \eta_{\mathsf{coef}}^2 \cdot O(dr^4)^{d+1} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2 \cdot \left(1 + \frac{1}{\sqrt{B}} \cdot O(\log(1/\delta))^d\right),$$

so the lemma follows by recalling that $\|V - V^*\|_F = d_P(V, V^*)$ so that

$$(\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2 \leq 4\|\mathbf{c} - \mathbf{c}^*\|_2^2 \cdot (1 \vee \rho_{\mathbf{c}}^2)$$

and taking $B = \Omega(\log(1/\delta))^{2d}$. □

We note that this is one of the first of many places where the fact that one cannot obtain a $\mathbf{c}$ whose error is much smaller than the "misspecification error" $d_P(V, V^*)$ incurred by the subspace V manifests: here, our bounds on the magnitudes of the gradient steps $\|\Delta^x_{\mathsf{coef}}\|$ inherently depend on $d_P(V, V^*)$, yet we require that the gradient steps have norm bounded by $\|\mathbf{c} - \mathbf{c}^*\|$.

2.6.2 Local Curvature

We begin by outlining our argument for proving Lemma 2.6.3. It will be helpful to first decompose $\langle \Delta_{\mathsf{coef}}, \mathbf{c} - \mathbf{c}^* \rangle$ into "dominant" and "non-dominant" terms.

Proposition 2.6.8. *For every monomial index I and any $x \in \mathbb{R}^n$, let*

$$({\Delta'_{\mathsf{coef}}}^x)_I \triangleq -2\eta_{\mathsf{coef}} \cdot \langle \nabla F_x(\Theta), \Theta^* - \Theta \rangle \cdot \phi_I(V^\top x) \quad \textit{and} \quad ({\Delta''_{\mathsf{coef}}}^x)_I \triangleq -2\eta_{\mathsf{coef}} \cdot \mathfrak{R}^x \cdot \phi_I(V^\top x) \; \forall \; I.$$

Then $\Delta^x_{\mathsf{coef}} = {\Delta'_{\mathsf{coef}}}^x + {\Delta''_{\mathsf{coef}}}^x$.

Proof. ${\Delta'_{\mathsf{coef}}}^x$ and ${\Delta''_{\mathsf{coef}}}^x$ correspond to the first-order and higher-order terms in the Taylor expansion of Δ^x_{coef}. Concretely, recall that

$$(\Delta^x_{\mathsf{coef}})_I = 2\eta_{\mathsf{coef}} \cdot (F_x(\Theta) - F_x(\Theta^*)) \cdot \phi_I(V^\top x).$$

We can decompose Δ^x_{coef} by Taylor expanding the factor $F_x(\Theta) - F_x(\Theta^*)$ around $\Theta^* = \Theta$ to get

$$F_x(\Theta^*) - F_x(\Theta) = \langle \nabla F_x(\Theta), \Theta^* - \Theta\rangle + \mathfrak{R}^{\Theta,x} \quad \text{for} \quad \mathfrak{R}^{\Theta,x} \triangleq \sum_{\ell=2}^{d+1} \frac{1}{\ell!}\big\langle \nabla^{[\ell]} F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell}\big\rangle, \tag{2.15}$$

from which the proposition follows. □

Motivated by Proposition 2.6.8, for any $x \in \mathbb{R}^n$ define

$$Y^x \triangleq \langle {\Delta'_{\mathsf{coef}}}^x, \mathbf{c} - \mathbf{c}^*\rangle, \quad \text{and} \quad E^x \triangleq \langle {\Delta''_{\mathsf{coef}}}^x, \mathbf{c} - \mathbf{c}^*\rangle.$$

To show Lemma 2.6.3, we will show that the random variables $\frac{1}{B}\sum_{i=0}^{B-1} Y^{x^i}$ and $\frac{1}{B}\sum_{i=0}^{B-1} E^{x^i}$ are respectively large and negligible with high probability. Eventually we will invoke the concentration inequalities of Lemmas 1.3.16 and 1.3.33 to control them, so we will compute the expectations (Section 2.6.2) and variances (Section 2.6.2) of their summands next.

Local Curvature in Expectation

In this section we give bounds for $\mu_Y \triangleq \mathbb{E}_x[Y^x]$ and $\mu_E \triangleq \mathbb{E}_x[E^x]$ in the following two lemmas. Throughout this section, we will omit the superscript x when the context is clear.

Lemma 2.6.9. $\mu_Y \geq 2\eta_{\mathsf{coef}} \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot \big(\|\mathbf{c} - \mathbf{c}^*\|_2 - O(r^{3/2}d) \cdot d_P(V, V^*)^2\big)$.

Lemma 2.6.10. $|\mu_E| \leq 2\eta_{\mathsf{coef}} \cdot O(dr^3)^{(d+1)/2} \cdot d_P(V, V^*) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot (d_P(V, V^*) + \|\mathbf{c} - \mathbf{c}^*\|_2)$.

In this section we will give the proof of Lemma 2.6.9; we will defer the proof of Lemma 2.6.10 to Appendix 2.10.2.

Proof of Lemma 2.6.9. We have that

$$\langle \Delta'_{\mathsf{coef}}, \mathbf{c} - \mathbf{c}^* \rangle = -2\eta_{\mathsf{coef}} \langle \nabla F_x(\Theta), \Theta^* - \Theta \rangle \cdot \delta(V^\top x) \tag{2.16}$$

Writing

$$\begin{aligned} \langle \nabla F_x(\Theta), \Theta^* - \Theta \rangle &= \langle \nabla^{\mathsf{vec}} F_x(\Theta), V^* - V \rangle + \langle \nabla^{\mathsf{coef}} F_x(\Theta), \mathbf{c}^* - \mathbf{c} \rangle \\ &= x^\top (V^* - V)\nabla + \delta(V^\top x) \\ &= x^\top \Pi_V^\perp (V^* - V)\nabla + x^\top \Pi_V (V^* - V)\nabla + \delta(V^\top x) \\ &= x^\top \Pi_V^\perp V^* \nabla + x^\top \Pi_V \cdot (V^* - V)\nabla + \delta(V^\top x), \end{aligned} \tag{2.17}$$

we see that (2.16) is given by $2\eta_{\mathsf{coef}}$ times

$$\underbrace{\big(\delta(V^\top x)\big)^2}_{\textcircled{A}} + \underbrace{\delta(V^\top x) \cdot \big(x^\top \Pi_V (V^* - V)\nabla\big)}_{\textcircled{B}} + \underbrace{\delta(V^\top x) \cdot \big(x^\top \Pi_V^\perp V^* \nabla\big)}_{\textcircled{C}} \tag{2.18}$$

Note that $x^\top \Pi_V$ and $x^\top \Pi_V^\perp$ are independent Gaussian vectors with mean zero and covariances Π_V and $\Pi_V^\perp$ respectively. So we readily conclude that

Observation 2.6.11. *For any V, the expectation of $\textcircled{C}$ with respect to x vanishes.*

The following is also immediate:

Observation 2.6.12. $\mathbb{E}[\textcircled{A}] = \mathbb{E}_{g \sim \mathcal{N}(0, Id_r)}[\delta(g)^2] = \|\mathbf{c} - \mathbf{c}^*\|_2^2$.

We now turn to bounding $\mathbb{E}[\textcircled{B}]$. We will make use of the following helper bound whose proof we defer to Appendix 2.10.3

Proposition 2.6.13. *If $\|V - V^*\| = d_P(V, V^*)$, then*

$$\mathbb{E}_g \left[\big(x^\top \Pi_V (V^* - V) \nabla p(V^\top x)\big)^2 \right]^{1/2} \le d_P(V, V^*)^2 \cdot O(r^{3/2} d).$$

Lemma 2.6.14. $\mathbb{E}[\textcircled{B}] \le O(r^{3/2} d) \cdot d_P(V, V^*)^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2$.

Proof. Note that

$$\begin{aligned}
|\mathbb{E}[\textcircled{B}]| &= \left|\mathbb{E}\left[\delta(V^\top x)\cdot\left(x^\top\Pi_V(V^*-V)\nabla p(V^\top x)\right)\right]\right| \\
&\le \mathbb{E}_g\left[\delta(g)^2\right]^{1/2}\cdot\mathbb{E}_g\left[\left(g^\top V^\top(V^*-V)\nabla p(g)\right)^2\right]^{1/2} \\
&\le \|\mathbf{c}-\mathbf{c}^*\|_2\cdot d_P(V,V^*)^2\cdot O(r^{3/2}d),
\end{aligned}$$

where the second step follows by Cauchy-Schwarz, and the third by Proposition 2.6.13. □

Lemma 2.6.9 now follows from (2.18), Observations 2.6.11 and 2.6.12, and Lemma 2.6.14.

□

Local Curvature with High Probability

In this section, we complete the proof of Lemma 2.6.3 by establishing high-probability bounds for Y^x and E^x. That is, we argue that with high probability, the dominant term given by Y is large and the error from Taylor approximation is small. Specifically, we will show:

Lemma 2.6.15. *For any $\delta > 0$ and $\gamma > 0$, if $B = \Omega(\log(1/\delta))^d \cdot O(\gamma^{-2})$, then*

$$\frac{1}{B}\sum_{i=1}^{B-1} Y^{x^i} \ge \eta_{\mathsf{coef}}\left(\|\mathbf{c}-\mathbf{c}^*\|_2^2 - O(r^{3/2}d)\cdot d_P(V,V^*)^2\cdot\|\mathbf{c}-\mathbf{c}^*\|_2 - \gamma\cdot d_P(V,V^*)\cdot\|\mathbf{c}-\mathbf{c}^*\|_2\right)$$

Lemma 2.6.16. *For any $\delta > 0$, if $B = \Omega(\log(1/\delta))^{2d}$, then*

$$\left|\frac{1}{B}\sum_{i=0}^{B-1} E^{x^i}\right| \le \eta_{\mathsf{coef}}\cdot O(dr^3)^{(d+1)/2}\cdot d_P(V,V^*)\cdot\|\mathbf{c}-\mathbf{c}^*\|_2\cdot\left(d_P(V,V^*)+\|\mathbf{c}-\mathbf{c}^*\|_2\right)$$

We defer their proofs to Appendices 2.10.4 and 2.10.5 respectively. We can finally deduce Lemma 2.7.3, completing the proof of Theorem 2.6.2 and thus Theorem 2.6.1.

Proof of Lemma 2.7.3. By Lemmas 2.6.15 and 2.6.16, and the earlier calculation showing that for any x, $\langle\Delta^x_{\mathsf{coef}}, V-V^*\rangle = Y^x + E^x$, we see that under our choice of B, (2.13) holds with probability $1-3\delta$. By replacing 3δ with δ, and absorbing the constant factors, the lemma follows. □

2.7 Guarantees for SubspaceDescent

Henceforth, fix a set of coefficients $\mathbf{c} \in \mathbb{R}^M$. In contrast with RealignPolynomial, the aim of SubspaceDescent is to take a frame $V^{(0)}$ of a subspace which is somewhat close to the true subspace and refine it to some $V^{(T)}$ which is slightly closer, using only the misspecified coefficients $\mathbf{c}$. It turns out that if the misspecification error of $\mathbf{c}$ is comparable to the subspace distance from $V^{(0)}$ to the true subspace, SubspaceDescent can indeed accomplish this, and this is the main result of this section.

Theorem 2.7.1. *There are absolute constants $c_{10}, c_{11} > 0$ and $c_{12} < 1/10$ such that the following holds for any $\delta > 0$. Let $V^{(0)} \in St^n_r$, and let $(\mathbf{c}^*, V^*)$ be the realization of $\mathcal{D}$ for which $d_P(V, V^*) = \|V - V^*\|_F$. Suppose*

$$d_P(V^{(0)}, V^*) \leq c_{12} \cdot \nu_{\mathsf{cond}} \cdot O(dr^3)^{-d-2}, \tag{2.19}$$

Let $\mathbf{c}$ be a set of coefficients satisfying

$$d_P(V^{(0)}, V^*) \geq \frac{1}{2}\|\mathbf{c} - \mathbf{c}^*\|_2 \tag{2.20}$$

Define $V^{(T)} =$ SubspaceDescent$(\mathcal{D}, V^{(0)}, \mathbf{c}, \delta)$, where in the specification of SubspaceDescent *we take*

$$\eta_{\mathsf{vec}} \triangleq \frac{\nu_{\mathsf{cond}}}{T \cdot n}\left(c_{11} \cdot dr^3 \ln(T/\delta)\right)^{-d-2} \tag{2.21}$$

$$T \triangleq \left(\frac{r}{\nu_{\mathsf{cond}}}\right)^2 \cdot \left(c_{10} \cdot d \cdot \log(1/\delta)\right)^{2c_1 d}. \tag{2.22}$$

Then with probability at least $1 - \delta$, we have that

$$1 - \frac{d_P(V^{(T)}, V^*)^2}{d_P(V^{(0)}, V^*)^2} \geq \frac{\nu_{\mathsf{cond}}}{n} \cdot \mathrm{poly}(\ln(1/\nu_{\mathsf{cond}}), r, d, \ln(1/\delta))^{-d}.$$

Furthermore, SubspaceDescent *draws $N \triangleq O(T)$ samples and runs in time $n \cdot r^{O(d)} \cdot N$.*

Henceforth, let $\delta, V^{(0)}, V^*, \mathbf{c}, \mathbf{c}^*, T, \eta_{\mathsf{vec}}$ satisfy the hypotheses of Theorem 2.7.1.

As discussed in Section 2.2.2, a single execution of SubspaceDescent should be thought of as a single step of stochastic gradient descent over a batch of size T. The only difference lies in the fact that the empirical risk we work with in each iteration of SubspaceDescent is slightly different, as our subspace estimate $V^{(t)}$ continues to update by a small amount. So just as we analyzed the individual steps of RealignPolynomial in Lemma 2.6.2 via local curvature and smoothness estimates, we would like to do the same for an entire execution of SubspaceDescent. That is, we want to show that with high probability, the steps $-\Delta^{\Theta^t,x^t}_{\mathsf{vec}}$ are 1) bounded, and 2) each correlated with the direction $V^* - V^{(t)}$ in which we want to move. Quantitatively, we claim that it suffices to show

Lemma 2.7.2 (Local Smoothness With High Probability)**.**

$$\|V^{(0)} - V^{(T)}\|_F^2 \le \eta_{\mathsf{vec}}^2 \cdot O(dr^3 \ln(T/\delta))^{d+2} \cdot O(n) \cdot d_P(V^{(0)}, V^*)^2.$$

with probability at least $1-\delta$*.*

Lemma 2.7.3 (Local Curvature with High Probability)**.**

$$\sum_{t=0}^{T-1} \left\langle \Delta^{\Theta^{(t)},x^t}_{\mathsf{vec}}, V^{(t)} - V^* \right\rangle \ge T \cdot \eta_{\mathsf{vec}} \cdot (\nu_{\mathsf{cond}}/4) \cdot d_P(V^{(0)}, V^*)^2$$

with probability at least $1-\delta$*.*

We verify that Lemmas 2.7.3 and 2.7.2 are enough to prove Theorem 2.7.1.

Proof of Theorem 2.7.1. For every $0 \le t < T$, we have

$$\|V^{(t+1)} - V^*\|_F^2 - \|V^{(t)} - V^*\|_F^2 = \|\Delta^{\Theta^{(t)},x^t}_{\mathsf{vec}}\|_F^2 - 2\left\langle \Delta^{\Theta^{(t)},x^t}_{\mathsf{vec}}, V^{(t)} - V^* \right\rangle. \tag{2.23}$$

If the event of Lemma 2.7.3 holds, then

$$\sum_{t=0}^{T-1} \left\langle \Delta^{\Theta^{(t)},x^t}_{\mathsf{vec}}, V^{(t)} - V^* \right\rangle \ge T \cdot (\nu_{\mathsf{cond}}/4) \cdot \eta_{\mathsf{vec}} \cdot d_P(V^{(0)}, V^*)^2.$$

If the event of Lemma 2.7.2 holds, then

$$\begin{aligned}\sum_{t=0}^{T-1} \|\Delta^{\Theta^{(t)},x^t}_{\mathsf{vec}}\|_F^2 &\le T \cdot \eta_{\mathsf{vec}}^2 \cdot O(dr^3 \ln(T/\delta))^{d+2} \cdot O(n) \cdot d_P(V^{(0)}, V^*)^2 \\ &\le O\left(\nu_{\mathsf{cond}} \cdot \eta_{\mathsf{vec}} \cdot d_P(V^{(0)}, V^*)^2\right).\end{aligned}$$

where the last step follows by the choice of η_{vec} in (2.21), and the constant factor in the last expression can be made arbitrarily small. By summing (2.23) over t, telescoping, and recalling that $\|V^{(0)} - V^*\|_F^2 = d_P(V^{(0)}, V^*)^2$, we conclude that

$$\|V^{(T)} - V^*\|_2^2 - d_P(V^{(0)}, V^*)^2 \leq -T \cdot (\nu_{\mathsf{cond}}/5) \cdot \eta_{\mathsf{vec}} \cdot d_P(V^{(0)}, V^*)^2,$$

from which we get, because $d_P(V^{(T)}, V^*) \leq \|V^{(T)} - V^*\|_F$, that

$$1 - \frac{d_P(V^{(T)}, V^*)^2}{d_P(V^{(0)}, V^*)^2} \geq T \cdot (\nu_{\mathsf{cond}}/5) \cdot \eta_{\mathsf{vec}}.$$

The claim follows by substituting the choice of η_{vec} and T in (2.21) and (2.22). □

We now proceed to show Lemma 2.7.2 and 2.7.3.

2.7.1 Local Smoothness

In this section we establish Lemma 2.7.2. We also show that $d_P(V^{(t)}, V^*)$ does not change much, both in expectation (Lemma 2.7.7) and with high probability (Lemma 2.7.6), as t varies. While we have already seen that Lemma 2.7.2 is needed to prove Theorem 2.7.1, Lemmas 2.7.6 and 2.7.7 will be crucial to our arguments in later sections, where we argue that at each step t we make progress scaling with the distance $d_P(V^{(t)}, V^*)$ and thus need that this distance is comparable to the initial distance $d_P(V^{(0)}, V^*)$.

For a fixed Θ, we will first show a high-probability bound on the norm of $\Delta_{\mathsf{vec}}^{\Theta,x}$, that is, we bound the size of the step made in a single iteration inside SubspaceDescent.

Where the context is clear, we will suppress superscript Θ, x. Then very naively, using the inequalities $1 - \cos(x) \leq x$ and $|\sin(x)| \leq x$ for all $x \geq 0$, we have

$$\|\Delta_{\mathsf{vec}}\|_F \leq (1 - \cos(\sigma\eta_{\mathsf{vec}}))(2\sqrt{r}) + |\sin(\sigma\eta_{\mathsf{vec}})| \leq 2\sqrt{r} \cdot \sigma\eta_{\mathsf{vec}} + \sigma\eta_{\mathsf{vec}} \leq 3\sqrt{r} \cdot \sigma\eta_{\mathsf{vec}}. \quad (2.24)$$

We first bound the moments of σ^2.

Lemma 2.7.4. *For all integers $q \geq 1$, $\mathbb{E}[\sigma^{2q}]^{1/q} \leq O(nrd) \cdot O(q^2 dr^3)^{d+2} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2$.*

Proof. Recall that $\sigma = 2(F_x(\Theta) - F_x(\Theta^*)) \cdot \|\Pi_V^\perp x\|_2 \cdot \|\nabla p(V^\top x)\|_2$. So by Cauchy-Schwarz,

$$\mathbb{E}[\sigma^{2q}]^{1/q} \le 4\mathbb{E}[(F_x(\Theta) - F_x(\Theta^*))^{4q}]^{1/2q} \cdot \mathbb{E}\left[\|\Pi_V^\perp x\|_2^{4q} \cdot \|\nabla p(V^\top x)\|_2^{4q}\right]^{1/2q}. \tag{2.25}$$

The second factor in (2.25) is simply

$$\begin{aligned}
&\mathbb{E}_{g'\sim\mathcal{N}(0,\Pi_V^\perp)}[\|g'\|_2^{4q}]^{1/2q} \cdot \mathbb{E}_{g\sim\mathcal{N}(0,Id_r)}[\|\nabla p(g)\|_2^{4q}]^{1/2q} \\
&\quad \le ((2q-1)\cdot(n-r+1)) \cdot \left(rd \cdot (4q-1)^d \cdot \mathbb{V}[p]\right) \\
&\quad \le O(n) \cdot qrd \cdot (4q)^d \cdot \mathbb{V}[p] \\
&\quad \le O(n) \cdot rd \cdot (4q)^{d+1},
\end{aligned}$$

where in the first step we used Corollary 1.3.17 and Lemma 2.3.8, and in the last step we used Fact 2.3.2 and triangle inequality to bound $\mathbb{V}[p] = O(1)$.

For the first factor in (2.25), we have that

$$\begin{aligned}
\mathbb{E}\left[(F_x(\Theta) - F_x(\Theta^*))^{4q}\right]^{1/2q} &\le (2q-1)^d \cdot \mathbb{E}\left[(F_x(\Theta) - F_x(\Theta^*))^4\right]^{1/2} \\
&\le O(qdr^3)^{d+1} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2
\end{aligned}$$

by Fact 1.3.15 and Lemma 2.6.5 respectively, from which the claim follows. □

As a result, the random variable σ^2 enjoys sub-Weibull-type concentration.

Corollary 2.7.5. *For any $0 < \delta' < 1$, let $\tau = \Omega(\ln(1/\delta'))^{d+2}$. Then*

$$Pr\left[\sigma^2 \ge \tau \cdot \Omega(n) \cdot \Omega(dr^3)^{d+2} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2\right] \le \delta'.$$

Proof. Let $\gamma \triangleq n \cdot O(rd^3)^{d+2} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2$. We wish to apply Lemma 1.3.16 to σ^2, which is a degree-$4d$ polynomial in x. By Lemma 2.7.4 above, $\mathbb{E}[\sigma^2] \le O(\gamma)$ and $\mathbb{V}[\sigma^2] \le \mathbb{E}[\sigma^4] \le O(\gamma^2)$. By Lemma 1.3.16 specialized to $T = 1$,

$$\Pr\left[\sigma^2 \ge O(\log(1/\delta'))^{2d} \cdot \gamma\right] \le \delta',$$

from which the lemma follows. □

From (2.24) we conclude that for any $0 < \delta < 1$,

$$\|\Delta_{\mathsf{vec}}\|_F \le 3\sqrt{r}\cdot\eta_{\mathsf{vec}}\cdot O(\ln(1/\delta))^{(d+2)/2}\cdot O(\sqrt{n})\cdot O(dr^3)^{(d+2)/2}\cdot(\|V-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2) \tag{2.26}$$

with probability at least $1-\delta$.

Now consider the sequence of iterates $\{\Theta^{(t)}\}_{0\le t\le T}$ in SubspaceDescent. In this subsection alone, for convenience define

$$\alpha \triangleq 3\sqrt{r}\cdot\eta_{\mathsf{vec}}\cdot O(\ln(1/\delta))^{(d+2)/2}\cdot O(\sqrt{n})\cdot O(dr^3)^{(d+2)/2}$$

For every $0 \le t < T$, let $\mathcal{E}_t$ be the event that (2.26) holds for $\Delta^{\Theta^{(t)},x^t}_{\mathsf{vec}}$, that is, that $\|\Delta^{\Theta^{(t)},x^t}_{\mathsf{vec}}\|_F \le \alpha(\|V^{(t)}-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2)$. If $\mathcal{E}_t$ held for every t, then by triangle inequality and induction, we would have that for every $0\le t<T$,

$$\begin{aligned}
\|\Delta^{\Theta^{(t)},x^t}_{\mathsf{vec}}\|_F &\le \alpha\left(\|V^{(0)}-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2+\sum_{s=0}^{t-1}\|\Delta^{\Theta^{(s)},x^s}_{\mathsf{vec}}\|_F\right)\\
&\le \alpha(1+\alpha)^t\left(\|V^{(0)}-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2\right)\\
&= \alpha(1+\alpha)^t\left(d_P(V^{(0)},V^*)+\|\mathbf{c}-\mathbf{c}^*\|_2\right)\\
&\le 3\alpha(1+\alpha)^t\cdot d_P(V^{(0)},V^*),
\end{aligned}$$

where the last step follows by (2.20). So

$$\sum_{t=0}^{T-1}\|\Delta^{\Theta^{(t)},x^t}_{\mathsf{vec}}\|_F \le 3\left((1+\alpha)^T-1\right)\cdot d_P(V^{(0)},V^*). \tag{2.27}$$

Taking δ' in Corollary 2.7.5 to be δ/T and applying a union bound, we deduce by monotonicity of L_p norms that Lemma 2.7.2 holds for our choice of η_{vec}, T. We also deduce the following crude bound.

Lemma 2.7.6. *$\|V^{(t)}-V^*\|_F \in [0.9,1.1]\cdot d_P(V^{(0)},V^*)$ for every $0\le t\le T$ with probability at least $1-\delta$.*

This modest level of control over how much the distance to the true subspace fluctuates over the course of SubspaceDescent will be sufficient for our subsequent analysis.

We pause to note that the assumption that the "misspecification error" $\|\mathbf{c}-\mathbf{c}^*\|_2$ incurred by the coefficients $\mathbf{c}$ must, by (2.20), be small relative to the subspace distance error incurred by the initial subspace $V^{(0)}$ is crucial here. Indeed, our bounds for the moments of σ^2, i.e. the moments of the size of the gradient steps, inherently scale with $\|\mathbf{c}-\mathbf{c}^*\|$, yet we need local smoothness in the sense that the gradient steps have norm comparable to $d_P(V^{(0)}, V^*)$.

Lastly, it will be useful to establish bounds on the moments of $\|V^{(t)} - V^*\|_F$ for each t.

Lemma 2.7.7. *For any absolute, integer-valued constant $q \geq 1$, $\mathbb{E}\left[\|V^{(t)} - V^*\|_F^q\right] \leq 1.1 \cdot d_P(V^{(0)}, V^*)^q$ for every $0 \leq t < T$, where the expectation is in the randomness of the samples $x^0, ..., x^{T-1}$ drawn in* SubspaceDescent.

We defer the proof of this to Appendix 2.11.2.

2.7.2 Local Curvature

We begin by outlining our argument for proving Lemma 2.7.3. As with the proof of Lemma 2.6.3 for RealignPolynomial, it will be helpful to first decompose $\langle \Delta_{\mathsf{vec}}, V - V^*\rangle$ into "dominant" and "non-dominant" terms. Here the "non-dominant" terms will be more complicated because of the trigonometric corrections associated with geodesic gradient descent.

Proposition 2.7.8. *For any Θ, x, define*

$${\Delta'_{\mathsf{vec}}}^{\Theta,x} \triangleq -2\eta_{\mathsf{vec}} \cdot \langle \nabla F_x(\Theta), \Theta^* - \Theta\rangle \cdot \Pi_V^\perp \cdot x \cdot (\nabla^{\Theta,x})^\top \quad \textit{and} \quad {\Delta''_{\mathsf{vec}}}^{\Theta,x} \triangleq -2\eta_{\mathsf{vec}} \cdot \mathfrak{R}^{\Theta,x} \cdot \Pi_V^\perp \cdot x \cdot (\nabla^{\Theta,x})^\top$$

and also

$$\begin{aligned}\mathcal{E}^{\Theta,x} &\triangleq \Delta_{\mathsf{vec}}^{\Theta,x} - {\Delta'_{\mathsf{vec}}}^{\Theta,x} - {\Delta''_{\mathsf{vec}}}^{\Theta,x} \\ &= \left(\cos(\sigma^{\Theta,x}\eta_{\mathsf{vec}}) - 1\right) V \cdot \widehat{\nabla}^{\Theta,x} (\widehat{\nabla}^{\Theta,x})^\top + \left(\sin(\sigma^{\Theta,x}\eta_{\mathsf{vec}}) - \sigma^{\Theta,x}\eta_{\mathsf{vec}}\right) \widehat{h}^{\Theta,x} (\widehat{\nabla}^{\Theta,x})^\top.\end{aligned}$$

Then $\Delta_{\mathsf{vec}}^{\Theta,x} = {\Delta'_{\mathsf{vec}}}^{\Theta,x} + {\Delta''_{\mathsf{vec}}}^{\Theta,x} = \mathcal{E}^{\Theta,x}$.

Proof. $\tilde{\Delta}^{\Theta,x}_{\mathsf{vec}} \triangleq {\Delta'_{\mathsf{vec}}}^{\Theta,x} + {\Delta''_{\mathsf{vec}}}^{\Theta,x}$ is the lowest-order term in the Taylor expansion of $\Delta^{\Theta,x}_{\mathsf{vec}}$ around $\eta_{\mathsf{vec}} = 0$, given by

$$\tilde{\Delta}^{\Theta,x}_{\mathsf{vec}} \triangleq \eta_{\mathsf{vec}} \cdot h^{\Theta,x}(\nabla^{\Theta,x})^\top.$$

Recalling the factor $F_x(\Theta) - F_x(\Theta^*)$ in the definition of h in (2.7), we Taylor expand around $\Theta^* = \Theta$ to get (2.15) from Section 2.6 and therefore the decomposition of $\tilde{\Delta}^{\Theta,x}_{\mathsf{vec}}$ into ${\Delta'_{\mathsf{vec}}}^{\Theta,x}$ and ${\Delta''_{\mathsf{vec}}}^{\Theta,x}$. □

$\widehat{\Delta}'_v$ Motivated by Proposition 2.7.8, for any $x \in \mathbb{R}^n$ and $\Theta = (\mathbf{c}, V)$ define

$$X^{\Theta,x} \triangleq \langle (\tilde{\Delta}'_{\mathsf{vec}})^{\Theta,x}, V - V^* \rangle, \qquad E_1^{\Theta,x} \triangleq \langle (\tilde{\Delta}''_{\mathsf{vec}})^{\Theta,x}, V - V^* \rangle, \qquad E_2^{\Theta,x} \triangleq \langle \mathcal{E}^{\Theta,x}, V - V^* \rangle.$$

Consider a sequence of iid samples $(x^0, y^0), ..., (x^{T-1}, y^{T-1}) \sim \mathcal{D}$ and iterates $\Theta^{(0)}, ..., \Theta^{(T-1)}$ in the execution of SubspaceDescent, where each $\Theta^{(t)}$ is given by $\Theta^{(t)} = (\mathbf{c}, V^{(t)})$. To show Lemma 2.7.3, we will show that the random variable $\sum_{t=0}^{T-1} X^{\Theta^{(t)},x^t}$ is large with high probability, while the random variables $\sum_{t=0}^{T-1} E_1^{\Theta^{(t)},x^t}$, and $\sum_{t=0}^{T-1} E_2^{\Theta^{(t)},x^t}$ are negligible with high probability. Eventually, we will invoke the martingale concentration inequalities of Lemmas 2.3.3 and 2.3.4 to control them. Before that, we first need to compute their expectations.

Local Curvature in Expectation- Single Step

In this section we give bounds on the *expected* correlation between the direction in which we would like to move, and a step taken in a *single* iteration in SubspaceDescent.

Given an iterate $\Theta = (\mathbf{c}, V)$, let $\mu_X(\Theta), \mu_{E_1}(\Theta), \mu_{E_2}(\Theta)$ be the expectations $\mathbb{E}[X^{\Theta,x}]$, $\mathbb{E}[E_1^{\Theta,x}]$, $\mathbb{E}[E_2^{\Theta,x}]$ with respect to $x \sim \mathcal{N}(0, \mathrm{Id}_n)$. In this section we will bound these quantities in terms of the distance between Θ and $(\mathbf{c}^*, V^*)$. As usual, we will omit the superscript Θ, x when the context is clear.

Lemma 2.7.9. $\mu_X(\Theta) \geq 2\eta_{\mathsf{vec}} \cdot (\nu_{\mathsf{cond}}/4) \cdot d_P(V, V^*)^2$.

Lemma 2.7.10.

$$|\mu_{E_1}(\Theta)| \leq O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot \|V - V^*\|_F \cdot d_P(V, V^*) \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2).$$

Lemma 2.7.11. *If $\eta_{\mathsf{vec}} \le O(1/n)$, then*

$$|\mu_{E_2}(\Theta)| \le O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{d+2} \cdot \|V - V^*\|_F \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2 .$$

At this point we pause to emphasize that Lemma 2.7.9 is the key reason why we must work with $\mathrm{G}(n, r)$ and not simply with the Euclidean space of $n \times r$ matrices, as Lemma 2.7.9 says that the local curvature with respect to the empirical risk in a neighborhood of a subspace V is dictated solely by its Procrustes distance to V^* rather than by $\|V - V^*\|_F$.

Additionally, note that once again, (2.20) is essential here, to ensure that the expectations from Lemmas 2.7.10 and 2.7.11 of the "non-dominant" terms do not overwhelm the expectation from Lemma 2.7.18 of the "dominant" term, which only depends on $d_P(V, V^*) \sim d_P(V^{(0)}, V^*)$.

We now turn to proving Lemma 2.7.9.

Proof of Lemma 2.7.9. Fix a sample $(x, y) \sim \mathcal{D}$. We have that

$$\begin{aligned} \langle \tilde{\Delta}'_{\mathsf{vec}}, V - V^* \rangle &= -2\eta_{\mathsf{vec}} \langle \nabla F_x(\Theta), \Theta^* - \Theta \rangle \cdot x^\top \Pi_V^\perp (V - V^*) \nabla \\ &= 2\eta_{\mathsf{vec}} \langle \nabla F_x(\Theta), \Theta^* - \Theta \rangle \cdot x^\top \cdot \Pi_V^\perp V^* \cdot \nabla \end{aligned} \tag{2.28}$$

By (2.17) we see that (2.28) is given by $2\eta_{\mathsf{vec}}$ times

$$\underbrace{\left(x^\top \Pi_V^\perp V^* \nabla\right)^2}_{\textcircled{\scriptsize A'}} + \underbrace{\left(x^\top \Pi_V (V^* - V) \nabla\right) \cdot \left(x^\top \Pi_V^\perp V^* \nabla\right)}_{\textcircled{\scriptsize B'}} + \underbrace{\delta(V^\top x) \cdot \left(x^\top \Pi_V^\perp V^* \nabla\right)}_{\textcircled{\scriptsize C'}} . \tag{2.29}$$

As in the proof of Lemma 2.6.9, note that $x^\top \Pi_V$ and $x^\top \Pi_V^\perp$ are independent Gaussan random vectors with mean zero and covariances Π_V and $\Pi_V^\perp$ respectively. So we immediately conclude that

Observation 2.7.12. *For any V, the expectations of $\textcircled{\scriptsize B'}$ and $\textcircled{\scriptsize C'}$ with respect to x vanish.*

We next bound $\mathbb{E}[\textcircled{\scriptsize A'}]$.

Lemma 2.7.13. $(\nu_{\mathsf{cond}}/4) \cdot d_P(V, V^*)^2 \le \mathbb{E}[\textcircled{\scriptsize A'}] \le 4 d_P(V, V^*)^2$.

Proof. Note that

$$\begin{aligned}
\mathbb{E}[\text{\textcircled{A'}}] &= \mathbb{E}\left[\left(x^\top \Pi_V^\perp V^* \nabla\right)^2\right] \\
&= \mathbb{E}_{\substack{h\sim\mathcal{N}(0,\Pi_V) \\ h_\perp\sim\mathcal{N}(0,\Pi_V^\perp)}} \left[\nabla p(V^\top h)^\top V^{*\top} h_\perp h_\perp^\top V^* \nabla p(V^\top h)\right] \\
&= \mathbb{E}_{h\sim\mathcal{N}(0,\Pi_V)} \left[\nabla p(V^\top h)^\top V^{*\top} \Pi_V^\perp V^* \nabla p(V^\top h)\right] \\
&= \mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{Id}_r)} \left[\nabla p(g)^\top \cdot \left(\mathrm{Id} - V^{*\top} V V^\top V^*\right) \cdot \nabla p(g)\right] \\
&= \left\langle \mathbb{E}_g\left[\nabla p(g) \nabla p(g)^\top\right], \mathrm{Id} - V^{*\top} V V^\top V^* \right\rangle \qquad (2.30)
\end{aligned}$$

where we used independence of $h, h_\perp$ in the third step. We wil need the following bound.

Lemma 2.7.14. *If* $\|\mathbf{c} - \mathbf{c}^*\|_2 \le O(r^{-3/2} d^{-1})$*, then we have that*

$$(\nu_{\mathsf{cond}}/2) \cdot \underset{r}{Id} \preceq \mathbb{E}_{g\sim\mathcal{N}(0,Id_r)}\left[\nabla p(g) \nabla p(g)^\top\right] \preceq 2 \cdot \underset{r}{Id}.$$

Proof. For convenience, let M and M_* denote $\mathbb{E}\left[\nabla p(g) \nabla p(g)^\top\right]$ and $\mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{Id}_r)}\left[\nabla p_*(g) \nabla p_*(g)^\top\right]$ respectively. For any $v \in \mathbb{S}^{r-1}$, we have that

$$\begin{aligned}
|v^\top M_* v - v^\top M v| &= \left|\mathbb{E}\left[\langle v, \nabla p_*(g)\rangle^2 - \langle v, \nabla p(g)\rangle^2\right]\right| \\
&= |\mathbb{E}\left[\langle v, \nabla \delta(g)\rangle \cdot \langle v, \nabla (p + p_*)(g)\rangle\right]| \\
&\le \mathbb{E}\left[\|\nabla \delta(g)\|_2^2\right]^{1/2} \cdot \left(\mathbb{E}\left[\|\nabla p(g)\|_2^2\right]^{1/2} + \mathbb{E}\left[\|\nabla p_*(g)\|_2^2\right]^{1/2}\right) \\
&\le rd \cdot \mathbb{V}[\delta]^{1/2} \cdot (\mathbb{V}[p]^{1/2} + \mathbb{V}[p_*]^{1/2}) \\
&< O(r^{3/2} d \cdot \|\mathbf{c} - \mathbf{c}^*\|_2),
\end{aligned}$$

where in the third step we used Cauchy-Schwarz, in the fourth step we used Lemma 2.3.8, and in the last step we upper bounded $\mathbb{V}[p]$ and $\mathbb{V}[p_*]$ by $O(r)$ using Corollary 2.3.2 and the fact that $\|\mathbf{c} - \mathbf{c}^*\|_2 = O(1)$. □

To conclude the proof of Lemma 2.7.13, we see that

$$\mathbb{E}[\text{\textcircled{A'}}] \in [\nu_{\mathsf{cond}}/2, 2] \cdot \mathrm{Tr}(\mathrm{Id} - V^{*\top} V V^\top V^*)$$

$$
\begin{aligned}
&= [\nu_{\mathsf{cond}}/2, 2] \cdot d_C(V, V^*)^2 \\
&\in [\nu_{\mathsf{cond}}/4, 4] \cdot d_P(V, V^*)^2, \qquad (2.31)
\end{aligned}
$$

where the first step follows by (2.30) and Lemma 2.7.14, the second step follows by the fact that $\mathrm{Tr}(\mathrm{Id} - V^{*\top} V V^\top V^*) = d - \|V^{*\top} V\|_F^2$, and the last step follows by Lemma 1.3.5. □

Lemma 2.7.9 now follows from (2.29), Observation 2.7.12, and Lemma 2.7.13. □

We defer the proofs of Lemmas 2.7.10 and 2.7.11, to Appendix 2.11.

Local Curvature in Expectation- All Iterations

In this section we extend the results of the previous section to give bounds on the sum *over all* t of the expected correlations between the direction in which we would like to move at time t, and the step we actually take at time t.

Specifically, for the sequence of iterates $\{\Theta^{(t)}\}_{0\le t\le T}$ in SUBSPACEDESCENT, we would like to bound $\mathbb{E}\left[\sum_{t=0}^{T-1} \mu_X(\Theta^{(t)})\right]$, $\left|\mathbb{E}\left[\sum_{t=0}^{T-1} \mu_{E_1}(\Theta^{(t)})\right]\right|$, and $\left|\mathbb{E}\left[\sum_{t=0}^{T-1} \mu_{E_2}(\Theta^{(t)})\right]\right|$. We emphasize that the expectation here is over the randomness of the samples $x^0, ..., x^{T-1}$, so e.g. $\mu_X(\Theta^{(t)})$ is a random variable depending on $x^0, ..., x^{t-1}$ and is itself an expectation over the next sample x^t.

Intuitively, for our choice (2.21) of small step size η_{vec} which scales with $O(1/T)$, Lemma 2.7.7 suggests that the expected behavior of the corresponding martingales should not be very different from that of a sum of iid random variables. That is, these expected sums should be not much different than T times the expectation of their *first* summand, corresponding to the first iteration which takes a step from $\Theta^{(0)}$. In Lemmas 2.7.15, 2.7.16, and 2.7.17, we show that this is indeed the case:

Lemma 2.7.15. $\mathbb{E}\left[\sum_{t=0}^{T-1} \mu_X(\Theta^{(t)})\right] \le T \cdot \eta_{\mathsf{vec}} \cdot (\nu_{\mathsf{cond}}/2.2) \cdot d_P(V^{(0)}, V^*)^2$.

Lemma 2.7.16. $\mathbb{E}\left[\sum_{t=0}^{T-1} \mu_{E_1}(\Theta^{(t)})\right] \le T \cdot O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot d_P(V^{(0)}, V^*)^3$.

Lemma 2.7.17. $\mathbb{E}\left[\sum_{t=0}^{T-1} \mu_{E_2}(\Theta^{(t)})\right] \le T \cdot O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{d+2} \cdot d_P(V^{(0)}, V^*)^3$.

We defer their proofs to Appendices 2.11.5, 2.11.6, and 2.11.7 respectively.

Local Curvature with High Probability

In this section, we complete the proof of Lemma 2.7.3 by establishing high-probability bounds for the MDS's corresponding to X, E_1, and E_2. That is, we argue that with high probability, the dominant term given by X is large, while the error terms from Taylor approximation and from the trigonometric corrections are small. Specifically, we show:

Lemma 2.7.18.

$$\sum_{t=0}^{T-1} X^{\Theta^{(t)},x^t} \geq T \cdot \eta_{\mathsf{vec}} \cdot (\nu_{\mathsf{cond}}/3) \cdot d_P(V^{(0)}, V^*)^2$$

with probabiliy at least $1 - \delta$*.*

Lemma 2.7.19.

$$\left| \sum_{t=0}^{T-1} E_1^{\Theta^{(t)},x^t} \right| \leq T \cdot \eta_{\mathsf{vec}} \cdot (c_{12} \cdot \nu_{\mathsf{cond}}) \cdot d_P(V^{(0)}, V^*)^2$$

with probability at least $1 - \delta$*.*

Lemma 2.7.20.

$$\left| \sum_{t=0}^{T-1} E_2^{\Theta^{(t)},x^t} \right| \leq T \cdot \eta_{\mathsf{vec}} \cdot (c_{12} \cdot \nu_{\mathsf{cond}}) \cdot d_P(V^{(0)}, V^*)^2$$

with probability at least $1 - \delta$*.*

We defer their proofs to Appendix 2.11.8. The key technical step in all three proofs is to upper bound the variance of the martingale differences, after which one can invoke the corresponding expectation bounds from Section 2.7.2 together with the martingale concentration inequalities of Lemma 2.3.4 for Lemma 2.11.8 and Lemma 2.3.3 for Lemmas 2.7.19 and 2.7.20. We emphasize that here we must again crucially use (2.20), this time to ensure that the variances of the martingale differences, which depend in part on $\|\mathbf{c} - \mathbf{c}^*\|_2$, do not swamp the expectation $\mu_X(\Theta)$ of the dominant term.

Also, we remark that it is in the proof of Lemma 2.7.19 and Lemma 2.7.20 that we finally use the assumption (2.19) that $d_P(V^{(0)}, V^*)$ is somewhat small.

Finally, we can deduce Lemma 2.7.3, completing the proof of Theorem 2.7.1.

Proof of Lemma 2.7.3. By Lemmas 2.7.18, 2.7.19, and 2.7.20, and the earlier calculation showing that for any $\Theta = (\mathbf{c}, V)$, $\langle \Delta^{\Theta,x}_{\mathsf{vec}}, V - V^* \rangle = X^{\Theta,x} + E_1^{\Theta,x} + E_2^{\Theta,x}$, we see that under our choice of T, η_{vec},

$$\sum_{t=0}^{T-1} \left\langle \Delta^{\Theta^{(t)},x^t}_{\mathsf{vec}}, V^{(t)} - V^* \right\rangle \geq \nu_{\mathsf{cond}} \left(\frac{1}{3} - 2c_{12} \right) \cdot T \cdot \eta_{\mathsf{vec}} \cdot d_P(V^{(0)}, V^*)^2$$

with probability $1 - 3\delta$. By replacing 3δ with δ, and absorbing the constant factors, the lemma follows. □

2.8 Putting Everything Together for GeoSGD

In this section we conclude the proof of Theorem 2.5.1 using Theorems 2.6.1 and 2.7.1.

There is one last subtlety we must address. In Theorem 2.6.1 on the distance $\|\mathbf{c} - \mathbf{c}^*\|$ between the coefficients $\mathbf{c}$ output by RealignPolynomial and the true coefficients $\mathbf{c}^*$, the upper bound is at best only in terms of the known parameter $\underline{\varepsilon}$. On the other hand, in Theorem 2.7.1 on the error $d_P(V^{(T)}, V^*)$ incurred by the subspace $V^{(T)}$ output by SubspaceDescent when initialized to $V^{(0)}$, the upper bound we can show only applies when (2.20) holds.

The scenario that these guarantees do not account for is when at some point in the middle of GeoSGD, we arrive upon a subspace $V^{(0)}$ for which $d_P(V^{(0)}, V^*) \ll \underline{\varepsilon}/2$, in which case running RealignPolynomial with $V^{(0)}$ gives coefficients $\mathbf{c}$ for which (2.20) fails to hold. Intuitively, this should be fine because $d_P(V^{(0)}, V^*) < \underline{\varepsilon}$, so GeoSGD has already produced a good enough estimate for the true subspace and we could just terminate. Unfortunately, it is not immediately obvious how to tell when this has happened and terminate accordingly.

Instead, we argue that local smoothness for SubspaceDescent (Lemma 2.7.2), implies that in this case, running SubspaceDescent initialized to $V^{(0)}$ will produce a subspace $V^{(T)}$ whose error is still good enough:

Lemma 2.8.1. *Suppose all of the assumptions of Theorem 2.7.1 hold except for (2.20). Then we still have that $d_P(V^{(T)}, V^*) \leq \|\mathbf{c} - \mathbf{c}^*\|_2$ with probability at least $1 - \delta$.*

Proof. Suppose the event of Lemma 2.7.2 occurs. We have that

$$\begin{aligned}
d_P(V^{(T)}, V^*) &\le d_P(V^{(0)}, V^*) + d_P(V^{(0)}, V^{(T)}) \\
&\le d_P(V^{(0)}, V^*) \cdot \left(1 + \eta_{\mathsf{vec}} \cdot O(dr^3 \ln(T/\delta))^{(d+2)/2} \cdot O(\sqrt{n})\right) \\
&\le \frac{1}{2}\|\mathbf{c} - \mathbf{c}^*\|_2 \cdot \left(1 + \eta_{\mathsf{vec}} \cdot O(dr^3 \ln(T/\delta))^{(d+2)/2} \cdot O(\sqrt{n})\right) \\
&= \frac{1}{2}\|\mathbf{c} - \mathbf{c}^*\| \cdot \left(1 + O\left(\frac{\nu_{\mathsf{cond}}}{T\sqrt{n}}\right)\right) \\
&< \|\mathbf{c} - \mathbf{c}^*\|_2,
\end{aligned}$$

where the first step follows by triangle inequality for Procrustes distance (Fact 1.3.4), the second by the assumption that the event of Lemma 2.7.2 holds, the third by the assumption that (2.20) does not hold, and the fourth by the definition of η_{vec} in (2.21). □

We can now complete the proof of Theorem 2.5.1.

Proof of Theorem 2.5.1. Let $\mathbf{c}^{(t)}$ and $V^{(t)}$ be the iterates of GeoSGD. Suppose for $0 \le t < T$ we had $d_P(V^{(t)}, V^*) \le c_{12} \cdot \nu_{\mathsf{cond}} \cdot O(dr^3)^{-d-2}$. By Theorem 2.6.1, we have that

$$\|\mathbf{c}^{(t+1)} - \mathbf{c}^*\|_2 < 2 \cdot \varepsilon/2 \vee d_P(V^{(t)}, V^*).$$

If $\|\mathbf{c}^{(t+1)} - \mathbf{c}^*\|_2 < \varepsilon$, then by Lemma 2.8.1, $d_P(V^{(t+1)}, V^*) < \varepsilon$. Otherwise, if $\|\mathbf{c}^{(t+1)} - \mathbf{c}^*\|_2 \le 2d_P(V^{(t)}, V^*)$, then (2.20) in Theorem 2.7.1 holds and we get that

$$d_P(V^{(t+1)}, V^*) \le (1 - \alpha) \cdot d_P(V^{(t)}, V^*),$$

where

$$\alpha \triangleq \frac{\nu_{\mathsf{cond}}}{n} \cdot \mathrm{poly}(\ln(1/\nu_{\mathsf{cond}}), r, d, \ln(1/\delta'))^{-d}$$

for $\delta' = \delta/(2T+1)$ as defined in GeoSGD.

In either case, $d_P(V^{(t+1)}, V^*) \le c_{12} \cdot \nu_{\mathsf{cond}} \cdot O(dr^3)^{-d-2}$. And furthermore, if we unroll this recurrence, we conclude that

$$d_P(V^{(T)}, V^*) \le \varepsilon \vee (1 - \alpha)^T \cdot d_P(V^{(0)}, V^*).$$

So by taking $T = \alpha^{-1} \cdot \log(1/\varepsilon)$, we get that $d_P(V^{(T)}, V^*) \leq \varepsilon$ as desired. This corresponds to the choice of T in (2.5). Lastly, we get that $\|\mathbf{c}^{(T)} - \mathbf{c}\|_2 \leq \varepsilon$ by one last application of Theorem 2.6.1. □

Proof of Theorem 2.5.2. This follows from the runtime and sample complexity guarantees of Theorems 2.6.1 and 2.7.1. □

2.9 Appendix: Martingale Concentration Inequalities

In this section we prove the two martingale concentration inequalities from Section 2.3.2 that are needed for the analysis of the boosting phase of our algorithm.

2.9.1 Proof of Lemma 2.3.3

We first prove the following more general statement.

Lemma 2.9.1. *Let $\sigma > 0$ and $0 < \alpha \leq 2$ be constants, and let $\mathcal{E}_i$ be the event that $\mathbb{E}[|Z_i|^q|\xi_1, ..., \xi_{i-1}] \leq \sigma^q \cdot q^{q/\alpha}$ for all $q \geq 1$.*

If $Pr[\mathcal{E}_i|\xi_1, ..., \xi_{i-1}] \geq 1 - \beta$ for each $i \in [T]$, then for any $t > 0$,

$$Pr\left[\max_{\ell\in[T]}\left|\sum_{i=1}^{\ell} Z_i\right| \geq t\cdot\sqrt{T}\cdot\sigma\right] \leq O\left(1 + t^2(1/\alpha)^{O(1/\alpha)}\right)\cdot\exp\left(-\left(t^2/32\right)^{\frac{\alpha}{2+\alpha}}\right) + T\cdot\beta. \quad (2.32)$$

In particular, there is an absolute constant $c_1 > 0$ such that for any $\delta > 0$,

$$Pr\left[\max_{\ell\in[T]}\left|\sum_{i=1}^{\ell} Z_i\right| \geq (\log(1/\delta)/\alpha)^{2c_1/\alpha}\cdot\sqrt{T}\cdot\sigma\right] \leq \delta + T\cdot\beta.$$

We first show that this implies Lemma 2.3.3.

Proof of Lemma 2.3.3. This is an immediate consequence of Lemma 2.9.1 together with Fact 1.3.15, which implies the requisite moment bounds for Lemma 2.9.1 for $\alpha = d/2$. □

To show Lemma 2.9.1, we require the following theorem on the concentration of martingales with sub-Weibull differences, which is a consequence of the main result of [Li18a].

Theorem 2.9.2 ([Li18a]). *Let $\sigma > 0$ and $0 < \alpha \leq 2$ be constants. Suppose that for every $i \in [T]$, we have that with probability one, $\mathbb{E}[|Z_i|^q|\xi_1, ..., \xi_{i-1}] \leq \sigma^q \cdot q^{q/\alpha}$ holds for all $q \geq 1$. Then for any $z > 0$,*

$$Pr\left[\max_{\ell \in [T]} \left|\sum_{i=1}^{\ell} Z_i\right| \geq t \cdot \sqrt{T} \cdot \sigma\right] \leq O\left(1 + t^2(1/\alpha)^{O(1/\alpha)}\right) \cdot \exp\left(-\left(t^2/32\right)^{\frac{\alpha}{2+\alpha}}\right) \tag{2.33}$$

We use a standard trick, see e.g. Lemma 3.1 of [Vu02], to relax the assumption that the differences are sub-Weibull almost surely to the assumption that they are sub-Weibull with high probability. It will also be more convenient for us to state the inequality in terms of moment bounds rather than Orlicz norm bounds.

Proof of Lemma 2.9.1. Given a realization ξ of the random variables $(\xi_1, ..., \xi_T)$, let i_ξ be the first index i, if any, for which $\mathcal{E}_i$ does not hold. Define $B_i \triangleq \{\xi : i_\xi = i\}$ and note that these sets are disjoint for different i. Let $Y'(\xi)$ be the function which agrees with $Y(\xi)$ for $\xi \in (\cup B_i)^c$ and which is equal to $\mathbb{E}_{B_i}[Y]$ for $\xi \in B_i$. Y' and Y have the same mean, so the lemma follows by union bounding over the events $\cup B_i$ together with the probability that the martingale Y' fails to concentrate. For the former probabilities, by definition $\Pr[B_i] \leq \beta$. And for the latter, because the martingale differences for Y' satisfy the assumptions of Theorem 2.9.2, Y' fails to concentrate with probability at most the right-hand side of (2.33). This yields (2.32). □

2.9.2 Proof of Lemma 2.3.4

To show Lemma 2.3.4, we require the following theorem due to [Ben03], which controls the tails of martingales whose differences are only bounded on one side.

Theorem 2.9.3 ([Ben03]). *Let $\{c_i\}_{i \in [T]}$ and $\{s_i\}_{i \in [T]}$ be collections of positive constants for which $Z_i \leq c_i$ and $\mathbb{E}[Z_i^2|\xi_1, ..., \xi_{i-1}] \leq s_i^2$ with probability one for every $i \in [T]$. Let $\sigma_i = c_i \vee s_i$, and define $\sigma^2 = \sum_i \sigma_i^2$. Then*

$$Pr\left[\sum_{i=1}^{T} Z_i \geq t \cdot \sigma\right] \leq \exp(-t^2/2).$$

Proof of Lemma 2.3.4. The proof is identical to that of Lemma 2.9.1, except instead of applying Theorem 2.9.2 to the auxiliary martingale, we apply Theorem 2.9.3 to get that for any $t > 0$,

$$\Pr\left[\sum_{i=1}^{T} Z_i \geq t \cdot \sigma\right] \leq \exp(-t^2/2) + T \cdot \beta.$$

The lemma follows by taking $t = \sqrt{2}\log(1/\delta)$. □

2.10 Appendix: Deferred Proofs from Section 2.6

2.10.1 Proof of Lemma 2.6.5

Proof.

$$\begin{aligned}
&\mathbb{E}\left[(F_x(\Theta) - F_x(\Theta^*))^4\right] \\
&\quad\leq \sum_{\ell_1,\ldots,\ell_4 \in [d+1]} \frac{1}{\prod_{\nu=1}^4 \ell_\nu!} \mathbb{E}\left[\prod_{\nu=1}^{4} \left\langle \nabla^{[\ell_\nu]} F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell_\nu} \right\rangle\right] \\
&\quad\leq \sum_{\ell_1,\ldots,\ell_4 \in [d+1]} \frac{1}{\prod_{\nu=1}^4 \ell_\nu!} \cdot 16 \cdot (8dr^2)^{2(d+1)} \cdot \|V - V^*\|_F^{\sum_\nu \ell_\nu} \cdot \left(1 + \frac{\|\mathbf{c} - \mathbf{c}^*\|_2}{\|V^* - V\|_F}\right)^4 \\
&\quad\leq 16 \cdot (8dr^2)^{2(d+1)} \left(\sum_{\ell=1}^{d+1} \frac{1}{\ell!} \cdot \|V - V^*\|_F^{\ell}\right)^4 \cdot \left(1 + \frac{\|\mathbf{c} - \mathbf{c}^*\|_2}{\|V^* - V\|_F}\right)^4 \\
&\quad\leq 16 \cdot (8dr^2)^{2(d+1)} \cdot (e \cdot (4r)^{d/2} \|V - V^*\|_F)^4 \cdot \left(1 + \frac{\|\mathbf{c} - \mathbf{c}^*\|_2}{\|V^* - V\|_F}\right)^4 \\
&\quad\leq (2e)^4 \cdot (32dr^3)^{2(d+1)} \cdot \left(\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2\right)^4,
\end{aligned}$$

where the second step follows by Lemma 2.5.4, the fourth by the fact that $\|V - V^*\|_F \leq 2\sqrt{r}$ and the fact that $\sum_{\ell=1}^{d+1} \frac{1}{\ell!} \cdot x^\ell \leq e \cdot (4r)^{d/2} \cdot x$ for $x \in [0, 2\sqrt{r}]$. □

2.10.2 Proof of Lemma 2.6.10

Proof. We have that

$$\frac{1}{2\eta_{\mathsf{coef}}} \left|\langle \Delta''_{\mathsf{coef}}, \mathbf{c} - \mathbf{c}^* \rangle\right| = \left|\mathbb{E}\left[\sum_{\ell=2}^{d+1} \frac{1}{\ell!} \left\langle \nabla^{[\ell]} F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell} \right\rangle \cdot \delta(V^\top x)\right]\right|$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\left(\sum_{\ell=2}^{d+1} \frac{1}{\ell!}\langle \nabla^{[\ell]} F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell}\rangle\right)^2\right]^{1/2} \cdot \mathbb{E}\left[\delta(V^\top x)^2\right]^{1/2} \\
&\leq O(dr^3)^{(d+1)/2} \cdot \|V - V^*\|_F \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \\
&= O(dr^3)^{(d+1)/2} \cdot d_P(V, V^*) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot (d_P(V, V^*) + \|\mathbf{c} - \mathbf{c}^*\|_2),
\end{aligned}
$$

where the second step follows by Cauchy-Schwarz, the third step follows by Lemma 2.11.1, and the last step follows by the assumption that $\|V - V^*\|_F = d_P(V, V^*)$. □

2.10.3 Proof of Proposition 2.6.13

Proof. Note that

$$
\begin{aligned}
\mathbb{E}_g\left[\left(x^\top \Pi_V (V^* - V)\nabla p(V^\top x)\right)^2\right]^{1/2} &\leq \|\mathrm{Id} - V^\top V^*\|_2 \cdot \mathbb{E}_g\left[\|g\|_2^2 \cdot \|\nabla p(g)\|_2^2\right]^{1/2} \\
&\leq d_P(V, V^*)^2 \cdot O(r^{3/2} d),
\end{aligned}
$$

where the second step follows by the second part of Lemma 2.3.13, Lemma 2.3.10, and the fact that

$$
\mathbb{V}[p]^{1/2} \leq \|\mathbf{c} - \mathbf{c}^*\|_2 + \mathbb{V}[p^*]^{1/2} \leq O(r)
$$

because $\|\mathbf{c} - \mathbf{c}^*\|_2 \leq 1$ by assumption and because of Corollary 2.3.2. □

2.10.4 Proof of Lemma 2.6.15

We will split up $\frac{1}{B}\sum_{i=0}^{B-1} Y^{x^i}$ according to the decomposition (2.18). That is, define

$$
\begin{aligned}
\textcircled{A}^x &\triangleq \left(\delta(V^\top x)\right)^2 \\
\textcircled{B}^x &\triangleq \delta(V^\top x) \cdot \left(x^\top \Pi_V (V^* - V)\nabla\right) \\
\textcircled{C}^x &\triangleq \delta(V^\top x) \cdot \left(x^\top \Pi_V^\perp V^* \nabla\right)
\end{aligned}
$$

so that for any x,

$$
\frac{1}{2\eta_{\mathsf{coef}}} Y^x = \textcircled{A}^x + \textcircled{B}^x + \textcircled{C}^x. \tag{2.34}
$$

We will show concentration for these three random variables separately.

Lemma 2.10.1. *For any $\delta > 0$, if $B = \Omega(\log(1/\delta)^2 \cdot 9^d)$, then*

$$\frac{1}{B}\sum_{i=0}^{B-1} \textcircled{A}^{x^i} \geq \frac{1}{2}\|\mathbf{c} - \mathbf{c}^*\|_2^2$$

with probability at least $1 - \delta$.

Lemma 2.10.2. *For any $\delta > 0$, if $B = \Omega(\log(1/\delta))^{2d}$, then*

$$\left|\frac{1}{B}\sum_{i=0}^{B-1} \textcircled{B}^{x^i}\right| \leq O(r^{3/2}d) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot d_P(V, V^*)^2$$

with probability at least $1 - \delta$.

Lemma 2.10.3. *For any $\delta > 0$ and $\gamma > 0$, if $B = \Omega(\log(1/\delta))^{2d} \cdot \gamma^{-2}$, then*

$$\left|\frac{1}{B}\sum_{i=0}^{B-1} \textcircled{C}^{x^i}\right| \leq \gamma \cdot d_P(V, V^*) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2$$

with probability at least $1 - \delta$.

We prove these in the subsequent Appendices 2.10.4, 2.10.4, and 2.10.4. Note that Lemma 2.6.15 immediately follows from these lemmas.

Proof of Lemma 2.6.15. By a union bound over the failure probabilities of Lemmas 2.10.1, 2.10.2, and 2.10.3, we see by triangle inequality and (2.34) that

$$\frac{1}{B}\sum_{i=0}^{B-1} Y^{x^i} \geq 2\eta_{\mathsf{coef}} \cdot \left(\frac{1}{2}\|\mathbf{c} - \mathbf{c}^*\|_2^2 - O(r^{3/2}d) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot d_P(V, V^*)^2 - \gamma \cdot d_P(V, V^*) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2\right)$$

with probability at least $1 - 3\delta$, provided $B = \Omega(\log(1/\delta))^d \cdot \gamma^{-2}$. The result follows by replacing 3δ with δ and absorbing constants. □

Proof of Lemma 2.10.1

Proof. Observe that $\frac{1}{B}\sum_{i=0}^{B-1}\left(\mathbb{E}_x[\textcircled{A}^x] - \textcircled{A}^{x^i}\right)$ is an average of B iid copies of a mean-zero random variable satisfying one-sided bounds, so we wish to apply Lemma 1.3.33.

To do so, we just need to bound the variances of the summands.

Lemma 2.10.4. $\mathbb{V}_x[\textcircled{A}^x] \le 9^d \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^4$.

Proof. Clearly $\mathbb{V}[\textcircled{A}^x] \le \mathbb{E}[(\textcircled{A}^x)^2]$, so it suffices to bound the latter. By Fact 1.3.15 applied to the degree-d polynomial δ,

$$\mathbb{E}[(\textcircled{A}^x)^2] = \mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{Id}_r)}[\delta(g)^4] \le 9^d \cdot \mathbb{E}[\delta(g)^2]^2 = 9^d \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^4 \tag{2.35}$$

as claimed. □

We can now complete the proof of Lemma 2.10.1.

By Lemma 1.3.33, Observation 2.6.12, and Lemma 2.10.4,

$$\frac{1}{B}\sum_{i=0}^{B-1} \textcircled{A}^{x^i} \ge \|\mathbf{c} - \mathbf{c}^*\|_2^2 - \frac{1}{\sqrt{B}} \cdot \sqrt{2}\log(1/\delta) \cdot 3^d \cdot \|\mathbf{c} - \mathbf{c}^*\|^2$$

with probability at least $1 - \delta$. The lemma follows by taking $B = \Omega(\log(1/\delta)^2$. □

Proof of Lemma 2.10.2

Proof. Note that $\textcircled{B}^x$ is a polynomial of degree $2d$ in x, so by Lemma 1.3.16, we just need to upper bound its variance.

Lemma 2.10.5. $\mathbb{V}_x[\textcircled{B}^x] \le 9^d \cdot O(r^{3/2}d) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^2 \cdot d_P(V, V^*)^4$.

Proof. We will upper bound $\mathbb{E}_x[(\textcircled{B}^x)^2]$ via

$$\begin{aligned}\mathbb{E}\left[\textcircled{B}^2\right] &\le \mathbb{E}\left[\delta(V^\top x)^4\right]^{1/2} \cdot \mathbb{E}\left[\left(x^\top \Pi_V (V^* - V)\nabla\right)^4\right]^{1/2} \\ &\le \mathbb{E}[\textcircled{A}^2] \cdot 3^d \cdot \mathbb{E}\left[\left(x^\top \Pi_V (V^* - V)\nabla\right)^2\right] \\ &\le 9^d \cdot O(r^3 d^2) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^2 \cdot d_P(V, V^*)^4,\end{aligned}$$

where in the first step we used Cauchy-Schwarz, in the second we used Proposition 2.6.13, and in the third we used (2.35). □

We can now complete the proof of Lemma 2.10.1.

By Lemma 1.3.16, Lemma 2.6.14, and Lemma 2.10.5,

$$\left|\frac{1}{B}\sum_{i=0}^{B-1} \textcircled{B}^{x^i}\right| \le O(r^{3/2}d)\cdot\|\mathbf{c}-\mathbf{c}^*\|_2\cdot d_P(V,V^*)^2\cdot\left(1+\frac{1}{\sqrt{B}}\cdot O(\log(1/\delta))^d\cdot 3^d\right),$$

with probability at least $1-\delta$. The lemma follows by taking $B=\Omega(\log(1/\delta))^{2d}\cdot\Omega(9^d)$. □

Proof of Lemma 2.10.3

Proof. Note that $\textcircled{C}^x$ is a polynomial of degree $2d$ in x, so by Lemma 1.3.16, we just need to upper bound its variance.

Lemma 2.10.6. *For any* Θ, $\mathbb{E}_x[(\textcircled{C}^{\Theta,x})^2] \le d_P(V,V^*)^2\cdot\|\mathbf{c}-\mathbf{c}^*\|_2^2\cdot\exp(O(d))$.

Proof. This is shown in Lemma 2.11.9 below. The proof involves calculations which are more pertinent to the behavior of SubspaceDescent, so we defer the details to there. □

We can now complete the proof of Lemma 2.10.3. By Lemma 1.3.16, Observation 2.6.11, and Lemma 2.10.6,

$$\left|\frac{1}{B}\sum_{i=1}^{B-1} \textcircled{C}^{x^i}\right| \le \frac{1}{\sqrt{B}}\cdot O(\log(1/\delta))^d\cdot d_P(V,V^*)\cdot\|\mathbf{c}-\mathbf{c}^*\|_2\cdot\exp(O(d))$$

with probability at least $1-\delta$. The lemma follows by taking $B=\Omega(\log(1/\delta))^{2d}\cdot\gamma^{-2}$. □

2.10.5 Proof of Lemma 2.6.16

Proof. Note that E^x is a polynomial of degree $2d$ in x, so by Lemma 1.3.16, we just need to upper bound its variance.

To do so, we will need the following helper lemma, which like Lemma 2.6.5 is a straightforward consequence of Lemma 2.5.4.

Lemma 2.10.7. $\mathbb{E}[(\mathfrak{R}^{\Theta,x})^4]^{1/2} \le O(dr^3)^{d+1}\cdot\|V-V^*\|_F^2\cdot\left(\|\mathbf{c}-\mathbf{c}^*\|_2+\|V^*-V\|_F\right)^2$

Proof. We have that

$$\begin{aligned}
\mathbb{E}\left[(\mathfrak{R}^{\Theta,x})^4\right]^{1/2} &= \left(\sum_{\ell_1,\dots,\ell_4>1}\frac{1}{\prod_{\nu=1}^4 \ell_\nu!}\mathbb{E}\left[\prod_{\nu=1}^4 \left\langle \nabla^{[\ell_\nu]}F_x(\Theta),(\Theta^*-\Theta)^{\otimes \ell_\nu}\right\rangle\right]\right)^{1/2} \\
&\le \left(\sum_{\ell_1,\dots,\ell_4>1}\frac{1}{\prod_{\nu=1}^4 \ell_\nu!}16\cdot(8dr^2)^{2(d+1)}\cdot\|V^*-V\|_F^{\sum_\nu \ell_\nu}\cdot\left(1+\frac{\|\mathbf{c}-\mathbf{c}^*\|_2}{\|V^*-V\|_F}\right)^4\right)^{1/2} \\
&= 4(8dr^2)^{d+1}\left(\sum_{\ell=2}^{d+1}\frac{1}{\ell!}\|V^*-V\|_F^\ell\right)^2\cdot\left(1+\frac{\|\mathbf{c}-\mathbf{c}^*\|_2}{\|V^*-V\|_F}\right)^2 \\
&\le 4(8dr^2)^{d+1}\cdot\left(e^2\cdot(4r)^{d-1}\|V^*-V\|_F^4\right)\cdot\left(1+\frac{\|\mathbf{c}-\mathbf{c}^*\|_2}{\|V^*-V\|_F}\right)^2 \\
&= 4e^2\cdot(32dr^3)^{d+1}\cdot\|V-V^*\|_F^2\cdot\left(\|\mathbf{c}-\mathbf{c}^*\|_2+\|V^*-V\|_F\right)^2,
\end{aligned}$$

where the second step follows by Lemma 2.5.4, and the fourth step follows by the fact that we always have $\|V-V^*\|_F\le 2\sqrt{r}$, and $\sum_{\ell=2}^{d+1}\frac{1}{\ell!}x^\ell < e\cdot(4r)^{(d-1)/2}\cdot x^2$ for $x\in[0,2\sqrt{r}]$. □

We can now show the variance bound.

Lemma 2.10.8. $\mathbb{E}_x[(E^x)^2]\le \eta_{\mathsf{coef}}^2\cdot O(dr^3)^{d+1}\cdot d_P(V,V^*)^2\cdot\|\mathbf{c}-\mathbf{c}^*\|_2^2\cdot(d_P(V,V^*)+\|\mathbf{c}-\mathbf{c}^*\|)^2$.

Proof. By Cauchy-Schwarz,

$$\begin{aligned}
\frac{1}{4\eta_{\mathsf{coef}}^2}\mathbb{E}\left[(E^x)^2\right] &\le \mathbb{E}[(\mathfrak{R}^{\Theta,x})^4]^{1/2}\cdot\mathbb{E}[\delta(g)^4]^{1/2} \\
&\le 4e^2\cdot(32dr^3)^{d+1}\cdot\|V-V^*\|_F^2\cdot\left(\|\mathbf{c}-\mathbf{c}^*\|_2+\|V^*-V\|_F\right)^2\cdot 3^d\cdot\|\mathbf{c}-\mathbf{c}^*\|_2^2 \\
&= O(dr^3)^{d+1}\cdot d_P(V,V^*)^2\cdot\|\mathbf{c}-\mathbf{c}^*\|_2^2\cdot(d_P(V,V^*)+\|\mathbf{c}-\mathbf{c}^*\|)^2
\end{aligned}$$

where the second step follows by Lemma 2.10.7 and the third step follows by the assumption that $\|V-V^*\|_F = d_P(V,V^*)$. □

Finally, by Lemma 1.3.16, Lemma 2.6.10, and Lemma 2.10.8,

$$\left|\frac{1}{B}\sum_{i=0}^{B-1}E^{x^i}\right|\le O(dr^3)^{(d+1)/2}\,d_P(V,V^*)\,\|\mathbf{c}-\mathbf{c}^*\|_2\,(d_P(V,V^*)+\|\mathbf{c}-\mathbf{c}^*\|_2)\cdot\left(1+\frac{1}{\sqrt{B}}\cdot O(\log(1/\delta))^d\right).$$

The lemma follows by taking $B = O(\log(1/\delta))^{2d}$. □

2.11 Appendix: Deferred Proofs from Section 2.7

2.11.1 Proof of Lemma 2.5.4

Proof. We begin by explicitly computing the higher-order terms in the Taylor-expansion of $F_x(\Theta) - F_x(\Theta^*)$. For any $\ell \in [d+1]$, recalling the notation of (2.3) and (2.4),

$$
\begin{aligned}
&\left\langle \nabla^{[\ell]} F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell} \right\rangle \\
&\quad = \sum_{\mathbf{i}\in[n]^\ell, \mathbf{j}\in[r]^\ell} \prod_{a=1}^{\ell} (V^*_{i_a,j_a} - V_{i_a,j_a}) \cdot \mathrm{D}_{\mathbf{i},\mathbf{j}}\, F_x(\Theta) + \sum_{I, \mathbf{i}\in[n]^\ell, \mathbf{j}\in[r]^{\ell-1}} \prod_{a=1}^{\ell-1} (V^*_{i_a,j_a} - V_{i_a,j_a}) \cdot (c^*_I - c_I) \cdot \mathrm{D}_{\mathbf{i},\mathbf{j}}\, F_x(\Theta) \\
&\quad = \sum_{\mathbf{i}\in[n]^\ell, \mathbf{j}\in[r]^\ell} \prod_{a=1}^{\ell} (V^*_{i_a,j_a} - V_{i_a,j_a}) \cdot x_{i_a} \cdot \mathrm{D}_{\mathbf{j}}\, p(V^\top x) + \sum_{\mathbf{i}\in[n]^\ell, \mathbf{j}\in[r]^{\ell-1}} \prod_{a=1}^{\ell-1} (V^*_{i_a,j_a} - V_{i_a,j_a}) \cdot x_{i_a} \cdot \mathrm{D}_{\mathbf{j}}\, \delta(V^\top x) \\
&\quad = \sum_{\mathbf{j}\in[r]^\ell} \prod_{a=1}^{\ell} \langle (V^* - V)_{j_a}, x\rangle \cdot \mathrm{D}_{\mathbf{j}}\, p(V^\top x) + \sum_{\mathbf{j}\in[r]^{\ell-1}} \prod_{a=1}^{\ell-1} \langle (V^* - V)_{j_a}, x\rangle \cdot \mathrm{D}_{\mathbf{j}}\, \delta(V^\top x)
\end{aligned}
\tag{2.36}
$$

From (2.36), we can rewrite the quantity in the expectation as

$$
\sum_{\substack{\mathbf{b}\in\{0,1\}^m \\ \{\mathbf{j}^{(\nu)}\}_{\nu\in[m]}}} \prod_{\nu=1}^{m} \left(\prod_{a=1}^{\ell_\nu - b_\nu} \left\langle (V^* - V)_{j^{(\nu)}_a}, x \right\rangle \right) \left(\mathbb{1}[b_\nu = 0] \cdot \mathrm{D}_{\mathbf{j}^{(\nu)}}\, p(V^\top x) + \mathbb{1}[b_\nu = 1] \cdot \mathrm{D}_{\mathbf{j}^{(\nu)}}\, \delta(V^\top x) \right).
$$

We will bound the expected absolute values of each of these summands individually, so henceforth fix an arbitrary $\mathbf{b}$, $\{\mathbf{j}^{(\nu)}\}$. For convenience, define $C_\nu \triangleq \big(\mathbb{1}[b_\nu = 0] \cdot \mathrm{D}_{\mathbf{j}^{(\nu)}}\, p(V^\top x) + \mathbb{1}[b_\nu = 1] \cdot \mathrm{D}_{\mathbf{j}^{(\nu)}}\, \delta(V$

By AM-GM, we have that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\prod_{\nu=1}^{m} |C_\nu| \right) \cdot \left(\prod_{\nu=1}^{m} \prod_{a=1}^{\ell_\nu - b_\nu} \left| \left\langle (V^* - V)_{j^{(\nu)}_a}, x \right\rangle \right| \right) \right] \\
&\qquad \le \mathbb{E}\left[\left(\prod_{\nu=1}^{m} C_\nu \right) \cdot \left(\prod_{\nu=1}^{m} \frac{1}{\ell_\nu - b_\nu} \sum_{a=1}^{\ell_\nu - b_\nu} \left| \left\langle (V^* - V)_{j^{(\nu)}_a}, x \right\rangle \right|^{\ell_\nu - b_\nu} \right) \right]
\end{aligned}
$$

$$\leq \mathbb{E}\left[\prod_{\nu=1}^{m} \frac{C_\nu^2}{(\ell_\nu - b_\nu)^2}\right]^{1/2} \cdot \mathbb{E}\left[\left(\sum_{\mathbf{a}\in\prod_\nu [\ell_\nu - b_\nu]} \prod_{\nu=1}^{m} \left|\left\langle (V^* - V)_{j^{(\nu)}_{a_\nu}}, x\right\rangle\right|^{\ell_\nu - b_\nu}\right)^2\right]^{1/2} \tag{2.37}$$

where the last inequality follows by Cauchy-Schwarz.

Defining $w_{\mathbf{b}} = \sum_\nu \ell_\nu - b_\nu$, we may write the second factor in (2.37) as

$$\mathbb{E}\left[\sum_{\mathbf{a}^1,\mathbf{a}^2} \prod_{\nu=1}^{m} \left|\left\langle (V^* - V)_{j^{(\nu)}_{a^1_\nu}}, x\right\rangle\right|^{\ell_\nu - b_\nu} \cdot \prod_{\nu=1}^{m} \left|\left\langle (V^* - V)_{j^{(\nu)}_{a^2_\nu}}, x\right\rangle\right|^{\ell_\nu - b_\nu}\right]^{1/2}$$
$$\leq (2w_{\mathbf{b}})^{w_{\mathbf{b}}/2} \|V^* - V\|_F^{w_{\mathbf{b}}} \cdot \prod_\nu (\ell_\nu - b_\nu) \leq (2m)^{m/2} \|V^* - V\|_F^{w_{\mathbf{b}}} \cdot \prod_\nu (\ell_\nu - b_\nu),$$

where we used the standard bound for moments of a univariate Gaussian, the fact that there are $\prod_\nu (\ell_\nu - b_\nu)^2$ pairs of summands $\mathbf{a}^1, \mathbf{a}^2$, and the fact that any column of $V^* - V$ has L_2 norm at most $\|V^* - V\|_F$.

By Holder's, we may upper bound the first factor in (2.37) by $\prod_{\nu=1}^{m} \frac{1}{\ell_\nu - b_\nu} \mathbb{E}\left[C_\nu^{2m}\right]^{1/2m}$.

By Corollary 2.3.7,

$$\mathbb{E}\left[\left(\mathrm{D}_{\mathbf{j}^{(\nu)}}\, \mathfrak{d}(V^\top x)\right)^{2m}\right]^{1/2m} \leq (2m)^{d/2} d^{(\ell_\nu - 1)/2} \cdot \mathbb{V}[\mathfrak{d}]^{1/2} \leq (2m)^{d/2} d^{\ell_\nu/2} \cdot \|\mathbf{c} - \mathbf{c}_*\|_2.$$

$$\mathbb{E}\left[\left(\mathrm{D}_{\mathbf{j}^{(\nu)}}\, p(V^\top x)\right)^{2m}\right]^{1/2m} \leq (2m)^{d/2} d^{\ell_\nu/2} \cdot \mathbb{V}[p]^{1/2} \leq 2 \cdot (2m)^{d/2} d^{\ell_\nu/2},$$

where in the last step we used that $\mathbb{V}[p]^{1/2} \leq \mathbb{V}[p*]^{1/2} + \mathbb{V}[\mathfrak{d}]^{1/2} \leq 2$. So the first factor in (2.37) is at most

$$\left(\prod_{\nu=1}^{m} \frac{1}{\ell_\nu - b_\nu}\right) \cdot 2^m \cdot (2m)^{md/2} d^{\sum_\nu \ell_\nu/2} \|\mathbf{c} - \mathbf{c}_*\|_2^{\sum_\nu b_\nu},$$

so (2.37) is at most $2^m \cdot (2m)^{m(d+1)/2} d^{m(d+1)/2} \cdot \|V^* - V\|_F^{w_{\mathbf{b}}} \cdot \|\mathbf{c} - \mathbf{c}_*\|_2^{\sum_\nu b_\nu}$. The proof follows by noting that

$$\sum_{\mathbf{b}} \|V^* - V\|_F^{w_{\mathbf{b}}} \cdot \|\mathbf{c} - \mathbf{c}_*\|_2^{\sum_\nu b_\nu} = \|V^* - V\|_F^{\sum_\nu \ell_\nu} \cdot \sum_{\mathbf{b}} \left(\frac{\|\mathbf{c} - \mathbf{c}_*\|_2}{\|V^* - V\|_F}\right)^{\sum_\nu b_\nu}$$

and summing (2.37) over all choices of $\mathbf{b}$ and all $\prod_\nu r^{\ell_\nu} \leq r^{m(d+1)}$ choices of $\{\mathbf{j}^{(\nu)}\}$. □

2.11.2 Proof of Lemma 2.7.7

Proof. Let

$$\alpha_q \triangleq 3\sqrt{r} \cdot \eta_{\mathsf{vec}} \cdot O(\sqrt{n}) \cdot O(dr^3)^{(d+2)/2}.$$

$(q = 1)$. Analogous to the derivation of (2.27), we have that

$$\begin{aligned}
\mathbb{E}\left[\|V^{(t)} - V^*\|_F\right] &\leq \mathbb{E}\left[\|V^{(t-1)} - V^*\|_F\right] + \mathbb{E}\left[\|\Delta_{\mathsf{vec}}^{\Theta^{(t-1)},x^{t-1}}\|_F\right] \\
&\leq \mathbb{E}\left[\|V^{(t-1)} - V^*\|_F\right] + 3\sqrt{r} \cdot \eta_{\mathsf{vec}} \mathbb{E}\left[(\sigma^{\Theta^{(t-1)},x^{t-1}})^2\right]^{1/2} \\
&\leq (1 + \alpha_1)\mathbb{E}\left[\|V^{(t-1)} - V^*\|_F\right] + \alpha_1 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \\
&\leq (1 + \alpha_1)^t \cdot \|V^{(0)} - V^*\|_F + \left((1 + \alpha_1)^t - 1\right) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \\
&= (1 + \alpha_1)^t \cdot d_P(V^{(0)}, V^*) + \left((1 + \alpha_1)^t - 1\right) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2
\end{aligned}$$

where in the second step we used Cauchy-Schwarz and (2.24), in the third step we used Lemma 2.7.4, in the fourth step we unrolled the recurrence, and in the last step we used the assumption that $\|V^{(0)} - V^*\|_F = d_P(V^{(0)}, V^*)$. The proof follows by taking η_{vec} small enough that

$$(1 + \alpha_1)^t + \left((1 + \alpha_1)^t - 1\right) \cdot \frac{\|\mathbf{c} - \mathbf{c}^*\|_2}{d_P(V^{(0)}, V^*)} \leq 1.1.$$

η_{vec} given by (2.21) will easily satisfy this.

(Larger q) We have that

$$\begin{aligned}
\mathbb{E}\left[\|V^{(t)} - V^*\|_F^q\right]^{1/q} &\leq \mathbb{E}\left[\|V^{(t-1)} - V^*\|_F^q\right]^{1/q} + \mathbb{E}\left[\|\Delta_{\mathsf{vec}}^{\Theta^{(t-1)},x^{t-1}}\|_F^q\right]^{1/q} \\
&\leq \mathbb{E}\left[\|V^{(t-1)} - V^*\|_F^q\right]^{1/q} + \mathbb{E}\left[\|\Delta_{\mathsf{vec}}^{\Theta^{(t-1)},x^{t-1}}\|_F^{2q}\right]^{1/2q} \\
&\leq \mathbb{E}\left[\|V^{(t-1)} - V^*\|_F^q\right]^{1/q} + \alpha_q \cdot \left(\mathbb{E}\left[\|V^{(t-1)} - V^*\|_F\right] + \|\mathbf{c} - \mathbf{c}^*\|_2\right) \\
&\leq \mathbb{E}\left[\|V^{(t-1)} - V^*\|_F^2\right]^{1/q} + 1.1\alpha_q \cdot \left(d_P(V^{(0)}, V^*) + \cdot\|\mathbf{c} - \mathbf{c}^*\|_2\right) \\
&\leq d_P(V^{(0)}, V^*) + 1.1t \cdot \alpha_q \cdot \left(d_P(V^{(0)}, V^*) + \cdot\|\mathbf{c} - \mathbf{c}^*\|_2\right)
\end{aligned}$$

where the first step follows by triangle inequality, the second by monotonicity of L_p norms, the third by Lemma 2.7.4, the fourth by Lemma 2.7.7, and the fifth by unrolling the recurrence and using the assumption that assumption that $\|V^{(0)} - V^*\|_F = d_P(V^{(0)}, V^*)$.

The proof follows by taking η_{vec} small enough that $1.1T \cdot \alpha_q \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \le O(\alpha_q \cdot T)$ is a negligible constant, which is certainly the case if η_{vec} satisfies (2.21) (with hidden constant factors there depending on q). □

2.11.3 Proof of Lemma 2.7.10

We first prove the following basic consequence of Lemma 2.5.4:

Lemma 2.11.1.

$$\mathbb{E}\left[\left(\sum_{\ell=2}^{d+1} \frac{1}{\ell!}\langle \nabla^{[\ell]} F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell}\rangle\right)^2\right]^{1/2} \le O(dr^3)^{(d+1)/2} \cdot \|V - V^*\|_F \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2) \quad (2.38)$$

Proof. The left-hand side of (2.38) can be rewritten as

$$\begin{aligned}
&\left(\sum_{\ell_1,\ell_2>1} \frac{1}{\ell_1!\ell_2!}\mathbb{E}\left[\prod_{\nu=1}^{2} \langle \nabla^{[\ell_\nu]} F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell_\nu}\rangle\right]\right)^{1/2} \cdot \mathbb{E}\left[(x^\top \cdot \Pi_V^\perp V^* \cdot \Delta)^2\right]^{1/2} \\
&\le \left(\sum_{\ell_1,\ell_2>1} \frac{1}{\ell_1!\ell_2!} 4 \cdot (4dr^2)^{d+1} \cdot \|V - V^*\|_F^{\ell_1+\ell_2} \cdot \left(1 + \frac{\|\mathbf{c} - \mathbf{c}^*\|_2}{\|V^* - V\|_F}\right)^2\right)^{1/2} \\
&= 2(4dr^2)^{(d+1)/2} \sum_{\ell>1} \frac{1}{\ell!}\|V - V^*\|_F^\ell \cdot \left(1 + \frac{\|\mathbf{c} - \mathbf{c}^*\|_2}{\|V^* - V\|_F}\right) \\
&\le 2e(16dr^3)^{(d+1)/2} \cdot \|V - V^*\|_F \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2),
\end{aligned}$$

where the first step follows by Lemma 2.5.4, and the last step follows by the fact that $\|V - V^*\|_F \le 2\sqrt{r}$ and the fact that $\sum_{\ell=2}^{d+1} \frac{1}{\ell!} x^\ell < e \cdot (4r)^{(d-1)/2} \cdot x^2$ for $x \in [0, 2\sqrt{r}]$. □

Proof of Lemma 2.7.10. We have that

$$\frac{1}{2\eta_{\mathsf{vec}}} \left|\langle \tilde{\Delta}''_V, V - V^*\rangle\right|$$

$$\begin{aligned}
&= \left|\mathbb{E}\left[\sum_{\ell=2}^{d+1}\frac{1}{\ell!}\left\langle \nabla^{[\ell]}F_x(\Theta), (\Theta^*-\Theta)^{\otimes \ell}\right\rangle \cdot x^\top \cdot \Pi_V^\perp V^* \cdot \Delta\right]\right| \\
&\le \mathbb{E}\left[\left(\sum_{\ell=2}^{d+1}\frac{1}{\ell!}\left\langle \nabla^{[\ell]}F_x(\Theta), (\Theta^*-\Theta)^{\otimes \ell}\right\rangle\right)^2\right]^{1/2} \cdot \mathbb{E}\left[(x^\top \cdot \Pi_V^\perp V^* \cdot \Delta)^2\right]^{1/2} \\
&\le O(dr^3)^{(d+1)/2} \cdot \|V-V^*\|_F \cdot (\|V-V^*\|_F + \|\mathbf{c}-\mathbf{c}^*\|_2) \cdot \mathbb{E}[\textcircled{\scriptsize A'}]^{1/2} \\
&\le O(dr^3)^{(d+1)/2} \cdot \|V-V^*\|_F \cdot (\|V-V^*\|_F + \|\mathbf{c}-\mathbf{c}^*\|_2) \cdot (2d_P(V,V^*)),
\end{aligned}$$

where the second step follows by Cauchy-Schwarz, the third by Lemma 2.38 and the definition of $\textcircled{\scriptsize A'}$, the fourth by the upper bound in (2.31). □

2.11.4 Proof of Lemma 2.7.11

By Holder's,

$$\begin{aligned}
&|\mathbb{E}\left[\langle \mathcal{E}, V-V^*\rangle\right]| \\
&\quad\le \mathbb{E}\left[|\cos(\sigma\eta_{\mathsf{vec}})-1|\right]\cdot \sup_{\widehat{\nabla}}\left|\langle V\cdot\widehat{\nabla}\widehat{\nabla}^\top, V-V^*\rangle\right| + \mathbb{E}\left[|\sin(\sigma\eta_{\mathsf{vec}})-\sigma\eta_{\mathsf{vec}}|\right]\cdot\sup_{\widehat{h},\widehat{\nabla}}\left|\langle \widehat{h}\widehat{\nabla}^\top, V-V^*\rangle\right| \\
&\quad\le O(\eta_{\mathsf{vec}}^2)\cdot\mathbb{E}[\sigma^2]\cdot\left(\sup_{\widehat{\nabla}}\left|\langle V\cdot\widehat{\nabla}\widehat{\nabla}^\top, V-V^*\rangle\right| + \sup_{\widehat{h},\widehat{\nabla}}\left|\langle\widehat{h}\widehat{\nabla}^\top, V-V^*\rangle\right|\right), \qquad (2.39)
\end{aligned}$$

where in the second step we used that $|\cos(x)-1|\le x^2/2$ and $|\sin(x)-x|\le x^2/\pi$ for all $x\ge 0$, and in the third step we invoked Lemmas 2.11.2 and 2.11.3 below.

Lemma 2.11.2. *For any $\widehat{\nabla}\in\mathbb{S}^{r-1}$, $\left|\langle V\cdot\widehat{\nabla}\widehat{\nabla}^\top, V-V^*\rangle\right|\le\|V-V^*\|_F$.*

Proof. We may write the quantity on the left-hand side as

$$\widehat{\nabla}^\top\cdot\left((V-V^*)^\top V\right)\cdot\widehat{\nabla} = \widehat{\nabla}^\top\left(\mathrm{Id}-V^{*\top}V\right)\widehat{\nabla}\le\|\mathrm{Id}-V^{*\top}V\|_2\le\|V-V^*\|_F,$$

where the last step follows by the first part of Lemma 2.3.13. □

Lemma 2.11.3. *For any $\widehat{\nabla}\in\mathbb{S}^{r-1}$ and $\widehat{h}\in\mathbb{S}^{n-1}$ for which $\widehat{h}$ lies in the orthogonal complement of the column span of V, $\left|\left\langle\widehat{h}\widehat{\nabla}^\top, V-V^*\right\rangle\right|\le d_P(V,V^*)$.*

Proof. Because $\Pi_V^{\perp}\widehat{h} = \widehat{h}$, The left-hand side can be rewritten as

$$\widehat{h}^{\top}(V - V^*)\widehat{\nabla} = \widehat{h}^{\top}\Pi_V^{\perp}(V - V^*)\widehat{\nabla},$$

it is upper-bounded by

$$\begin{aligned}\sigma_{\max}(\Pi^{\perp}(V - V^*)) &\le \mathrm{Tr}((V - V^*)^{\top}(\mathrm{Id} - VV^{\top})(V - V^*)^{1/2} \\ &= \mathrm{Tr}(\mathrm{Id} - V^*V^{\top}VV^{*\top})^{1/2} \\ &= d_C(V, V^*) \le d_P(V, V^*),\end{aligned}$$

where the last step follows by Lemma 1.3.5. □

Proof of Lemma 2.7.11. We have

$$|\mathbb{E}\left[\langle \mathcal{E}, V - V^*\rangle\right]| \le O(\eta_{\mathsf{vec}}^2) \cdot O(n) \cdot O(dr^3)^{d+2} \cdot \|V - V^*\|_F \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}_*\|_2)^2 \,,$$

by (2.39), Lemmas 2.11.2, 2.11.3, and 2.7.4. The lemma follows by taking $\eta_{\mathsf{vec}} \le O(1/n)$. □

2.11.5 Proof of Lemma 2.7.15

Proof. We will bound each $\mathbb{E}_{x^0,\dots,x^{t-1}}[\mu_X(\Theta^{(t)}]$ individually. By Lemma 2.7.9, for any realization of $x^0, \dots, x^{t-1}$ giving rise to iterate $\Theta^{(t)} = (\mathbf{c}, V^{(t)})$, $\mu_X(\Theta^{(t)}) \ge (\nu_{\mathsf{cond}}/4) \cdot d_P(V^{(t)}, V^*)^2$. We have that

$$\begin{aligned}&\mathbb{E}\left[d_P(V^{(t)}, V^*)^2\right] \\ &\qquad \ge \mathbb{E}\left[\left(d_P(V^{(t-1)}, V^*) - d_P(V^{(t)}, V^{(t-1)})\right)^2\right] \\ &\qquad \ge \mathbb{E}\left[d_P(V^{(t-1}, V^*)^2\right] - 2\mathbb{E}\left[d_P(V^{(t-1)}, V^*)^2\right]^{1/2} \cdot \mathbb{E}\left[d_P(V^{(t)}, V^{(t-1)})^2\right]^{1/2} \\ &\qquad \ge \mathbb{E}\left[d_P(V^{(t-1}, V^*)^2\right] - 2\mathbb{E}\left[d_P(V^{(t-1)}, V^*)^2\right]^{1/2} \cdot \mathbb{E}\left[\|\Delta_{\mathsf{vec}}^{\Theta^{(t-1)},x^{t-1}}\|_F^2\right]^{1/2} \\ &\qquad \ge \mathbb{E}\left[d_P(V^{(t-1}, V^*)^2\right] - 6\sqrt{r} \cdot \eta_{\mathsf{vec}}\mathbb{E}\left[d_P(V^{(t-1)}, V^*)^2\right]^{1/2} \cdot \mathbb{E}\left[(\sigma^{\Theta^{(t-1)},x^{t-1}})^2\right]^{1/2} \quad (2.40)\end{aligned}$$

where the first step follows by triangle inequality (Fact 1.3.4), the second by Cauchy-Schwarz, the third by the definition of Procrustes distance, and the fourth by (2.24). By Lemma 2.7.4 and Lemma 2.7.7,

$$\begin{aligned}
&6\sqrt{r}\cdot\eta_{\mathsf{vec}}\mathbb{E}\left[(\sigma^{\Theta^{(t-1)},x^{t-1}})^2\right]^{1/2}\\
&\quad\le 6\sqrt{r}\cdot\eta_{\mathsf{vec}}\cdot O(\sqrt{n})\cdot(dr^3)^{(d+2)/2}\cdot\left(\mathbb{E}\left[\|V^{(t-1)}-V^*\|_F\right]+\|\mathbf{c}-\mathbf{c}^*\|_2\right)\\
&\quad\le 6\sqrt{r}\cdot O(\sqrt{n})\cdot(dr^3)^{(d+2)/2}\cdot\left(1.1d_P(V^{(0)},V^*)+\|\mathbf{c}-\mathbf{c}^*\|_2\right)\\
&\quad\le \frac{1}{100T}d_P(V^{(0)},V^*),
\end{aligned}$$

where the last step follows by our choice of η_{vec} in (2.21). So by (2.40) we conclude that as long as $\mathbb{E}[d_P(V^{(s)},V^*)^2] > d_P(V^{(0)},V^*)^2/1.1$ for all $s<t$,

$$\begin{aligned}
\mathbb{E}\left[d_P(V^{(t)},V^*)^2\right] &\ge \left(1-\frac{\sqrt{1.1}}{100T}\right)\mathbb{E}\left[d_P(V^{(t-1)},V^*)^2\right]\\
&\ge \left(1-\frac{\sqrt{1.1}}{100T}\right)^t d_P(V^{(0)},V^*)^2\\
&\ge d_P(V^{(0)},V^*)^2/1.1.
\end{aligned}$$

By induction, $d_P(V^{(t)},V^*)^2 \ge d_P(V^{(0)},V^*)^2/1.1$ for all $0\le t<T$. Recalling that $\mu_X(\Theta^{(t)}) \ge (\nu_{\mathsf{cond}}/4)\cdot d_P(V^{(t)},V^*)^2$, we conclude that

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\mu_X(\Theta^{(t)})\right] \ge T\cdot(\nu_{\mathsf{cond}}/4)\cdot\left(d_P(V^{(0)},V^*)^2/1.1\right)$$

as desired. □

2.11.6 Proof of Lemma 2.7.16

Proof. We will bound each $\mathbb{E}_{x^0,\dots,x^{t-1}}[\left|\mu_{E_1}(\Theta^{(t)}\right|]$ individually and apply triangle inequality.

By Lemma 2.7.10, for any realization of $x^0, ..., x^{t-1}$ giving rise to iterate $\Theta^{(t)} = (\mathbf{c}, V^{(t)})$,

$$\left|\mu_{E_1}(\Theta^{(t)})\right| \leq O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot \|V^{(t)} - V^*\|_F \cdot d_P(V^{(t)}, V^*) \cdot \left(\|V^{(t)} - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2\right).$$

$$\leq O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot \left(\|V^{(t)} - V^*\|_F^3 + \|V^{(t)} - V^*\|_F^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2\right).$$

By Lemma 2.7.7 and (2.20), we conclude that

$$\begin{aligned}\mathbb{E}\left[\left|\mu_{E_1}(\Theta^{(t)})\right|\right] &\leq O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot \left(1.1 d_P(V^{(0)}, V^*)^3 + 1.1 d_P(V^{(0)}, V^*)^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2\right) \\ &\leq O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot d_P(V^{(0)}, V^*)^3.\end{aligned}$$

The claim follows by summing over t. □

2.11.7 Proof of Lemma 2.7.17

Proof. We will bound each $\mathbb{E}_{x^0,...,x^{t-1}}[\left|\mu_{E_2}(\Theta^{(t)}\right|]$ individually and apply triangle inequality.

By Lemma 2.7.11, for any realization of $x^0, ..., x^{t-1}$ giving rise to iterate $\Theta^{(t)} = (\mathbf{c}, V^{(t)})$,

$$\begin{aligned}&\left|\mu_{E_2}(\Theta^{(t)})\right| \\ &\quad\leq \eta_{\mathsf{vec}} \cdot O(dr^3)^{d+2} \cdot \|V - V^*\|_F \cdot \left(\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2\right)^2. \\ &\quad\leq \eta_{\mathsf{vec}} \cdot O(dr^3)^{d+2} \cdot \left(\|V^{(t)} - V^*\|_F^3 + 2\|V^{(t)} - V^*\|_F^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 + \|V^{(t)} - V^*\|_F \cdot \|\mathbf{c} - \mathbf{c}^*\|_2\right).\end{aligned}$$

By Lemma 2.7.7 and (2.20), we conclude that

$$\mathbb{E}\left[\left|\mu_{E_2}(\Theta^{(t)})\right|\right] \leq O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{d+2} \cdot d_P(V^{(0)}, V^*)^3$$

The claim follows by summing over t. □

2.11.8 Proof of Lemma 2.7.18

Analogous to the proof of Lemma 2.6.15 in Appendix 2.10.4, we will prove concentration by decomposing the MDS $\{\mu_X(\Theta^{(t)}) - X^{\Theta^{(t)},x^t}\}_{0\leq t<T}$ into components corresponding to the decomposition (2.29). That is, define $\textcircled{\text{A'}}^{\Theta,x}$, $\textcircled{\text{B'}}^{\Theta,x}$, $\textcircled{\text{C'}}^{\Theta,x}$ to be the quantities in (2.29) for an iterate Θ and sample x. So by Observation 2.7.12, $\left\{\frac{1}{2\eta_{\mathsf{vec}}}\mu_X(\Theta^{(t)}) - \textcircled{\text{A'}}^{\Theta^{(t)},x^t}\right\}$, $\{\textcircled{\text{B'}}^{\Theta^{(t)},x^t}\}$, and $\{\textcircled{\text{C'}}^{\Theta^{(t)},x^t}\}$ are MDS's, and for any Θ, x,

$$\frac{1}{2\eta_{\mathsf{vec}}} X^{\Theta,x} = \textcircled{\text{A'}}^{\Theta,x} + \textcircled{\text{B'}}^{\Theta,x} + \textcircled{\text{C'}}^{\Theta,x}$$

by (2.29). We will show concentration for these MDS's separately.

Lemma 2.11.4.

$$\sum_{t=0}^{T-1} \textcircled{A'}^{\Theta^{(t)},x^t} \geq T \cdot (\nu_{\mathsf{cond}}/5) \cdot d_P(V^{(0)}, V^*)^2$$

with probability at least $1-\delta$.

Lemma 2.11.5.

$$\left|\sum_{t=0}^{T-1} \textcircled{B'}^{\Theta^{(t)},x^t}\right| \leq T \cdot (\nu_{\mathsf{cond}}/60) \cdot d_P(V^{(0)}, V^*)^2$$

with probability at least $1-\delta$.

Lemma 2.11.6.

$$\left|\sum_{t=0}^{T-1} \textcircled{C'}^{\Theta^{(t)},x^t}\right| \leq T \cdot (\nu_{\mathsf{cond}}/60) \cdot d_P(V^{(0)}, V^*)^2$$

with probability at least $1-\delta$.

We prove these in the subsequent Appendices 2.11.8, 2.11.8, and 2.11.8. Note that Lemma 2.7.18 follows easily from these three lemmas:

Proof of Lemma 2.7.18. The claim follows immediately from Lemmas 2.11.4, 2.11.5, and 2.11.6; triangle inequality; replacing 3δ in the resulting union bound with δ; and absorbing constant factors. □

Proof of Lemma 2.11.4

Proof. Observe that $\left\{\frac{1}{2\eta_{\mathsf{vec}}}\mu_X(\Theta^{(t)}) - \textcircled{A'}^{\Theta^{(t)},x^t}\right\}$ is an MDS which satisfies one-sided bounds, as $\textcircled{A'}^{\Theta,x} \geq 0$ with probability one for any Θ, x, so we wish to apply Lemma 2.3.4. To do so, we just need to bound the variances of the differences.

Lemma 2.11.7. *For any* Θ, $\mathbb{V}_x[\textcircled{A'}^{\Theta,x}] \leq 2^{4d+4} \cdot d_P(V, V^*)^4$.

Proof. We will suppress superscripts Θ, x in this proof. $\mathbb{V}[\textcircled{A'}] \leq \mathbb{E}[\textcircled{A'}^2]$, so it suffices to bound the latter. But note that $x^\top \Pi_V^\perp V^* \nabla p(V^\top x)$ is a polynomial, call it $f(x)$, of degree d in the Gaussians $x_1, ..., x_n$. By Fact 1.3.15,

$$\mathbb{E}[\textcircled{A'}^2] = \mathbb{E}[f(x)^4] \leq \left(4^{d/2} \cdot \mathbb{E}[f(x)^2]^{1/2}\right)^4 \leq 2^{4d} \cdot \mathbb{E}[f(x)^2]^2 = 2^{4d} \cdot \mathbb{E}[\textcircled{A'}]^2 \leq 2^{4d+4} \cdot d_P(V, V^*)^4, \tag{2.41}$$

where the last step is by Lemma 2.7.13. □

We can now complete the proof of Lemma 2.11.4. By Lemma 2.7.6 and Lemma 2.11.7, if η_{vec} satisfies (2.21), then with probability $1 - \delta$ we have that for all $0 \leq t < T$,

$$\frac{1}{2\eta_{\mathsf{vec}}}\mu_X(\Theta^{(t)}) - \textcircled{A'}^{\Theta^{(t)},x^t} \leq \frac{1}{2\eta_{\mathsf{vec}}}\mu_X(\Theta^{(t)}) \leq 4 d_P(V^{(t)}, V^*)^2 \leq 4.84 d_P(V^{(0)}, V^*)^2.$$

$$\mathop{\mathbb{V}}_{x^t}[\textcircled{A'}^{\Theta^{(t)},x^t}] \leq 1.1^4 \cdot 2^{4d+4} \cdot d_P(V^{(0)}, V^*)^4$$

Applying Lemma 2.3.4 with the parameter σ^2 taken to be $T \cdot 1.1^4 \cdot 2^{4d+4} \cdot d_P(V^{(0)}, V^*)^4$, we get

$$\Pr\left[\sum_{t=0}^{T-1} \textcircled{A'}^{\Theta^{(t)},x^t} \geq \frac{1}{2\eta_{\mathsf{vec}}} \sum_{t=0}^{T-1} \mathbb{E}\left[\mu_X\left(\Theta^{(t)}\right)\right] - O\left(4^d \log(1/\delta)\sqrt{T} \cdot d_P(V^{(0)}, V^*)^2\right)\right] \geq 1 - 2\delta,$$

where the expectation in $\mathbb{E}\left[X\left(\Theta^{(t)}\right)\right]$ is over the randomness of the samples $x^0, ..., x^{t-1}$.

By Lemma 2.7.15, we conclude that

$$\sum_{t=0}^{T-1} \textcircled{A'}^{\Theta^{(t)},x^t} \geq T \cdot (\nu_{\mathsf{cond}}/4.4) \cdot d_P(V^{(0)}, V^*)^2 - O\left(4^d \log(1/\delta)\sqrt{T} \cdot d_P(V^{(0)}, V^*)^2\right) \tag{2.42}$$

with probability at least $1 - 2\delta$. Taking T according to (2.22) will certainly ensure the right-hand side of (2.42) is at least $T \cdot (\nu_{\mathsf{cond}}/5) \cdot d_P(V^{(0)}, V^*)^2$. The proof is completed by replacing 2δ in the above with δ and absorbing the resulting constant factors. $\square$

Proof of Lemma 2.11.5

Proof. For fixed $x^1, ..., x^{t-1}$, the martingale difference $\textcircled{B'}^{\Theta^{(t)},x^t}$ is a polynomial of degree $2d$ in x^t, so by Lemma 2.3.3 we just need to upper bound the second moments of the differences, which we do in the following lemma.

Lemma 2.11.8. *For any Θ, $\mathbb{E}_x[(\textcircled{B'}^{\Theta,x})^2] \le d_P(V, V^*)^2 \cdot \|V - V^*\|_F^2 \cdot O(r^2) \cdot \exp(O(d))$.*

Proof. By Cauchy-Schwarz,

$$\begin{aligned}
\mathbb{E}\left[\textcircled{B'}^2\right] &\le \mathbb{E}\left[\left(x^\top \Pi_V (V^* - V)\nabla\right)^4\right]^{1/2} \cdot \mathbb{E}\left[\left(x^\top \Pi_V^\perp V^* \nabla\right)^4\right]^{1/2} \\
&= \mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{Id}_r)}\left[\left(g^\top V^\top (V^* - V)\nabla p(g)\right)^4\right]^{1/2} \cdot \mathbb{E}\left[\textcircled{A'}^2\right]^{1/2} \\
&\le \mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{Id}_r)}\left[\left(g^\top (\mathrm{Id} - V^\top V^*)\nabla p(g)\right)^4\right]^{1/2} \cdot 2^{2d+2} \cdot d_P(V, V^*)^2, \qquad (2.43)
\end{aligned}$$

where the third step follows by (2.41). It remains to bound the first factor in (2.43). As this factor is independent of n, we do not need a particularly sharp bound. We have

$$\begin{aligned}
\mathbb{E}_g\left[\left(g^\top (\mathrm{Id} - V^\top V^*)\nabla p(g)\right)^4\right]^{1/2} &\le \|\mathrm{Id} - V^\top V^*\|_2^2 \cdot \mathbb{E}_g[\|g\|_2^4 \cdot \|\nabla p(g)\|_2^4]^{1/2} \\
&\le \|V - V^*\|_F^2 \mathbb{E}_g[\|g\|_2^8]^{1/4} \cdot \mathbb{E}_g[\|\nabla p(g)\|_2^8]^{1/4} \\
&\le \|V - V^*\|_F^2 \cdot 3(r+1) \cdot (rd \cdot 7^d \cdot \mathbb{V}[p]) \\
&\le \|V - V^*\|_F^2 \cdot O(r^2 d \cdot 7^d),
\end{aligned}$$

where the second step follows by Lemma 2.3.13, the third step follows by Corollary 1.3.17 and Lemma 2.3.8 applied to $q = 4$, and the last step follows by noting that $\mathbb{V}[p] = O(1)$ by triangle inequality and absorbing constant factors. The claimed bound follows. $\square$

We now complete the proof of Lemma 2.11.5. By Lemma 2.7.6, $d_P(V^{(t)}, V^*) \le \|V^{(t)} - V^*\|_F \le 1.1 \cdot d_P(V^{(0)}, V^*)$ for every $0 \le t \le T$ with probability at least $1 - \delta$, in which case

Lemma 2.11.8 implies that for every $0 \le t < T$,

$$\mathbb{E}\left[\left(\textcircled{B'}^{\Theta^{(t)},x^t}\right)^2 \,\middle|\, x^1, ..., x^{t-1}\right] \le d_P(V^{(0)}, V^*)^4 \cdot O(r^2) \cdot \exp(O(d))$$

with probability at least $1 - \delta$. So by Lemma 2.3.3,

$$\left|\sum_{t=0}^{T-1} \textcircled{B'}^{\Theta^{(t)},x^t}\right| \le (\log(1/\delta) \cdot d)^{c_1 d} \cdot \sqrt{T} \cdot d_P(V^{(0)}, V^*)^2 \cdot O(r) \cdot \exp(O(d))$$

with probability at least $1 - 2\delta$. By taking T according to (2.22), we ensure that this quantity is upper bounded by a negligible multiple of $T \cdot (\nu_{\mathsf{cond}}/5) \cdot d_P(V^{(0)}, V^*)^2$ as desired. The proof is completed by replacing 2δ in the above with δ and absorbing the resulting constant factors. □

Proof of Lemma 2.11.6

Proof. As in the proof of Lemma 2.11.5, for fixed $x^1, ..., x^{t-1}$, the martingale difference $\textcircled{C'}^{\Theta^{(t)},x^t}$ is a polynomial of degree $2d$ in x^t, so by Lemma 2.3.3 we just need to upper bound the second moments of the differences, which we do in the following lemma.

Lemma 2.11.9. *For any Θ, $\mathbb{E}_x[(\textcircled{C'}^{\Theta,x})^2] \le d_P(V, V^*)^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^2 \cdot \exp(O(d))$.*

Proof. By Cauchy-Schwarz,

$$\begin{aligned}
\mathbb{E}\left[\textcircled{C'}^2\right] &\le \mathbb{E}\left[(\updelta(V^\top x)^4\right]^{1/2} \cdot \mathbb{E}\left[\left(x^\top \Pi_V^\perp V^* \nabla\right)^4\right]^{1/2} \\
&= \mathbb{E}_{g \sim \mathcal{N}(0, \mathrm{Id}_r)}\left[\updelta(g)^4\right]^{1/2} \cdot \mathbb{E}\left[\textcircled{A'}^2\right]^{1/2} \\
&\le \left(3^d \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^2\right) \cdot \left(2^{2d+2} \cdot d_P(V, V^*)^2\right),
\end{aligned}$$

where the third step follows by Fact 1.3.15 and (2.41). □

We now complete the proof of Lemma 2.11.6. By Lemma 2.7.6, $d_P(V^{(t)}, V^*) \le \|V^{(t)} - V^*\|_F \le 1.1 \cdot d_P(V^{(0)}, V^*)$ for every $0 \le t \le T$ with probability at least $1 - \delta$, in which case

Lemma 2.11.9 implies that for every $0 \le t < T$,

$$\mathbb{E}[(\textcircled{C'}^{\Theta^{(t)},x^t})^2 | x^1, ..., x^{t-1}] \le d_P(V^{(0)}, V^*) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot \exp(O(d))$$

with probability at least $1 - \delta$. So by Lemma 2.3.3,

$$\left| \sum_{t=0}^{T-1} \textcircled{C'}^{\Theta^{(t)},x^t} \right| \le (\log(1/\delta) \cdot d)^{c_1 d} \cdot \sqrt{T} \cdot d_P(V^{(0)}, V^*) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot \exp(O(d))$$

with probability at least $1 - 2\delta$. By taking T satisfying the bound in the lemma statement and invoking (2.20), we ensure that this quantity is upper bounded by a negligible multiple of $T \cdot (\nu_{\mathsf{cond}}/5) \cdot d_P(V^{(0)}, V^*)^2$ as desired. As in the proof of Lemma 2.11.4, the proof is completed by replacing 2δ in the above with δ and absorbing the resulting constant factors. □

Proof of Lemmas 2.7.19 and 2.7.20

We will apply Lemma 2.3.3 to the MDS's $\left\{ E_1^{\Theta^{(t)},x^t} - \mu_{E_1}(\Theta^{(t)}) \right\}$ and $\left\{ E_2^{\Theta^{(t)},x^t} - \mu_{E_2}(\Theta^{(t)}) \right\}$. As in the analysis of the MDS's for Lemmas 2.11.5 and 2.11.6, the differences in these MDS's are polynomials of degree at most $2d$, so we just need to bound the second moments of their differences. We do so in the following two lemmas.

Lemma 2.11.10. *For any* Θ*,*

$$\mathbb{E}_x[(E_1^{\Theta,x})^2] \le O(\eta_{\mathsf{vec}}^2) \cdot O(dr^3)^{d+1} \cdot \|V - V^*\|_F^2 \cdot d_P(V, V^*)^2 \cdot (\|\mathbf{c} - \mathbf{c}^*\|_2 + \|V^* - V\|_F)^2$$

Proof. We have that

$$\begin{aligned}
\frac{1}{4\eta_{\mathsf{vec}}^2} \mathbb{E}\left[\left(E_1^{\Theta,x} \right)^2 \right] &= \mathbb{E}\left[(\mathfrak{R}^{\Theta,x})^2 \cdot \left(x^\top \cdot \Pi_V^\perp V^* \cdot \Delta \right)^2 \right] \\
&\le \mathbb{E}\left[(\mathfrak{R}^{\Theta,x})^4 \right]^{1/2} \cdot \mathbb{E}\left[\left(x^\top \cdot \Pi_V^\perp V^* \cdot \Delta \right)^4 \right]^{1/2} \\
&= \mathbb{E}\left[(\mathfrak{R}^{\Theta,x})^4 \right]^{1/2} \cdot \mathbb{E}[\textcircled{A'}^2]^{1/2} \\
&\le O(dr^3)^{d+1} \cdot \|V - V^*\|_F^2 \cdot d_P(V, V^*)^2 \cdot (\|\mathbf{c} - \mathbf{c}^*\|_2 + \|V^* - V\|_F)^2,
\end{aligned}$$

where the second step follows by Cauchy-Schwarz, the third step follows by definition of $\textcircled{A'}$, and the fourth step follows by Lemma 2.6.5. □

Lemma 2.11.11. *For any Θ, if $\eta_{\mathsf{vec}} \le O(1/n)$, then*

$$\mathbb{E}_x[(E_2^{\Theta,x})^2] \le O(\eta_{\mathsf{vec}}^2) \cdot (64dr^3)^{2d+4} \cdot \|V - V^*\|_F^2 \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^4$$

Proof. By triangle inequality and Jensen's, $\mathbb{E}[(E_2^{\Theta,x})^2]^{1/2} = \mathbb{E}[\langle \mathcal{E}, V - V^*\rangle^2]^{1/2}$ is at most

$$\mathbb{E}\left[(\cos(\sigma\eta_{\mathsf{vec}}) - 1)^2 \cdot \langle V \cdot \widehat{\nabla}\widehat{\nabla}^\top, V - V^*\rangle^2\right]^{1/2} + \mathbb{E}\left[(\sin(\sigma\eta_{\mathsf{vec}}) - \sigma\eta_{\mathsf{vec}})^2 \cdot \langle \widehat{h}\widehat{\nabla}^\top, V - V^*\rangle^2\right]^{1/2}$$

By Holder's and the fact that $|\cos(x) - 1| \le x^2/2$ and $|\sin(x) - x| \le x^2/\pi$ for all $x \ge 0$, we may upper bound the first term by

$$\mathbb{E}\left[(\cos(\sigma\eta_{\mathsf{vec}}) - 1)^2\right]^{1/2} \cdot \max_{\widehat{\nabla}} \left|\langle V \cdot \widehat{\nabla}\widehat{\nabla}^\top, V - V^*\rangle\right| \le O(\eta_{\mathsf{vec}}^2) \cdot \mathbb{E}[\sigma^4]^{1/2} \cdot \max_{\widehat{\nabla}} \left|\langle V \cdot \widehat{\nabla}\widehat{\nabla}^\top, V - V^*\rangle\right|$$

and the second term by

$$\mathbb{E}\left[(\sin(\sigma\eta_{\mathsf{vec}}) - \sigma\eta_{\mathsf{vec}})^2\right]^{1/2} \cdot \max_{\widehat{h},\widehat{\nabla}} \left|\langle \widehat{h}\widehat{\nabla}^\top, V - V^*\rangle\right| \le O(\eta_{\mathsf{vec}}^2) \cdot \max_{\widehat{h},\widehat{\nabla}} \left|\langle \widehat{h}\widehat{\nabla}^\top, V - V^*\rangle\right|.$$

So $\mathbb{E}[(E_2^{\Theta,x})^2]^{1/2}$ is at most

$$\begin{aligned}
&O(\eta_{\mathsf{vec}}^2) \cdot \mathbb{E}[\sigma^4]^{1/2} \cdot \left(\max_{\widehat{\nabla}} \left|\langle V \cdot \widehat{\nabla}\widehat{\nabla}^\top, V - V^*\rangle\right| + \max_{\widehat{h},\widehat{\nabla}} \left|\langle \widehat{h}\widehat{\nabla}^\top, V - V^*\rangle\right|\right) \\
&\quad\le O(\eta_{\mathsf{vec}}^2) \cdot O(n) \cdot (64dr^3)^{d+2} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2 \cdot \|V - V^*\|_F \\
&\quad\le O(\eta_{\mathsf{vec}}) \cdot (64dr^3)^{d+2} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2 \cdot \|V - V^*\|_F,
\end{aligned}$$

where the first step follows by Lemma 2.7.4, Lemma 2.11.2, and Lemma 2.11.3, and the sixth follows by the assumption that $\eta_{\mathsf{vec}} \le O(1/n)$. □

We are now ready to complete the proofs of Lemma 2.7.19 and 2.7.20.

Proof of Lemma 2.7.19. By Lemma 2.7.6, $d_P(V^{(t)}, V^*) \le \|V^{(t)} - V^*\|_F \le 1.1 \cdot d_P(V^{(0)}, V^*)$ for every $0 \le t \le T$ with probability at least $1 - \delta$, in which case Lemma 2.11.10 implies that for every $0 \leq t < T$,

$$\mathbb{E}[(E_1^{\Theta^{(t)},x^t})^2 | x^1, ..., x^{t-1}] \leq O(\eta_{\mathsf{vec}}^2) \cdot O(dr^3)^{d+1} \cdot d_P(V^{(0)}, V^*)^4 \cdot \left(\|\mathbf{c} - \mathbf{c}^*\|_2 + d_P(V^{(0)}, V^*)\right)^2$$

$$\leq O(\eta_{\mathsf{vec}}^2) \cdot O(dr^3)^{d+1} \cdot d_P(V^{(0)}, V^*)^6$$

with probability at least $1 - \delta$. So by Lemma 2.3.3,

$$\left|\sum_{t=0}^{T-1} \left(E_1^{\Theta^{(t)},x^t} - \mathbb{E}\left[\mu_{E_1}(\Theta^{(t)})\right]\right)\right|$$
$$\leq (\log(1/\delta) \cdot d)^{c_1 d} \cdot \sqrt{T} \cdot O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot d_P(V^{(0)}, V^*)^3$$

with probability at least $1 - 2\delta$. By Lemma 2.7.16, we conclude that

$$\left|\sum_{t=0}^{T-1} E_1^{\Theta^{(t)},x^t}\right| \leq O(\sqrt{T} \cdot \eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot d_P(V^{(0)}, V^*)^3 \cdot \left((\log(1/\delta) \cdot d)^{c_1 d} + \sqrt{T}\right)$$

By taking T according to (2.22) and using the bound (2.19), we ensure that this quantity is upper bounded by a negligible multiple of $T \cdot \eta_{\mathsf{vec}} \cdot (\nu_{\mathsf{cond}}/3) \cdot d_P(V^{(0)}, V^*)^2$ as desired. As usual, the proof is completed by replacing 2δ in the above with δ and absorbing the resulting constant factors. □

Proof of Lemma 2.7.20. By Lemma 2.7.6, $\|V^{(t)} - V^*\|_F \leq 1.1 \cdot d_P(V^{(0)}, V^*)$ for every $0 \leq t \leq T$ with probability at least $1 - \delta$, in which case Lemma 2.11.11 implies that for every $0 \leq t < T$,

$$\mathbb{E}[(E_2^{\Theta^{(t)},x^t})^2 | x^1, ..., x^{t-1}] \leq O(\eta_{\mathsf{vec}}^2) \cdot O(dr^3)^{2d+4} \cdot d_P(V^{(0)}, V^*)^2 \cdot \left(\|\mathbf{c} - \mathbf{c}^*\|_2 + d_P(V^{(0)}, V^*)\right)^4$$

$$\leq O(\eta_{\mathsf{vec}}^2) \cdot O(dr^3)^{2d+4} \cdot d_P(V^{(0)}, V^*)^6$$

with probability at least $1-\delta$. So by Lemma 2.3.3,

$$\left|\sum_{t=0}^{T-1}\left(E_2^{\Theta^{(t)},x^t}\sum_{t=0}^{T-1}\mathbb{E}\left[\mu_{E_2}(\Theta^{(t)})\right]\right)\right|$$
$$\leq(\log(1/\delta)\cdot d)^{c_1 d}\cdot\sqrt{T}\cdot O(\eta_{\mathsf{vec}})\cdot O(dr^3)^{d+2}\cdot d_P(V^{(0)},V^*)^3$$

with probability at least $1-2\delta$. By (2.7.17), we conclude that

$$\left|\sum_{t=0}^{T-1}E_2^{\Theta^{(t)},x^t}\right|\leq O(\sqrt{T}\cdot\eta_{\mathsf{vec}})\cdot O(dr^3)^{d+2}\cdot d_P(V^{(0)},V^*)^3\cdot\left((\log(1/\delta)\cdot d)^{c_1 d}+\sqrt{T}\right)$$

By taking T according to (2.22) and using the bound (2.19), we ensure that this quantity is upper bounded by a negligible multiple of $T\cdot\eta_{\mathsf{vec}}\cdot(\nu_{\mathsf{cond}}/3)\cdot d_P(V^{(0)},V^*)^2$ as desired. As usual, the proof is completed by replacing 2δ in the above with δ and absorbing the resulting constant factors. □

Chapter 3

Deep ReLU Networks

3.1 Introduction

In this chapter, we turn from low-rank polynomials to the following class of concepts, originally introduced in Definition 1.2.4:

Definition 3.1.1 (ReLU Networks)**.** *Let $\mathcal{C}_S$ denote the concept class of (feedforward) ReLU networks over $\mathbb{R}^d$ of size S. Specifically, $F \in \mathcal{C}_S$ if there exist weight matrices $\mathbf{W}_0 \in \mathbb{R}^{k_0 \times d}, \mathbf{W}_1 \in \mathbb{R}^{k_1 \times k_0}, \ldots, \mathbf{W}_L \in \mathbb{R}^{k_L \times k_{L-1}}, \mathbf{W}_{L+1} \in \mathbb{R}^{1 \times k_L}$ for which*

$$F(x) \triangleq \mathbf{W}_{L+1}\phi\left(\mathbf{W}_L\phi\left(\cdots\phi(\mathbf{W}_0 x)\cdots\right)\right),$$

where $\phi(z) \triangleq \max(z, 0)$ is the ReLU activation applied entrywise, and $k_0 + \cdots + k_L = S$. In this case we say that F is computed by a ReLU network with depth $L + 2$. We will refer to the rank of $\mathbf{W}_0$ as k, to emphasize that the value of F only depends on a k-dimensional subspace of $\mathbb{R}^d$. We will also let $k_{L+1} = 1$.

When the weight matrices of two ReLU networks $F, F' \in \mathcal{C}_S$ have the same dimensions (at all layers), then we say that F and F' *have the same architecture.*

For example, a depth two ReLU network of size S in d-dimensions is a function $F : \mathbb{R}^d \to \mathbb{R}$ of the form

$$F(x) = \sum_{i=1}^{S} \lambda_i \phi(\langle w_i, x\rangle),$$

where $\lambda_i \in \mathbb{R}$ are scalars and $w_i \in \mathbb{R}^d$ are arbitrary vectors.

Note that any Boolean function $F : \{\pm 1\}^n \to \{\pm 1\}$ can be computed by an n-layer ReLU network (Lemma 3.5.2). In particular, if F is a junta depending only on k variables, then it can be computed by a k-layer ReLU network with size that depends only on k.

Learning ReLU Networks The problem of PAC learning an unknown ReLU network from labeled examples is a central challenge in the theory of machine learning. Given samples from a distribution of the form $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ where $y = F(x)$ with F an unknown size-S ReLU network,[1] and x is drawn according to a distribution $\mathcal{D}$, the goal is to output a function $f : \mathbb{R}^d \to \mathbb{R}$ with small *test* error, i.e., $\mathbb{E}_{x,y}[(y - f(x))^2] \leq \varepsilon\,\mathbb{E}[y^2]$. In this thesis, we focus on the widely studied case where the input distribution on x is Gaussian.

Ideally, we would like an algorithm with sample complexity and running time that is polynomial in all the relevant parameters. Even for learning arbitrary sums of ReLUs, i.e. depth two ReLU networks where we additionally assume the $\mathbf{W}_1$ has all positive entries, it remains a major open question to obtain a polynomial-time algorithm (see [DK20] for the strongest-known result). As a first step, one could ask for an algorithm that at least depends polynomially on the *ambient dimension* (it is often easy to obtain brute-force search algorithms that run in time exponential in the dimension[2]). In the absence of additional assumptions however, even this goal has remained elusive: it was not known how to achieve a subexponential-time algorithm even for learning *general* depth two ReLU networks, let alone ReLU networks of higher depth.

In this chapter, we address this gap by giving the first algorithm for learning ReLU networks whose running time is a fixed polynomial in the dimension, regardless of the depth of the network. Our algorithm is *fixed-parameter tractable*: we show that we can *properly learn* (i.e., the output hypothesis is also a ReLU network) ReLU networks with sample complexity and running time that is a fixed polynomial in the dimension and an exponential function of the network's *parameters*.

More precisely, our main result is as follows. We will also make the (as it turns out

[1] It should not be difficult to extend our techniques to the setting where $y = F(x) + \mathcal{N}(0, \sigma^2)$, but we focus on the noiseless case for simplicity in this thesis.

[2] Although in our specific case even this type of search turns out to be nontrivial.

necessary) assumption that the ReLU network has a bounded Lipschitz constant (recall that a function $f : \mathbb{R}^d \to \mathbb{R}$ is Λ-Lipschitz if $|f(x) - f(x')| \le \Lambda \|x - x'\|_2$ for all x, x').

Theorem 3.1.2 (Main, see Theorem 3.4.2 for formal statement)**.** *Let $\mathcal{D}$ be the distribution over pairs $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ where $x \sim \mathcal{N}(0, Id)$ and $y = F(x)$ for a size-S ReLU network F with depth $L + 2$, Lipschitz constant at most Λ, rank of* bottom *weight matrix $\mathbf{W}_0$ being k, and whose weight matrices all have spectral norm at most B.*

There is an algorithm that draws $d \log(1/\delta) \exp(\mathrm{poly}(k, S, \Lambda/\varepsilon))\, B^{O(Lk)}$ samples, runs in time $\widetilde{O}(d^2 \log(1/\delta)) \exp(\mathrm{poly}(k, S, \Lambda/\varepsilon))\, B^{O(LkS^2)}$, and outputs a ReLU network $\widetilde{F}$ such that $\mathbb{E}[(y - \widetilde{F}(x))^2] \le \varepsilon$ with probability at least $1 - \delta$.[3]

Note that the sample complexity is linear while the run-time is quadratic in the ambient dimension. In particular, in the well-studied special case where the product of the spectral norms of the weight matrices is a constant (see e.g. [GRS18]), in which case the Lipschitz constant of the network is also constant, we can obtain the following result as an immediate consequence of the formal version of the above theorem (Theorem 3.4.2):

Corollary 3.1.3. *Let $\mathcal{D}$ be the distribution over pairs $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ where $x \sim \mathcal{N}(0, Id)$ and $y = F(x)$ for a size-S ReLU network F for which the product of the spectral norms of its weight matrices is a constant.*

Then there is an algorithm that draws $N = d \log(1/\delta) \exp(O(k^3/\varepsilon^2 + kS))$ samples, runs in time $\widetilde{O}(d^2 \log(1/\delta)) \exp(O(k^3 S^2/\varepsilon^2 + kS^3))$, and outputs a ReLU network $\widetilde{F}$ such that $\mathbb{E}[(y - \widetilde{F}(x))^2] \le \varepsilon$ with probability at least $1 - \delta$.

As mentioned earlier, no algorithms that were sub-exponential in d were known even for S, B, ε being constants.

Before going further, we note that a dependence on the Lipschitz constant of the network is necessary even for learning depth two ReLU networks with respect to Gaussians:

Example 3.1.4. *Let $\Lambda > 0$. Consider the size-3, depth two ReLU network $F : \mathbb{R}^2 \to \mathbb{R}$ given by*

$$F(x_1, x_2) = \phi(x_1 + \Lambda x_2) + \phi(3x_1 + \Lambda x_2) - 2\phi(-x_1 + \Lambda x_2).$$

The Lipschitz constant of F is $\Theta(\Lambda)$: $F(0, 1/\Lambda) = 1$ and $F(1, 1/\Lambda) = 2$. Furthermore, note

[3] See Remark 3.4.3 for a discussion of why this guarantee is scale-invariant.

that for $(x_1, x_2) \in \mathbb{S}^1$*,* $F(x_1, x_2) = 0$ *unless* $x_2 \in [-3/\Lambda, 3/\Lambda]$*. By rotational symmetry, for* $(x_1, x_2) \sim \mathcal{N}(0, Id)$*,* $F(x_1, x_2) \neq 0$ *with probability at most* $O(1/\Lambda)$*.*

Note that for depth two ReLU networks with positive weights, no such dependence on the Lipschitz constant is necessary intuitively because without cancellations between the hidden units, one cannot devise "spiky" functions F which simultaneously have small variance but attain a large value at some bounded-norm x.

Interestingly, our techniques are also general enough to handle the more general family of all continuous piecewise-linear functions (see Definition 3.3.4 for a formal definition):

Theorem 3.1.5 (See Theorem 3.4.1 for formal statement)**.** *Let* $\mathcal{D}$ *be the distribution over pairs* $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ *where* $x \sim \mathcal{N}(0, Id)$ *and* $y = F(x)$ *for a continuous piecewise-linear function* F *which only depends on the projection of* x *to a* k*-dimensional subspace* V*, has at most* M *linear pieces, and is* Λ*-Lipschitz.*

There is an algorithm that draws $d \log(1/\delta) \cdot \mathrm{poly}\left(\exp\left(k^3\Lambda^2/\varepsilon^2\right), M^k\right)$ *samples, runs in time* $\widetilde{O}(d^2 \log(1/\delta)) \cdot M^{M^2} \cdot \mathrm{poly}\left(\exp\left(k^4\Lambda^2/\varepsilon^2\right), M^{k^2}\right)$*, and outputs a piecewise-linear function* $\widetilde{F}$ *such that* $\mathbb{E}[(y - \widetilde{F}(x))^2] \leq \varepsilon$ *with probability at least* $1 - \delta$*.*

Note that a size-S ReLU network is a continuous piecewise-linear function with at most 2^S linear pieces. Specializing Theorem 3.1.5 to ReLU networks gives a guarantee which is incomparable to Theorem 3.1.2: we obtain an algorithm that depends doubly exponentially on S but has no dependence on the norms of the weight matrices.

3.1.1 Prior Work on Provably Learning Neural Networks

Algorithmic Results Algorithms for learning neural networks (obtaining small *test error*) have been intensely studied in the literature. In the last few years alone there have been many papers giving provable results for learning restricted classes of neural networks under various settings [JSA15, ZLJ16, ZSJ+17, BG17, GKKT17, LY17, ZPS17, Tia17, GKM18, DLT18, GLM17, GKLW18, MR18, BJW18, GK19, AZLL19, VW19, ZYWG19, DGK+20, GMOV18, LMZ20, DK20].

The predominant techniques are spectral or tensor-based dimension reduction [JSA15, ZSJ+17, BJW18, DKKZ20], kernel methods [ZLJ16, GKKT17, Dan17, MR18, GK19], and gradient-based methods [GLM17, GKLW18, VW19]. All prior work takes distributional and/or architectural assumptions, the most common one being that the inputs come from a standard Gaussian. We will also work in this setting.[4]

As pointed out in [GGJ+20, DGK+20], all existing algorithmic results for Gaussian inputs hold *only for depth two networks* and make at least one of two assumptions on the unknown network F in question:

Assumption (1) Weight matrix $\mathbf{W}_0$ is well-conditioned and, in particular, full rank.

Assumption (2) The vector at the output layer ($\mathbf{W}_1$ when $L = 0$) has all positive entries.

Assumption (1) allows one to use tensor decomposition to recover the parameters of the network and hence PAC learn, an idea that has inspired a long line of works [JSA15, ZSJ+17, GLM17, GKLW18, BJW18]. However, the assumption is not necessary for PAC learning or achieving low-prediction error. For instance, consider a pathological case where $\mathbf{W}_0$ has repeated rows. Here, while parameter recovery is not possible it is still possible to PAC learn. To our knowledge, the only work that can PAC learn depth two networks over Gaussian inputs without a condition number bound on $\mathbf{W}_0$ is [DKKZ20]. However, their work still requires assumption (2) (and only holds for depth two networks). Our work shows that assumption (2) is neither information-theoretically nor computationally necessary.

Limitations of Gradient-Based Methods As discussed in Section 1.2.1, two recent works [GGJ+20, DKKZ20] showed that a broad family of algorithms, namely *correlational statistical query (CSQ) algorithms*, fail to PAC learn even depth two ReLU networks; that is, functions of the form $F(x) = \sum_{i=1}^{k} \lambda_i \phi(\langle v_i, x\rangle)$ with respect to Gaussian inputs in time polynomial in d where d is the ambient dimension (in fact, [DKKZ20] rules out running time $d^{o(k)}$). Informally, a CSQ algorithm is limited to using noisy estimates of statistics of the form $\mathbb{E}[y \cdot \sigma(x)]$ for arbitrary bounded σ, where the expectation is over examples (x, y) and $y = F(x)$ is computed by the network. The point is that this already rules out a wide range of algorithmic approaches in theory and practice, including gradient descent

[4] Other works such as [AZLL19] or kernel-based methods [ZLJ16, GKKT17] require strong norm-based assumptions on the inputs and weights.

on overparameterized networks (i.e., using neural tangent kernels [JGH18] or the mean-field approximation for gradient dynamics [MMN18]). Note that the algorithms of [DKKZ20] for learning depth two ReLU networks with positive coefficients are CSQ algorithms as well.

Note that as a consequence of Theorem 3.1.2, for any ε a function of k, our algorithm can learn the lower bound instances in [GGJ+20, DKKZ20] to error ε in time $g(k) \cdot \mathrm{poly}(d)$ for some g (note that the norm bounds and Lipschitz constants for these instances are upper bounded by functions of k), which is impossible for any CSQ algorithm. We explain why our algorithm is not a CSQ algorithm in Section 3.2.

For the classification version of this problem (i.e., taking a softmax) where we observe $Y \in \{0,1\}$ such that $\mathbb{E}[Y|X] = \sigma(f(X))$ where σ is say sigmoid and $f(X)$ is a depth two ReLU network, Goel et al. [GGJ+20] show that even general SQ algorithms cannot achieve a runtime with polynomial dependence on the dimension. We also remark there is an extensive literature of previous work showing various hardness results for learning certain classes of neural networks [BR89, Vu06, KS09, LSSS14, GKKT17, SVWX17, SSSS17, Sha18, VW19, GKK19, DV20]. We refer the reader to [GGJ+20] for a discussion of how these prior works relate to the above CSQ lower bounds.

3.1.2 Other Related Work and Discussion

Multi-Index Models Functions computed by ReLU networks where $\mathbf{W}_0$ has fewer rows than columns are a special case of a *multi-index model*, that is, a function $F : \mathbb{R}^d \to \mathbb{R}$ given by $F(x) = f(\mathbf{W}^\top x)$ for some matrix $\mathbf{W} \in \mathbb{R}^{k \times d}$ and some function $f : \mathbb{R}^k \to \mathbb{R}$. In the theoretical computer science literature, these are sometimes referred to as *subspace juntas* [VX11, DMN19, DMN20].

One result in this line of work which is close in spirit to the setting we consider is that of [DH18], which gives various conditions on f under which one can recover $\mathbf{W}$ (under Gaussian inputs) in the special case where $k = 1$, as well as a vector in the row span of $\mathbf{W}$ in the case of general k (although these results do not hold for ReLU). In general, the literature on multi-index models is vast, and we refer to [DH18] for a comprehensive overview of this body of work. Many works were inspired by a simple but powerful connection to Stein's lemma [Li92, Bri12, PV16], which was also a key ingredient in the above algorithms for learning neural networks using tensor decomposition.

Another relevant line of work in this literature is the series of results on learning intersections of halfspaces (and indicators of convex sets more generally) over structured input distributions, see e.g. [KLT09, KOS08, BK94, Vem10a, Vem10b]. For Gaussian inputs, when the number of halfspaces (or more generally the dimension of the convex set's hidden subspace) is bounded, it was shown in [Vem10a] that one can essentially read off the row span of $\mathbf{W}$ from the eigendecomposition of $\mathbb{E}[y \cdot (xx^\top - \mathrm{Id})]$ where for a given x, $y = 0$ if x lies in the convex set and $y = 1$ otherwise. By the CSQ lower bounds of [GGJ+20, DKKZ20], such an algorithm provably cannot learn general ReLU networks.

Alternatively, one could also try generalizing the approach of [Vem10a] to our real-valued setting by restricting to level sets S of the ReLU network and forming the matrix $\mathbb{E}[\mathbb{1}[x \in S](xx^\top - \mathrm{Id})]$. We remark however that the analysis in [Vem10a] for such an approach crucially uses convexity of the underlying concept and is therefore not applicable to our setting. Note that this technique is also known as *sliced inverse regression* [BB+18, Li91] in the multi-index model literature, and while it is related to the techniques that we employ, we explain in Remark 2.2.1 why the state of the art here also falls short.

Non-Gaussian Component Analysis As we discuss in Section 3.2, the general approach we take is to find careful reweightings of the distribution over x that will look non-Gaussian in some important direction, i.e., in the row span of $\mathbf{W}_0$. There have been several works on *non-Gaussian component analysis* (see, e.g., [TV18, GS19] and the references therein), but this line of work is not relevant to our result. We also remark that the work [VX11] gives some moment-based conditions under which it is possible to learn multi-index models over Gaussian inputs via non-Gaussian component analysis. However, it seems highly nontrivial to verify whether such conditions hold for ReLU networks, and in addition, their results seem tailored to $\{0, 1\}$-valued functions.

Piecewise-Linear Regression We mention that previous works on *segmented regression* (see, e.g., [ADLS16] on the references therein) study regression for piecewise-linear functions but work with a different notion of piecewise-linearity that is unrelated to our setting.

Non-Homogeneous ReLU Networks We leave as an open question whether our result can be extended to *non-homogeneous* networks of the form $F(x) \triangleq \mathbf{W}_{L+1}\phi(\mathbf{W}_L\phi(\cdots\phi(\mathbf{W}_0x+b_0)+b_1)\cdots+b_L)$, where $b_0,\ldots,b_L \in \mathbb{R}$ are unknown bias parameters. We stress that, over Gaussian inputs, we are not aware of any positive results even for learning non-homogeneous networks of depth two. As for negative results, the recent work of [DV21] rules out polynomial-time algorithms for learning non-homogeneous ReLU networks, even of depth three, assuming local PRGs with polynomial stretch and constant distinguishing advantage exist [App12]. While this hardness result does not preclude the existence of a fixed-parameter tractable algorithm for non-homogeneous ReLU networks, it does give a compelling explanation for the lack of algorithmic progress in the non-homogeneous case.

3.2 Proof Overview

The conceptual novelty of our work is that we go beyond standard CSQ-based algorithms like gradient descent on square loss to give a fundamentally new algorithm for learning neural networks. There are a number of technical novelties to our approach we will describe over the course of outlining our algorithm and analysis in this section.

Suppose we are given samples (x, y) where $y = F(x)$ is computed by a size S ReLU network as in Definition 3.1.1. Let $V \subseteq \mathbb{R}^d$ denote the span of the rows of $\mathbf{W}_0$ and let k be its dimension. We will call V the *relevant subspace*, because the value of F only depends on the projection of x to V. In particular, we can write $y = F'(\Pi_V(x))$ for some function $F' : V \to \mathbb{R}$ that is itself a size S ReLU network and Π_V denotes the projection operator onto V. The main focus of our algorithm will be in figuring out the relevant subspace V given samples (x, y). This is the hardest part of the algorithm, because once we learn the relevant subspace to high enough accuracy, we can grid-search over ReLU networks in this subspace. Even this grid search turns out to be non-trivial to analyze and entails proving new *stability* results for piecewise-linear functions.

Filtered PCA Our algorithm builds upon the *filtered PCA* approach from the previous chapter (we explain in Remark 3.4.14 why a straightforward application of the algorithm there cannot work, necessitating a far more involved approach in the present work). For any $\psi : \mathbb{R} \to \mathbb{R}$, let $\mathbf{M}_\psi \triangleq \mathbb{E}[\psi(Y)(XX^T - \mathrm{Id})]$. As in the previous chapter, we would like to design ψ so that the top principal components of $\mathbf{M}_\psi$ reveal information about the relevant subspace V.

Threshold Filter. Our starting point, as before, is to consider ψ given by a univariate threshold, that is, $\psi(z) = \mathbb{1}[|z| > \tau]$ for suitable τ. For brevity, for $\tau \in \mathbb{R}$ define $\mathbf{M}_\tau = \mathbb{E}_{x,y}[\mathbb{1}[|y| > \tau](xx^T - \mathrm{Id})]$. Then we have that

$$\langle \Pi_V, \mathbf{M}_\tau \rangle = \mathop{\mathbb{E}}_{x,y}\left[\mathbb{1}[|y| > \tau] \cdot (\|\Pi_V x\|^2 - k)\right].$$

In particular, as we discussed in the previous chapter, if one could choose τ for which $|F(x)| > \tau$ only if $\|\Pi_V x\|^2 \geq 2k$ [5], then we would conclude that $\langle \Pi_V, \mathbf{M}_\tau \rangle \geq k \cdot \Pr[|y| > \tau]$, so some singular value of $\mathbf{M}_\tau$ is at least $\Pr[|y| > \tau]$. If F is Λ-Lipschitz, we can simply choose τ to be $\sqrt{2k} \cdot \Lambda$, and provided $\Pr[|y| > \tau]$ is reasonably large, then we conclude that $\mathbf{M}_\tau$ has some reasonably large singular value. Recall that in the previous chapter, our analysis here depended on a compactness argument, but in this chapter we will obtain more quantitative bounds. Namely, to lower bound $\Pr[|y| > \tau]$, we prove an *anti-concentration* result for piecewise linear functions over Gaussian space (Lemma 3.4.4).

Unfortunately, all that the above analysis tells us is that the trace of $\mathbf{M}_\tau$ is non-negligible which in turn helps us guarantee that we identify at least one direction in V. It is not at all clear whether the above threshold approach is enough to identify more than just one vector in the relevant subspace. Indeed, recovering the full relevant subspace turns out to be significantly more challenging, and the core technical contribution of the work in this chapter is to show how to do this.

Learning the Full Subspace: What Doesn't Work One might hope that a more refined analysis shows that for a suitable τ, the spectrum of $\mathbf{M}_\tau$ can identify the entire subspace V. Given that we can already learn some $w \in V$ with the threshold approach above, a first step would be to try to find a direction in V orthogonal to w, by lower bounding

[5] The choice of $2k$ here is for exposition; any bound noticeably more than k, e.g., $k+1$ will do.

the contribution to the Frobenius norm of $\mathbf{M}_\tau$ from vectors orthogonal to w. Concretely, letting $\Pi_{V\backslash\{w\}}$ denote the projector to the orthogonal complement of w in V, we have that

$$\langle \Pi_{V\backslash\{w\}}, \mathbf{M}_\tau \rangle = \mathop{\mathbb{E}}_{x,y}\left[\mathbb{1}[|y| > \tau] \cdot (\|\Pi_{V\backslash\{w\}}x\|^2 - (k-1))\right]. \tag{3.1}$$

As before, if one could choose τ for which $|F(x)| > \tau$ only if $\|\Pi_{V\backslash\{w\}}x\|^2 \geq k$, and if we could lower bound $\Pr[|y| > \tau]$, then we would conclude that $\langle \Pi_{V\backslash\{w\}}, \mathbf{M}_\tau \rangle \geq \Pr[|y| > \tau]$, so $\mathbf{M}_\tau$ has some other singular vector, orthogonal to w, with non-negligible singular value. The issue is that such a τ typically does not exist! For x satisfying $\|\Pi_{V\backslash\{w\}}x\|^2 \leq k$, $F(x)$ can be arbitrarily large, because $\|\Pi_w x\|$ can be arbitrarily large.

It may be possible to lower bound the quantity in (3.1) using a more refined argument, but for general deep ReLU networks or piecewise linear functions, this seems very challenging. At the very least, one must be careful not to prove something too strong, like showing that $v^\top \mathbf{M}_\tau v$ is non-negligible for *any* unit vector $v \in V$. For instance, even when $L = 0$, it could be that all but one of the rows of $\mathbf{W}_0$ lie in a proper subspace $W \subsetneq V$, and for the remaining row u of $\mathbf{W}_0$, $\|\Pi_{V\backslash W}u\|/\|u\|$ is arbitrarily small. In this case, for v in the direction of $\Pi_{V\backslash W}u$, the quadratic form $v^\top \mathbf{M}_\tau v$ is arbitrarily small, and it would be impossible to recover all of V from a reasonable number of samples.

More generally, any proposed algorithm for learning all of V had better be consistent with the fact that it is impossible to recover the full subspace V within a reasonable number of samples if almost all of the variance of F is explained by some proper subspace $W \subsetneq V$, or equivalently, if the "leftover variance" $\mathbb{E}_x[(F(x) - F(\Pi_W x))^2]$ is negligible. We emphasize that this is a key subtlety that does not manifest in previous works that consider full-rank, well-conditioned weight matrices.

Learning the Full Subspace: Our Approach We now explain our approach. At a high level, we try to learn orthogonal directions inside the relevant subspace in an iterative fashion. The threshold filter approach above already gives us a single direction in V. Suppose inductively that we've learned some orthogonal vectors $w_1, ..., w_\ell \in V$ spanning a subspace $W \subseteq V$ and want to learn another (note that technically we can only guarantee $w_1, ..., w_\ell$

are approximately within V, but let us temporarily ignore this for the sake of exposition). Motivated by the above consideration regarding "leftover variance," we proceed by a win-win argument: either the leftover variance already satisfies $\mathbb{E}_x[(F(x) - F(\Pi_W x))^2] \le \varepsilon$ in which case we are already done, or we can learn a new direction via the following crucial modification of the threshold filter.

First, as a thought experiment, consider the following matrix

$$\mathbf{M}_\tau^W \triangleq \Pi_{W^\perp} \mathop{\mathbb{E}}_{x,y}\big[\mathbb{1}[|y - F(\Pi_W x)| > \tau] \cdot (xx^\top - \mathrm{Id})\big]\Pi_{W^\perp}.$$

Note the critical fact that we threshold on $y - F(\Pi_W x)$ as opposed to just on y. As before, it is not hard to show that if this matrix is nonzero, then its singular vectors with nonzero singular value must lie in $\mathbf{W}_0$ and be orthogonal to W; thus giving us a new direction in $\mathbf{W}_0$. We claim that if the leftover variance is non-negligible, then the above matrix will give us a new direction in W.

The intuition behind the above matrix is as follows. Let $V \backslash W$ denote the subspace of V orthogonal to W. We can write $F(x) = F(\Pi_V x) = F(\Pi_W x + \Pi_{V\backslash W} x)$. Now, as F is Lipschitz, we can bound $G(x) = y - F(\Pi_W x) = F(\Pi_W x + \Pi_{V\backslash W} x) - F(\Pi_W x)$ as $|G(x)| \le \Lambda \|\Pi_{V\backslash W} x\|^2$, where Λ is the Lipschitz constant of F. In other words, $G(x)$ is bounded over x for which $\|\Pi_{V\backslash W} x\|$ is bounded. Recall that the fact that $F(x)$ is not bounded over such x was the key obstacle to using the original threshold filter approach to learn the full subspace.

The upshot is that for a suitably large τ, the only contribution to the matrix $\mathbf{M}_\tau^W$ should be from inputs x that have large projection in $V \setminus W$. We are now in a position to adapt the analysis lower bounding $\langle \Pi_V, \mathbf{M}_\tau \rangle$ to lower bounding $\langle \Pi_{V\backslash W}, \mathbf{M}_\tau^W \rangle$. In particular, we can apply the aforementioned anti-concentration for piecewise linear functions *to the function* G and argue that, provided the leftover variance $\mathbb{E}_x[(F(x) - F(\Pi_W x))^2] = \mathbb{E}_x[G(x)^2]$ is non-negligible, the top singular vector of $\mathbf{M}_\tau^W$ will give us a new vector in $V \setminus W$.

That being said, an obvious obstacle in implementing the above is that along with not knowing the true subspace $\mathbf{W}_0$, we also don't know the true function F. This precludes us from forming the matrix $\mathbf{M}_\tau^W$ as defined above.

To get around this, we will enumerate over a sufficiently fine net of ReLU networks $\widetilde{F}$

with relevant subspace W, one of which will be close to the ReLU network $F(\Pi_W x)$. For each $\widetilde{F}$, we will form the matrix

$$\widetilde{\mathbf{M}}^W_\tau \triangleq \Pi_{W^\perp} \mathop{\mathbb{E}}_{x,y}\Big[\mathbb{1}[|y - \widetilde{F}(\Pi_W x)| > \tau] \cdot (xx^\top - \mathrm{Id})\Big]\Pi_{W^\perp}. \tag{3.2}$$

and output the top singular vector as our new direction only if it has non-negligible singular value.

Arguing soundness, i.e. that this procedure doesn't yield a "false positive" in the form of an erroneous direction lying far from V, is not too hard. However, analyzing completeness, i.e. that this procedure will find *some* new direction, is surprisingly subtle (see Lemma 3.4.12). Formally, we need to argue that if we have an approximation $\widetilde{F}$ to the true F (under some suitable metric), then the corresponding matrix $\widetilde{\mathbf{M}}^W_\tau$ is close to the matrix $\mathbf{M}^W_\tau$. This is further complicated by the fact that ultimately, we will only have access to a subspace W which is *approximately* in V, as every direction we find in our iterative procedure is only guaranteed to *mostly* lie within V.

Our key step in proving this is showing a new stability property of affine thresholds of piecewise linear functions and makes an intriguing connection to *lattice polynomials* in tropical geometry.

Stability of Piecewise Linear Functions Following the above discussions, to complete our analysis we need to show *stability* of affine thresholds of ReLU networks in the following sense: if $F, \tilde{F} : \mathbb{R}^d \to \mathbb{R}$ are two RELU networks that are close in some structural sense (i.e., under some parametrization), then $\mathbb{E}[\mathbb{1}[|F(x)| > \tau](xx^T - Id)] \approx \mathbb{E}[\mathbb{1}[|\tilde{F}(x)| > \tau](xx^T - Id)]$. A natural way to approach the above is to upper bound $\Pr[|F(x)| > \tau \wedge |\widetilde{F}(x)| \leq \tau]$. That is, affine thresholds of ReLU networks that are structurally close disagree with low probability.

A natural way to parametrize closeness is to require the weight matrices of the two networks $F, \tilde{F}$ to be close to each other. While such a statement is not too difficult to show for depth two networks (by a union bound over pairs of ReLUs), proving such a statement for general ReLU networks using a direct approach seems quite challenging. We instead look at proving such a statement for a more general class of functions - continuous piecewise-linear functions which allows us to do a certain kind of hybrid argument more naturally.

Concretely, we show that affine thresholds of piecewise-linear functions that are close in some appropriate structural sense disagree with low probability over Gaussian space. We will elaborate upon the notion of structural closeness we consider momentarily, but for now it is helpful to keep in mind that it specializes to L_2 distance for linear functions.

Lemma 3.2.1 (Informal, see Lemma 3.4.6). *Let $F, \widetilde{F} : \mathbb{R}^d \to \mathbb{R}$ be piecewise-linear functions, both consisting of at most m linear pieces, which are "(m, η)-structurally-close" (see Definition 3.3.12). For any $\tau > 0$,*

$$\Pr_{x \sim \mathcal{N}(0, Id)} \left[|F(x)| > \tau \wedge |\widetilde{F}(x)| \le \tau \right] \le O(\eta m^2 / \tau). \tag{3.3}$$

To get a sense for this, suppose $F, \widetilde{F}$ were even close in the sense that the polyhedral regions over which F is linear are *identical* to those over which $\widetilde{F}$ is linear, and furthermore $\mathbb{E}_x[(F(x) - \widetilde{F}(x))^2]^{1/2} \le \eta$. Then if we take for granted that Lemma 3.2.1 holds when $m = 1$, i.e. when $F, \widetilde{F}$ are linear (see Lemma 1.3.35), it is not hard to show an $O((\eta m/\tau)^c)$ upper bound in (3.3) under this very strong notion of closeness for some $c < 1$. Because F and $\widetilde{F}$ are L_2-close as functions, for any $t > 0$ we have that with probability $1 - O(\eta^2/t^2)$ the input $x \sim \mathcal{N}(0, \mathrm{Id})$ lies in a polyhedral region for which the corresponding linear functions for F and $\widetilde{F}$ are t-close. By the $m = 1$ case of Lemma 3.2.1, over any one of these at most m regions, the affine thresholds $\mathbb{1}[|F(x)| > \tau]$ and $\mathbb{1}[|F(x)| > \tau]$ disagree with probability $O(t/\tau)$. Union bounding over these regions as well as the event of probability η^2/t^2 that x does not fall in such a polyhedral region, we can upper-bound the left-hand side of (3.3) by $O(\eta^2/t^2 + mt/\tau)$, and by taking $t = (\eta^2\tau/m)^{1/3}$, we get a bound of $(\eta m^2/\tau)^{2/3}$.

The issues with this are twofold. First, recall the function $\widetilde{F}$ that we want to apply Lemma 3.4.6 to is obtained from some enumeration over a fine net of ReLU networks. As such there is no way to guarantee that the polyhedral regions defining F and $\widetilde{F}$ are exactly the same, making adapting the above argument far more difficult, especially for general ReLU networks.

Second, we stress that the *linear* scaling in $O(\eta)$ in (3.2.1) is essential. If one suffered any polynomial loss in this bound as in the above argument, then upon applying Lemma 3.2.1 k

times over the course of our iterative algorithm for recovering V, we would incur time and sample complexity *doubly exponential* in k. The reason is as follows.

Recall that in the final argument we can only ensure that the directions $w_1, \ldots, w_\ell$ we have found so far are *approximately* within V, and the parameter η will end up scaling with an appropriate notion of subspace distance between W and the true space V. On the other hand, the bound we can show on how far $\widetilde{M}^W_\tau$ deviates from M^W_τ in spectral norm will essentially scale with the right-hand side of (3.2.1). So if we could only ensure $\widetilde{M}^W_\tau$ and M^W_τ are $O(\eta^c)$-close in spectral norm for $c < 1$, then if we append the top eigenvector of $\widetilde{M}^W_\tau$ to the list of directions $w_1, ..., w_\ell$ we have found so far, the resulting span will only be $O(\eta^c)$-close in subspace distance. Iterating, we would conclude that for the final output of the algorithm to be sufficiently accurate, we would need the error incurred by the very first direction w_1 found to be doubly exponentially small in k!

Lattice Polynomials It turns out that there is a clean workaround to both issues: passing to the *lattice polynomial* representation for piecewise-linear functions. Specifically, we exploit the following powerful tool:

Theorem 3.2.2 ([Ovc02], Theorem 4.1; see Theorem 3.3.11 below)**.** *If F is continuous piecewise-linear, there exist linear functions $\{g_i\}_{i\in[M]}$ and subsets $\mathcal{I}_1, ..., \mathcal{I}_m \subseteq [M]$ for which*

$$F(x) = \max_{j\in[m]} \min_{i\in\mathcal{I}_j} g_i(x). \tag{3.4}$$

In fact, our notion of "structural closeness" will be built around this structural result. Roughly speaking, we say two piecewise linear functions are structurally close if they have lattice polynomial representations of the form (3.4) with the same set of clauses and whose corresponding linear functions are pairwise close in L_2 (see Definition 3.3.12).

At a high level, Theorem 3.2.2 will then allow us to implement a hybrid argument in the proof of Lemma 3.2.1 and carefully track how the affine threshold computed by a piecewise-linear function changes as we interpolate between F and $\widetilde{F}$. In this way, we end up with the desired linear dependence on η in (3.2.1).

With Lemma 3.2.1 in hand, we can argue that even with only access to a subspace W

approximately within V and with only a function $\widetilde{F}$ that approximates $F(\Pi_W x)$, the top singular vector of (3.2) mostly lies within V, and we can make progress.

Finally, we remark that as an added bonus, Theorem 3.2.2 also gives us a way to enumerate over general continuous piecewise-linear functions! In this way, we can adapt our algorithm for learning ReLU networks to learning arbitrary piecewise-linear functions, with some additional computational overhead (see Theorem 3.4.1).

Enumerating Over Piecewise-Linear Functions and ReLU Networks There is in fact one more subtlety to implementing the above approach for ReLU networks and getting singly exponential dependence on k.

First note that whereas one can always enumerate over functions computed by lattice polynomials of the form (3.4) in time $\exp(\text{poly}(M))$ (see Lemma 3.3.16), for ReLU networks of size S this can be as large as doubly exponential in S. Instead, we enumerate over ReLU networks in the naive way, that is, enumerating over the $\exp(O(S))$ many possible architectures and netting over weight matrices with respect to spectral norm, giving us only singly exponential dependence on S.

Here is the subtlety. Obviously two ReLU networks with the same architecture and whose weight matrices are pairwise close in spectral norm will be close in L_2. But how do we ensure that the corresponding lattice polynomials guaranteed by Theorem 3.2.2 are structurally close? In particular, getting anything quantitative would be a nightmare if the clause structure of these lattice polynomials depended in some sophisticated, possibly discontinuous fashion on the precise entries of the weight matrices.

Our workaround is to open up the black box of Theorem 3.2.2 and give a proof for the special case of ReLU networks from scratch. In doing so, we will find out that there are lattice polynomial representations for ReLU networks which only depend on the architecture and the *signs* of the entries of the weight matrices (see Theorem 3.3.17). In this way, we can guarantee that a moderately fine net will contain a network which is structurally close to the true network.

3.3 Technical Preliminaries

In this section we collect some tools that will be useful specifically for this chapter.

3.3.1 Miscellaneous Tools

Clipping For $\eta > 0$, let $\mathsf{clip}_\eta : \mathbb{R} \to \mathbb{R}$ denote the function given by

$$\mathsf{clip}_\eta(z) = \begin{cases} z & \text{if } |z| \le \eta \\ 0 & \text{otherwise} \end{cases}$$

Overloading notation, given a vector $v \in \mathbb{R}^m$, we will use $\mathsf{clip}_\eta(v)$ to refer to the vector in $\mathbb{R}^m$ obtained by applying clip_η entrywise.

We will use the following basic property of the clipping operation:

Fact 3.3.1. *Suppose $v, v' \in \mathbb{R}^m$ satisfy $\|v - v'\|_\infty \le \eta$, and define $v'' \triangleq \mathsf{clip}_\eta(v')$. Then for any $i \in [m]$, $v_i v''_i \ge 0$.*

Proof. If $v''_i > 0$, then $v''_i = v'_i > \eta$ and by triangle inequality, $v_i > 0$. Similarly, if $v''_i < 0$, then $v''_i = v'_i < -\eta$ and by triangle inequality, $v_i < 0$. □

Lattice Polynomials Recall our notation for max/min (see the beginning of Section 1.3. The following class of functions will be useful for us.

Definition 3.3.2. *The set of* lattice polynomials over the reals *is the set of real-valued functions defined inductively as follows: for any $d \ge 1$, any constant real-valued function $\mathbb{R}^d \to \mathbb{R}$ is a lattice polynomial, and any function $h : \mathbb{R}^d \to \mathbb{R}$ which can be written as $h(x) = f(x) \vee g(x)$ or $h(x) = f(x) \wedge g(x)$ for two lattice polynomials $f, g : \mathbb{R}^d \to \mathbb{R}$ is also a lattice polynomial.*

3.3.2 Continuous Piecewise-Linear Functions and Lattice Polynomials

In this section, we introduce tools for reasoning about continuous piecewise-linear functions, culminating in a structural result (Theorem 3.3.17) giving an explicit representation of arbitrary ReLU networks as lattice polynomials (see Definition 3.3.2).

Basic Notions

We will work with functions which only depend on some low-dimensional projection of the input.

Definition 3.3.3 (Subspace juntas)**.** *A function $F : \mathbb{R}^d \to \mathbb{R}$ is a* subspace junta *if there exist $v_1, ..., v_k \in \mathbb{S}^{d-1}$ and a function $h : \mathbb{R}^k \to \mathbb{R}$ for which $F(x) = h(\langle v_1, x\rangle, ..., \langle v_k, x\rangle)$ for all $x \in \mathbb{R}^d$. We will refer to $V \triangleq \mathrm{span}(v_1, ..., v_k)$ as the* relevant subspace *of F, to $v_1, ..., v_k$ as the* relevant directions *of F, and to h as the* link function *of F.*

Definition 3.3.4 (Piecewise Linear Functions)**.** *Given vector space W, a function $h : W \to \mathbb{R}$ is said to be* piecewise-linear *(resp. piecewise-affine-linear) if there exist finitely many linear (resp. affine linear) functions $\{g_i : W \to \mathbb{R}\}_{i\in[M]}$ and a partition of W into finitely many polyhedral cones $\{S_i\}_{i\in\mathcal{I}}$ such that $G(x) = \sum_i \mathbb{1}[x \in S_i]g_i(x)$. We will say that h is* realized by M pieces $\{(g_i, S_i)\}$ *(note that h can have infinitely many realizations). If each g_i is given by $g_i(x) = \langle u_i, x\rangle + b_i$ for some $u_i \in W, b_i \in \mathbb{R}$, then we will also refer to the pieces of h by $\{(\langle u_i, \cdot\rangle + b_i, S_i)\}$.*

We are now ready to define the concept class we will work with in this chapter.

Definition 3.3.5 ("Kickers")**.** *We call a subspace junta F with link function h a* kicker *if h is continuous piecewise-linear. Note that a kicker is itself a continuous piecewise-linear function, and for any realization of its link function by M pieces, there is a realization of F by M pieces.*

Henceforth, fix a subspace junta $F : \mathbb{R}^d \to \mathbb{R}$ with link function h and relevant directions $v_1, ..., v_k$ spanning relevant subspace $V \subset \mathbb{R}^d$.

Example 3.3.6 (ReLU Networks)**.** *Feedforward ReLU networks as defined in Definition 3.1.1* *are kickers with relevant subspace of dimension at most k, where k is the row span of the weight matrix $\mathbf{W}_0$, the link function is defined by*

$$h(z) = \mathbf{W}_{L+1}\phi(\mathbf{W}_L\phi(\cdots \mathbf{W}_1\phi(z)\cdots)),$$

and the pieces in one possible realization of h correspond to the different possible sign patterns that the activations could take on, that is the different possible values of the vector

$$\{\mathbf{W}_a\phi(\mathbf{W}_{a-1}\phi(\cdots \mathbf{W}_1\phi(z)\cdots))\}_{0\le a\le L} \in \prod_{a=0}^{L}\{\pm 1\}^{k_a}$$

as z ranges over $\mathbb{R}^k$.

Lemma 3.3.7. *If F is a Λ-Lipschitz kicker, then for any realization of its link function h by pieces $\{(\langle w_i, \cdot\rangle, S_i)\}$, there is a realization by pieces $\{(\langle w'_i, \cdot\rangle, S_i)\}$ for which* $\max_i \|g_i\| \le L$*..*

Proof. Consider any piece $(\langle w_i, \cdot\rangle, S_i)$. If there is some $x \in S_i$ for which there exists a ball of nonzero radius r around x contained in S_i, then clearly $L \ge \|w_i\|$: take x and $x + r \cdot w_i/\|w_i\|$ and note that

$$L \ge \frac{F(x + r \cdot w_i/\|w_i\|) - F(x)}{\|(x + r \cdot w_i/\|w_i\|) - x\|} = \frac{r\|w_i\|}{r} = \|w_i\|.$$

If no such x and ball exist, then S_i is not full-dimensional and therefore contained in a hyperplane $W \subset V$. Then if we replace $(\langle w_i, \cdot\rangle, S_i)$ in the realization of h with $(\langle \Pi_W w_i, \cdot\rangle, S_i)$, this is still a realization of h. Again, it would suffice for there to exist a ball, now in the subspace W, of nonzero radius around some point in S_i. If this is not the case, then S_i is not a full-dimensional subset of W and thus lies in a codimension 1 subspace of W. Continuing thus, we eventually obtain some (possibly zero) vector w'_i for which replacing $(\langle w_i, \cdot\rangle, S_i)$ in the realization of h with $(\langle w'_i, \cdot\rangle, S_i)$ still gives a realization of h, and furthermore $\|w'_i\| \le L$. □

Definition 3.3.8 (Restrictions)**.** *Given any nonzero linear subspace $W \subseteq V$, let $F|_W : W \to \mathbb{R}$ denote the restriction of F to the subspace W. By abuse of notation, we will sometimes also regard $F|_W$ as a function over $\mathbb{R}^d$ given by $F|_W(x) = F(\Pi_W x)$.*

One of the main properties of kickers that we exploit is *positive homogeneity*:

Fact 3.3.9 (Positive homogeneity)**.** *For any $\lambda \geq 0$ and $x \in \mathbb{R}^k$, $F(\lambda \cdot x) = \lambda F(x)$.*

The following property of restrictions of Lipschitz functions will be important.

Lemma 3.3.10. *For any nonzero linear subspace $W \subseteq V$, and Λ-Lipschitz function $F : \mathbb{R}^d \to \mathbb{R}$,*

$$\sup_{x:\|\Pi_{V\backslash W}x\|\leq 1} |F(x) - F(\Pi_W x)| \leq \Lambda.$$

Proof. Because $F(x) = F(\Pi_V x)$ and $F(\Pi_W x) = F(\Pi_W \Pi_V x)$, we may assume without loss of generality that $x \in V$. For any $x \in V$ for which $\|\Pi_{V\backslash W}x\| \leq 1$, we have that

$$|F(x) - F(\Pi_W x)| \leq \Lambda\|x - \Pi_W x\| = \Lambda\|\Pi_{V\backslash W}x\| \leq \Lambda,$$

as claimed. □

A Generic Lattice Polynomial Representation

Essential to our analysis is the following structural result from [Ovc02] which says that, perhaps surprisingly, *any* piecewise linear function can be expressed as a relatively simple lattice polynomial.

Theorem 3.3.11 ([Ovc02], Theorem 4.1)**.** *If $h : \mathbb{R}^n \to \mathbb{R}$ is a continuous piecewise-linear function which has a realization by pieces $\{(g_i, S_i)\}_{i\in[M]}$, there exists a collection of* clauses *$\mathcal{I}_1, ..., \mathcal{I}_m \subseteq [M]$ for which*

$$h(x) = \max_{j\in[m]} \min_{i\in\mathcal{I}_j} g_i(x) \tag{3.5}$$

We will work with the following notion of approximation for such lattice polynomials:

Definition 3.3.12. *Two continuous piecewise-linear functions $G, \widetilde{G} : \mathbb{R}^d \to \mathbb{R}$ are (M, η)-*structurally-close *if there exist linear functions $g_1, ..., g_M$ and $\widetilde{g}_1, ..., \widetilde{g}_M$ and subsets $\mathcal{I}_1, ..., \mathcal{I}_m \subseteq [M]$ for which*

$$G(x) = \max_{j\in[m]} \min_{i\in\mathcal{I}_i} g_i(x) \qquad \widetilde{G}(x) = \max_{j\in[m]} \min_{i\in\mathcal{I}_i} \widetilde{g}_i(x)$$

and $\|g_i - \widetilde{g}_i\| \leq \eta$ for all i.

Structural closeness of continuous piecewise-linear functions in the above sense is stronger than L_2-closeness.

Lemma 3.3.13. *Take continuous piecewise-linear functions $G, \widetilde{G} : \mathbb{R}^m \to \mathbb{R}$ which are (M,η)-structurally-close. Then $\|G - \widetilde{G}\| \le \eta\sqrt{m}$. In particular, if G is a piecewise-linear function which is realized by pieces $\{(\langle u_i, \cdot\rangle, S_i)\}$ satisfying $\|u_i\| \le \eta$, then $\|G\| \le \eta\sqrt{m}$.*

To show this, we need the following helper lemma:

Lemma 3.3.14. *If $\{g_i\}_{i\in[M]}$ and $\{\widetilde{g}_i\}_{i\in[M]}$ are two collections of linear functions, then for any x,*

$$|\max_{j\in[m]}\min_{i\in\mathcal{I}_j} g_i(x) - \max_{j\in[m]}\min_{i\in\mathcal{I}_j} \widetilde{g}_i(x)| \le \max_i |g_i(x) - \widetilde{g}_i(x)|$$

Proof. This simply follows by induction using the fact that if $f_1, f_2 : \mathbb{R}^a \to \mathbb{R}$ are both 1-Lipschitz with respect to L_∞, then $f_1 \vee f_2$ and $f_1 \wedge f_2$ are as well. □

Proof of Lemma 3.3.13. Let $\{(\langle u_i, \cdot\rangle, S_i)\}_{i\in[M]}$ and $\{(\langle \widetilde{u}_i, \cdot\rangle, S_i)\}_{i\in[M]}$ be the realizations of $G, \widetilde{G}$ for which $\|u_i - \widetilde{u}_i\| \le \eta$. By Lemma 3.3.14 applied to these pieces, together with Cauchy-Schwarz, for any x we have that $|G(x) - \widetilde{G}(x)| \le \eta\|x\|$. So $\|G - \widetilde{G}\| \le \eta \cdot \mathbb{E}[\|x\|^2]^{1/2} = \eta\sqrt{m}$. □

As discussed in Section 3.2, for our application to learning general kickers, we will leverage the lattice polynomial representation in Theorem 3.3.11 to grid over piecewise-linear functions. Note that *a priori*, even if we knew exactly the set of linear functions $\{g_i\}_{i\in[M]}$ in a realization of a piecewise-linear function, enumerating over all lattice polynomials of the form (3.5) would require time doubly exponential in M, as there are 2^M possible clauses $\mathcal{I}_j$ and 2^{2^M} possible sets of clauses $\{\mathcal{I}_j\}$.

By being slightly more careful, we can enumerate over piecewise linear functions in time $\exp(\mathrm{poly}(M))$.

Definition 3.3.15. *An* order type on n elements *is specified by a function $\omega : [n] \to [n]$ for which every element from 1 to $\max_i \omega(i)$ is present. We say that a set of n real numbers $z_1, ..., z_n$ has order type ω (denoted $\{z_1, ..., z_n\} \vdash \omega$ if $z_i = z_j$ (resp. $z_i > z_j$, $z_i < z_j$) if and only if $\omega(i) = \omega(j)$ (resp. $\omega(i) > \omega(j)$, $\omega(i) < \omega(j)$). Denote the set of order types on n elements by Ω_n. Note that any set of real numbers has exactly one order type.*

Lemma 3.3.16. *If F has a realization by pieces $\{(g_i, S_i)\}_{i\in[M]}$, then there is a function*

$A : \Omega_M \to [M]$ such that for any x,

$$F(x) = \sum_{\omega \in \Omega_M} \mathbb{1}\left[\{g_i(x)\}_{i\in[M]} \vdash \omega\right] \cdot g_{A(\omega)}(x).$$

Proof. Let $F(x) = \max_{j\in[m]} \min_{i\in\mathcal{I}_j} g_i(x)$ be the max-min representation guaranteed by Theorem 3.3.11. This representation implies that for a fixed order type ω, there is some index $i \in [M]$ for which $F(x) = g_i(x)$ for all x satisfying $\{g_i(x)\}_{i\in[M]} \vdash \omega$. This gives the desired mapping A. □

Note that the set of functions $A : \Omega_M \to [M]$ is only of size $(M!)^M \le M^{M^2}$, so by Lemma 3.3.16, to enumerate over piecewise-linear functions with M pieces we can simply enumerate over linear functions $\{g_i\}$ together with all possible functions A (see Algorithm 8 below).

Lattice Polynomials for ReLU Networks

Here we give an explicit proof of Theorem 3.3.11 in the special case of ReLU networks. We emphasize that the specific nature of the construction exhibited in this theorem will be important in the proof of our main result for learning ReLU networks, and that simply applying Theorem 3.3.11 in a black-box fashion will not suffice for our purposes.

Theorem 3.3.17. *If $F \in \mathcal{C}_S$ is a ReLU network with weight matrices $\mathbf{W}_0 \in \mathbb{R}^{k_0\times d}, \mathbf{W}_1 \in \mathbb{R}^{k_1\times k_0}, \dots, \mathbf{W}_L \in \mathbb{R}^{k_L\times k_{L-1}}, \mathbf{W}_{L+1} \in \mathbb{R}^{1\times k_L}$, and if F' is a ReLU network with the same architecture as F, with weight matrices $\mathbf{W}'_0, ..., \mathbf{W}'_{L+1}$, such that*

$$(\mathbf{W}_a)_{i,j} \cdot (\mathbf{W}'_a)_{i,j} \ge 0 \qquad \forall\ 0 \le a \le L+1, (i,j) \in [k_a] \times [k_{a-1}],$$

then there exist vectors $v_1, ..., v_M, v'_1, ..., v'_M$ and clauses $\mathcal{I}_1, ..., \mathcal{I}_m \subseteq [M]$, where $M = 2^S$, for which

$$F(x) = \max_{j\in[m]} \min_{i\in\mathcal{I}_j} \langle v_i, x\rangle$$

$$F'(x) = \max_{j\in[m]} \min_{i\in\mathcal{I}_j} \langle v'_i, x\rangle.$$

Specifically, $v_1, ..., v_M$ consist of all vectors of the form $\mathbf{W}_{L+1}\mathbf{\Sigma}_L\mathbf{W}_L\mathbf{\Sigma}_{L-1} \cdot \cdots \cdot \mathbf{\Sigma}_0\mathbf{W}_0$ for diagonal matrices $\mathbf{\Sigma}_i \in \{0,1\}^{k_i \times k_i}$, and $v'_1, ..., v'_M$ are defined analogously.

We prove Theorem 3.3.17 by induction by exhibiting max-min representations for ReLUs, scalings, and sums of max-min formulas. Let $G : \mathbb{R}^d \to \mathbb{R}$ be a piecewise-linear function given by $G(x) \triangleq \max_{j\in[m]} \min_{i\in\mathcal{I}_j} \langle u_i, x\rangle$ for some subsets $\{\mathcal{I}_1, ..., \mathcal{I}_m\}$ of $[M]$ and vectors $\{u_1, ..., u_M\}$ in $\mathbb{R}^d$.

Lemma 3.3.18. *Let $u_{M+1} = 0$ and let $\mathcal{I}_{m+1} = \{M+1\}$. Then for all $x \in \mathbb{R}^d$,*

$$\phi(G(x)) = \max_{j\in[m+1]} \min_{i\in\mathcal{I}_j} \langle u_i, x\rangle.$$

Proof. This is immediate from the definition of ϕ. □

Lemma 3.3.19. *For any $\lambda \in \mathbb{R}$, there exist subsets $\{\mathcal{J}_1, ..., \mathcal{J}_{m'}\}$ of $[M]$ such that for all $x \in \mathbb{R}^d$,*

$$\lambda G(x) = \max_{j\in[m']} \min_{i\in\mathcal{J}_j} \langle \lambda u_i, x\rangle.$$

Furthermore, these subsets only depend on $\mathcal{I}_1, ..., \mathcal{I}_m$ and the sign of λ.

Proof. For $\lambda > 0$, we have $\mathcal{J}_j = \mathcal{I}_j$ for all j. So it remains to show the claim for $\lambda = -1$. We can write $-G(x)$ as $\min_{j\in[m]} \max_{i\in\mathcal{I}_j} \langle u_i, x\rangle$. This is a lattice polynomial over the reals, and any lattice polynomial over a distributive lattice can be written in disjunctive normal form as $\max_{j\in[m']} \min_{i\in\mathcal{J}_j} \langle u_i, x\rangle$ for some subsets $\{\mathcal{J}_j\}$ (see e.g. [Bir40, Section II.5, Lemma 3]), from which the claim follows. □

Lemma 3.3.20. *For any $k' \in \mathbb{N}$ and $b \in [k']$, let $G_b(x) = \max_{j\in[m_b]} \min_{i\in\mathcal{I}^b_j} \langle u^b_i, x\rangle$ for some subsets $\{\mathcal{I}^b_j\}$ of $[M_b]$ and vectors $\{u^b_i\}$ in $\mathbb{R}^d$. For all $x \in \mathbb{R}^d$,*

$$\sum_{b=1}^{k'} G_b(x) = \max_{(j_1,...,j_{k'})\in[m_1]\times\cdots\times[m_{k'}]} \ \min_{(i_1,...,i_{k'})\in\mathcal{I}_{j_1}\times\cdots\times\mathcal{I}_{j_{k'}}} \langle u^1_{i_1} + \cdots + u^{k'}_{i_{k'}}, x\rangle. \tag{3.6}$$

Proof. Take any $x \in \mathbb{R}^d$, and for $b \in [k']$ suppose that $G_b(x) = \langle u^b_{i^*_b}, x\rangle$ for some index

$i^*_b \in [M]$. Note that for any $\mathcal{I}^1_{j_1}, \ldots, \mathcal{I}^{k'}_{j_{k'}}$ containing $i^*_1, \ldots, i^*_{k'}$ respectively,

$$\min_{(i_1,\ldots,i_{k'}) \in \mathcal{I}^1_{j_1} \times \cdots \times \mathcal{I}^{k'}_{j_{k'}}} \langle u^1_{i_1} + \cdots + u^{k'}_{i_{k'}}, x\rangle = \langle u^{k'}_{i^*_{k'}} + \cdots + u^{k'}_{i^*_{k'}}, x\rangle.$$

This shows that the right-hand side of (3.6) is lower bounded by the left-hand side.

We now show the other direction. For any $i'_1, \ldots, i'_{k'}$ for which $\langle u^1_{i'_1} + \ldots + u^{k'}_{i'_{k'}}, x\rangle > G_1(x) + \cdots + G_{k'}(x)$, we must have $\langle u^b_{i'_b}, x\rangle > G_b(x)$ for some $b \in [k']$. In this case, we know that for every clause $\mathcal{I}^b_{j_b}$ in G_b which contains i'_b, there is some $i \in \mathcal{I}^b_{j_b}$ for which $\langle u^b_i, x\rangle < \langle u^b_{i'_b}, x\rangle$. So for any $\mathcal{I}^1_{j_1}, \ldots, \mathcal{I}^{k'}_{j_{k'}}$ containing $i'_1, \ldots, i'_{k'}$ respectively, the corresponding clause on the right-hand side of (3.6) satisfies $\min_{(i_1,\ldots,i_{L'}) \in \mathcal{I}^1_{j_1} \times \cdots \times \mathcal{I}^{L'}_{j_{L'}}} \langle u^1_{i_1} + \cdots + u^{L'}_{i_{L'}}, x\rangle < \langle u^1_{i'_1} + \cdots + u^{L'}_{i'_{L'}}, x\rangle$. This concludes the proof that the left-hand side of (3.6) is upper bounded by the left-hand side. □

We can now prove Theorem 3.3.17:

Proof. The claim is trivially true for $L = -1$. Suppose inductively that for some layer $0 \le a \le L$, we have that for all $b \in [k_a]$, if we denote

$$F_{a,b} \triangleq \mathbf{W}^b_a \phi\left(\mathbf{W}_{a-1}\phi\left(\cdots\phi\left(\mathbf{W}_0 x\right)\right)\right)$$
$$F'_{a,b} \triangleq \mathbf{W}'^b_a \phi\left(\mathbf{W}'_{a-1}\phi\left(\cdots\phi\left(\mathbf{W}'_0 x\right)\right)\right),$$

where $\mathbf{W}^b_a$ denotes the b-th row of $\mathbf{W}_a$, then $F_{a,b}$ and $F'_{a,b}$ can be expressed as max-min formulas $\max_{j\in[m_{a,b}]}\min_{i\in\mathcal{I}^{a,b}_j}\langle v^{a,b}_i, \cdot\rangle$ and $\max_{j\in[m_{a,b}]}\min_{i\in\mathcal{I}^{a,b}_j}\langle v'^{a,b}_i, \cdot\rangle$ for some clauses $\{\mathcal{I}^{a,b}_j\}$ and vectors $v^{a,b}_i, v'^{a,b}_i$ comprised respectively of vectors of the form $\mathbf{W}^b_a\mathbf{\Sigma}_{a-1}\cdots\mathbf{\Sigma}_0\mathbf{W}_0$ and $\mathbf{W}'^b_a\mathbf{\Sigma}_{a-1}\cdots\mathbf{\Sigma}_0\mathbf{W}'_0$ for all possible diagonal matrices $\mathbf{\Sigma}_i \in \{0,1\}^{k_i\times k_i}$. Then for any $b \in [k_{a+1}]$, note that $F_{a+1,b} = \mathbf{W}^b_{a+1}\phi(F_{a,1}, \ldots, F_{a,k_a})$ and $F'_{a+1,b} = \mathbf{W}'^b_{a+1}\phi(F'_{a,1}, \ldots, F'_{a,k_a})$. By Lemma 3.3.18 and Lemma 3.3.19, if the entries of $\mathbf{W}^b_a$ and $\mathbf{W}'^b_a$ are $w_1, \ldots, w_{k_{a+1}}$ and $w'_1, \ldots, w'_{k_{a+1}}$ respectively, then for every $b' \in [k_a]$, if $w_{b'} \cdot w'_{b'} \ge 0$, then there exist max-min representations for $w_{b'}\phi(F_{a,b'})$ and $w'_{b'}\phi(F_{a,b'})$ with the same set of clauses.

Finally, by Lemma 3.3.20, there exist max-min representations for the scalar-valued functions $F_{a+1,b} = \sum_{b'=1}^{k_a} w_{b'}\phi(F_{a,b'})$ and $F'_{a+1,b} = \sum_{b'=1}^{k_a} w'_{b'}\phi(F'_{a,b'})$ with the same set of clauses. And the vectors in this max-min representation consist of all vectors of the form

$\mathbf{W}^b_{a+1}\mathbf{\Sigma}_a \cdot \cdots \cdot \mathbf{\Sigma}_0 \mathbf{W}_0$ and $\mathbf{W}'^b_{a+1}\mathbf{\Sigma}_a \cdot \cdots \cdot \mathbf{\Sigma}_0 \mathbf{W}'_0$ respectively for $\mathbf{\Sigma}_i \in \{0,1\}^{k_i \times k_i}$. This completes the inductive step. □

3.4 Filtered PCA

In this section we prove our main results on learning kickers and ReLU networks. Throughout, we will make the following base assumption about the function F.

Assumption 3. *F is a kicker which is Λ-Lipschitz for some $\Lambda \geq 1$ and has at most M pieces.*

While our techniques are general enough to work under just this assumption, for our main application to learning ReLU networks (Definition 3.1.1), we can obtain improved runtime guarantees by making the following additional assumption on F.

Assumption 4. *F is computed by a size-S ReLU network*[6] *with depth $L+2$ and weight matrices $\mathbf{W}_0 \in \mathbb{R}^{k_0 \times d}, \ldots \mathbf{W}_L \in \mathbb{R}^{k_L \times k_{L-1}}, \mathbf{W}_{L+1} \in \mathbb{R}^{1 \times k_L}$ satisfying $\|\mathbf{W}_i\|_2 \leq B$ for all $0 \leq i \leq L+1$, for some $B \geq 1$.*[7]

In this section, unless stated otherwise, we will only assume F satisfies Assumption 3, but in certain parts of the proof (e.g. Section 3.4.5), we will get better bounds by additionally making Assumption 4. Formally, our main results are the following:

Theorem 3.4.1. *Given access to samples from the distribution $\mathcal{D}$ corresponding to kicker F satisfying Assumption 3,* FilteredPCAv2*$(\mathcal{D}, \varepsilon, \delta)$ outputs a kicker $\widetilde{F}$ for which $\mathbb{E}[(y - \widetilde{F}(x))^2] \leq \varepsilon^2$ with probability at least $1-\delta$. Furthermore,* FilteredPCAv2 *has sample complexity*

$$d \log(1/\delta) \cdot \mathrm{poly}\left(\exp\left(k^3\Lambda^2/\varepsilon^2\right), M^k\right)$$

and runtime

$$\widetilde{O}(d^2 \log(1/\delta)) \cdot M^{M^2} \cdot \mathrm{poly}\left(\exp\left(k^4\Lambda^2/\varepsilon^2\right), M^{k^2}\right).$$

Theorem 3.4.2. *Given access to samples from the distribution $\mathcal{D}$ corresponding to feedfor-*

[6] Note that this implies $M \leq 2^S$.

[7] Recall from Definition 3.1.1 that we will refer to the rank of $\mathbf{W}_0$ as k to emphasize that F is a kicker with relevant subspace V of dimension k.

ward ReLU network F satisfying Assumption 4, FilteredPCAv2*$(\mathcal{D}, \varepsilon, \delta)$ outputs a ReLU network $\widetilde{F}$ for which $\mathbb{E}[(y - \widetilde{F}(x))^2] \le \varepsilon^2$ with probability at least $1 - \delta$. Furthermore,* FilteredPCAv2 *has sample complexity*

$$d \log(1/\delta) \mathrm{poly}\left(\exp\left(k^3\Lambda^2/\varepsilon^2\right), 2^{kS}, \left(B^{(L+2)}/\Lambda\right)^k\right)$$

and runtime

$$\widetilde{O}(d^2 \log(1/\delta)) \cdot \mathrm{poly}\left(\exp\left(k^3 S^2 \Lambda^2/\varepsilon^2\right), 2^{kS^3}, \left(B^{L+2}/\Lambda\right)^{kS^2}\right).$$

Remark 3.4.3 (Scale Invariance)**.** *Often, guarantees for PAC learning ReLU networks are stated scale-invariantly in terms of the relative error $\mathbb{E}[(y - \widetilde{F}(x))^2]/\mathbb{E}[y^2]$, or equivalently the absolute error $\mathbb{E}[(y - \widetilde{F}(x))^2]$ for the true F satisfying $\mathbb{E}[y^2] = 1$.*

In our general setting, recall from Example 3.1.4 that some dependence on the Lipschitz constant of F is needed. One standard way to achieve this is to normalize the weight matrices of the true underlying network F to have operator norm at most B, in which case the Lipschitz constant of F is at most B^{L+2} and, with our techniques, we can obtain guarantees depending just on B by using Theorem 3.4.1. To obtain improved guarantees, we can additionally assume a better bound of Λ on the Lipschitz constant, and this gives rise to Theorem 3.4.2 above.

Under this normalization in terms of Λ and B, note that the sample complexity and runtime in Theorem 3.4.2 are scale invariant as the quantities Λ/ε and B^{L+2}/Λ are invariant under arbitrary rescalings of the $L + 2$ weight matrices of F. Also note that Λ can be any upper bound *on the actual Lipschitz constant of F, that is, the runtime guarantee in Theorem 3.4.2 does not degrade with the actual Lipschitz constant of F.*

In Section 3.4.1, we prove an anti-concentration result for piecewise-linear functions. We use this in Section 3.4.2 to prove that in an idealized scenario where we had exact access to some ℓ-dimensional $W \subset V$ as well as exact query access to $F|_W$, we would be able to approximately recover a vector in $V \backslash W$ by running one iteration of the main loop of FilteredPCAv2. In the remaining sections, we show how to pass from this idealized scenario to the setting we actually care about, in which we only samples $(x, F(x))$. In

Section 3.4.3 we show that affine thresholds of piecewise-linear functions are stable under small perturbations of the function. Then in Section 3.4.4, we show how to grid over the set of kickers, and in Section 3.4.5 we show how to grid over ReLU networks more efficiently and formally state our algorithm. In Section 3.4.6 we combine these ingredients to argue that as long as we have sufficiently good approximate access to W and $F|_W$, a single iteration of the main loop of FilteredPCAv2 will approximately recover a vector from $V\backslash W$. Lastly, in Section 3.4.7 we conclude the proofs of Theorem 3.4.1 and 3.4.2. At the very end, we discuss briefly why merely adapting the approach of the previous chapter does not work.

3.4.1 Anti-Concentration of Piecewise Linear Functions

In this section, we show that for any continuous piecewise-linear function with some variance, the probability that it exceeds any given threshold is non-negligible.

Lemma 3.4.4. *If $G : \mathbb{R}^m \to \mathbb{R}$ is continuous piecewise-linear and Λ-Lipschitz and $\mathbb{E}[G^2] \geq \sigma^2$, then for any $s \geq 0$,*

$$\Pr[|G| > s] \geq \Omega(\exp(-3ms^2/\sigma^2)) \cdot \frac{s\sigma}{\sqrt{m}\Lambda^2}.$$

Proof. Let $\{(g_i, S_i)\}$ be the pieces of some realization G, and for every i let $u_i \in \mathbb{R}^m$ be the vector for which $g_i(\cdot) = \langle u_i, \cdot\rangle$. By Lemma 3.3.7, we can assume $\|u_i\| \leq \Lambda$ for all i.

Take any i and define

$$\sigma_i^2 \triangleq \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,\mathrm{Id})}[\langle u_i, x\rangle^2 \mid x \in S_i]$$

Note that if i is chosen with probability $\Pr[x \in S_i]$, then $\mathbb{E}_i[\sigma_i^2] \geq \sigma^2$. Because each S_i is a polyhedral cone, sampling $x \sim \mathcal{N}(0, \mathrm{Id})$ conditioned on $x \in S_i$ is equivalent to sampling $r \sim \chi_m^2$, independently sampling $\widehat{x} \sim \mathbb{S}^{m-1}$ conditioned on $\widehat{x} \in S_i$, and outputting $r^{1/2} \cdot \widehat{x}$. It follows that

$$\sigma_i^2 = \mathop{\mathbb{E}}_{r\sim\chi_m^2, \widehat{x}\sim\mathbb{S}^{m-1}}[r\cdot\langle u_i, \widehat{x}\rangle^2 \mid \widehat{x} \in S_i] = \mathop{\mathbb{E}}_{r\sim\chi_m^2}[r]\cdot \mathop{\mathbb{E}}_{\widehat{x}\sim\mathbb{S}^{m-1}}[\langle u_i, \widehat{x}\rangle^2 \mid \widehat{x} \in S_i] = m\cdot \mathop{\mathbb{E}}_{\widehat{x}\sim\mathbb{S}^{m-1}}[\langle u_i, \widehat{x}\rangle^2 \mid \widehat{x} \in S_i].$$

By Fact 1.3.31, $\Pr[|\langle u_i, \widehat{x}\rangle| \geq \sigma_i/\sqrt{2m} \mid \widehat{x} \in S_i] \geq \frac{\sigma_i^2}{2m\|u_i\|^2}$. We conclude that for any $s > 0$,

$$\begin{aligned}\Pr[|\langle u_i, x\rangle| \geq s \mid x \in S_i] &\geq \Pr_{r\sim\chi^2_m}\left[r > 2ms^2/\sigma_i^2\right] \cdot \frac{\sigma_i^2}{2m\|u_i\|^2} \\ &\geq \mathrm{erfc}(s\sqrt{2m}/\sigma_i) \cdot \frac{\sigma_i^2}{2m\Lambda^2} \qquad (3.7)\end{aligned}$$

By Fact 1.3.14, the right-hand side of (3.7) is convex as a function of σ_i^2, so

$$\begin{aligned}\Pr[|G(x)| > s] &\geq \mathop{\mathbb{E}}_i\left[\mathrm{erfc}(s\sqrt{2m}/\sigma_i) \cdot \frac{\sigma_i^2}{2m\Lambda^2}\right] \\ &\geq \mathrm{erfc}(s\sqrt{2m}/\mathop{\mathbb{E}}_i[\sigma_i^2]^{1/2}) \cdot \frac{\mathbb{E}_i[\sigma_i^2]}{2m\Lambda^2} \\ &\geq \mathrm{erfc}(s\sqrt{2m}/\sigma) \cdot \frac{\sigma^2}{2m\Lambda^2} \\ &\geq \sqrt{2/\pi} \cdot \frac{s\sqrt{2m} \cdot \exp(-ms^2/\sigma^2)}{\sigma \cdot (2ms^2/\sigma^2 + 1)} \cdot \frac{\sigma^2}{2m\Lambda^2} \\ &\geq \Omega(\exp(-3ms^2/\sigma^2)) \cdot \frac{s\sigma}{\sqrt{m}\Lambda^2},\end{aligned}$$

where the second step follows by Jensen's and the fourth step follows by Fact 1.3.13. □

3.4.2 An Idealized Calculation

Suppose we had access to an orthonormal collection of vectors $w_1, \ldots, w_\ell$ that are *exactly* in V. Let W denote their span. Suppose further that we had access to the matrix

$$\mathbf{M}^W_\tau \triangleq \Pi_{W^\perp} \mathop{\mathbb{E}}_{x,y}\left[\mathbb{1}[|y - F(\Pi_W x)| > \tau] \cdot (xx^\top - \mathrm{Id})\right]\Pi_{W^\perp}.$$

When the threshold τ is clear from context, we will just refer to this matrix as $\mathbf{M}^W$.

As we will see, if this matrix is nonzero, then its singular vectors with nonzero singular value must lie in V and be orthogonal to $w_1, \ldots, w_\ell$. The main challenge will be to show that this matrix is nonzero. The following proof also applies to the case of $\ell = 0$, in which case $F(\Pi_W x)$ specializes to the zero function and (3.8) specializes to

$$\mathbf{M}^\emptyset_\tau \triangleq \mathop{\mathbb{E}}_{x,y}\left[\mathbb{1}[|y| > \tau] \cdot (xx^\top - \mathrm{Id})\right]. \qquad (3.8)$$

In particular, (3.8) is a matrix we actually have access to at the beginning of the algorithm, and one consequence of the warmup argument below is an algorithm for finding a single vector in V.

We first show that for appropriately chosen τ, either the top singular value of $\mathbf{M}^W_\tau$ is non-negligible, or $\mathbb{E}[(F(x)-F(\Pi_W x)^2]$ is small, that is, F is already sufficiently well-approximated by the function $F|_W$.

Lemma 3.4.5. *Suppose $\mathbb{E}_{x\sim\mathcal{N}(0,Id)}[(F(x)-F(\Pi_W x))^2] \geq \rho^2$ for some $\rho > 0$. For any $\tau > 0$, if a vector is not in the kernel of $\mathbf{M}^W_\tau$, then it must lie in $V\backslash W$. For $\tau \geq \sqrt{2(k-\ell)}\cdot\Lambda$,*

$$\left\langle \mathbf{M}^W_\tau, \Pi_{V\backslash W}\right\rangle \geq \Omega\left(e^{-3k\tau^2/\rho^2}\right)\cdot\frac{(k-\ell)\tau\rho}{\sqrt{k}\Lambda^2}. \tag{3.9}$$

In particular, for this choice of τ, the top singular vector of $\mathbf{M}^W_\tau$ lies in $V\backslash W$ and has singular value at least $\lambda^{(\ell)}_\tau \triangleq \Omega\left(e^{-3k\tau^2/\rho^2}\right)\cdot\frac{\tau\rho}{\sqrt{k}\Lambda^2}$.

Proof. The first part just follows from the fact that any $u \in \Pi_W$ is clearly in the kernel, and for any $u \in \mathbb{S}^{d-1}$ orthogonal to V, $\langle u, x\rangle$ and $F(x)$ are independent, so

$$u^\top \mathbf{M}^W_\tau u = \mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,1)}[g^2-1]\cdot\mathop{\mathbb{E}}_x[\mathbb{1}[|F(x)-F(\Pi_W v)| > \tau]] = 0.$$

For (3.9), we would like to apply Lemmas 3.3.10 and 3.4.4 to the continuous piecewise-linear function $G(x) \triangleq F(x) - F(\Pi_W x)$. Pick an orthonormal basis $w_{\ell+1},\ldots,w_k$ for $V\backslash W$. For any x for which $\|\Pi_{V\backslash W} x\| \leq 1$, Lemma 3.3.10 implies $|G(x)| \leq \Lambda$. So by positive homogeneity (see Fact 3.3.9) of $G(x)$ and the definition of τ, $|G(x)| > \tau$ only if $\|\Pi_{V\backslash W}x\|^2 \geq 2(k-\ell)$, so

$$\begin{aligned}\sum_{i=\ell+1}^{k} w_i^\top \mathbf{M}^W_\tau w_i &= \mathop{\mathbb{E}}_x\left[\mathbb{1}[|G(x)| > \tau]\cdot\left(\|\Pi_{V\backslash W}x\|^2 - (k-\ell)\right)\right] \\ &\geq (k-\ell)\cdot\Pr_x[G(x) > \tau].\end{aligned}$$

(3.9) then follows from Lemma 3.4.4 applied to G.

The final statement in Lemma 3.4.5 follows by averaging. □

If ε is the target L_2 error to which we want to learn F, we will only ever work with

$\rho \geq \Omega(\varepsilon)$. In the sequel, we will take

$$\tau = c\sqrt{k} \cdot \Lambda \tag{3.10}$$

for sufficiently large absolute constant $c > 0$. As a result, we have that

$$\lambda_{\tau}^{(\ell)} \geq \Omega\left(e^{-O(k^2\Lambda^2/\varepsilon^2)}\right) \cdot (\varepsilon/\Lambda) \triangleq \underline{\lambda}. \tag{3.11}$$

3.4.3 Stability of Piecewise Linear Threshold Functions

To get an iterative algorithm for finding all relevant directions of F, we need to show an analogue of Lemma 3.9 in the setting when we only have access to directions $\widetilde{w}_1, \ldots, \widetilde{w}_\ell$ which are *close* to the span of V, and when we only have access to an *approximation* of the function $F|_W$.

In this section, we show the following stability result for affine thresholds of piecewise-linear functions:

Lemma 3.4.6. *Let $f, g, g' : \mathbb{R}^d \to \mathbb{R}$ be piecewise-linear functions. For any $\tau > 0$, if g, g' are (m, η)-structurally-close and f has a realization with at most m pieces, then*

$$\Pr_{x \sim \mathcal{N}(0,Id)}[|g(x) - f(x)| > \tau \wedge |g'(x) - f(x)| \leq \tau] \leq 9\eta m^2/\tau \tag{3.12}$$

An important building block of the proof is the special case where $f = 0$ and g, g' are linear, which was shown in Lemma 1.3.35.

Proof of Lemma 3.4.6. The left-hand side of (3.12) is at most

$$\Pr_{x \sim \mathcal{N}(0,\mathrm{Id})}[g(x) - f(x) > \tau \wedge g'(x) - f(x) \leq \tau] + \Pr_{x \sim \mathcal{N}(0,\mathrm{Id})}[g(x) - f(x) < -\tau \wedge g'(x) - f(x) \geq -\tau], \tag{3.13}$$

and by symmetry it suffices to upper bound the former probability on the right-hand side of (3.13) by $O(\eta m^2/\tau)$.

By definition of (m, η)-structural-closeness, we can express g and g' as $\max_j \min_{i \in \mathcal{I}_j} \langle u_i, \cdot \rangle$ and $\max_j \min_{i \in \mathcal{I}_j} \langle u'_i, \cdot \rangle$ respectively, for vectors $\{u_i\}_{i \in [m]}$ and $\{u'_i\}_{i \in [m]}$ for which $\|u_i - u'_i\| \leq \eta$

for all i.

We proceed via a hybrid argument. Take any $0 \leq i \leq m$. Let $u_1^{(i)}, \ldots, u_{i-1}^{(i)}$ be $u_1, \ldots, u_{i-1}$, and let $u_i^{(i)}, \ldots, u_m^{(i)}$ be the vectors $u'_i, \ldots, u'_m$. Define the function $g^{(i)} = \max_a \min_{b \in \mathcal{I}_a} \langle u_i^{(i)}, x \rangle$ so that $g^{(0)}(x) = \max_a \min_{b \in \mathcal{I}_a} \langle u'_b, x \rangle$ and $g^{(m)}(x) = \max_a \min_{b \in \mathcal{I}_a} \langle u_b, x \rangle$.

We claim that for any x, $g^{(i-1)}(x)$ and $g^{(i)}(x)$ are sandwiched between $\langle u'_i, x \rangle$ and $\langle u_i, x \rangle$, in the sense that

$$\langle u'_i, x \rangle \geq g^{(i-1)}(x) \geq g^{(i)}(x) \geq \langle u_i, x \rangle \qquad \text{or} \qquad \langle u'_i, x \rangle \leq g^{(i-1)}(x) \leq g^{(i)}(x) \leq \langle u_i, x \rangle. \tag{3.14}$$

This would imply

$$\Pr[g^{(i)}(x) - f(x) > \tau \wedge g^{(i-1)}(x) - f(x) \leq \tau] \leq \Pr[\langle u_i, x \rangle - f(x) > \tau \wedge \langle u'_i, x \rangle - f(x) \leq \tau] \tag{3.15}$$

because either the left-hand side of (3.15) is zero, or or the event on the left-hand side immediately implies the one on the right-hand side.

Denote by $\{(\langle w_i, \cdot \rangle, S_i)\}_{i \in [m]}$ the pieces of some realization of f. We would then have

$$\begin{aligned}
&\Pr[g(x) - f(x) > \tau \wedge g'(x) - f(x) \leq \tau] \\
&\quad \leq \sum_{i=1}^{m} \Pr[\langle u_i, x \rangle - f(x) > \tau \wedge \langle u'_i, x \rangle - f(x) \leq \tau] \\
&\quad = \sum_{\ell=1}^{m} \sum_{i=1}^{m} \Pr[x \in S_\ell \wedge \langle u_i - w_\ell, x \rangle > \tau \wedge \langle u'_i - w_\ell, x \rangle \leq \tau] \\
&\quad \leq \sum_{\ell=1}^{m} \sum_{i=1}^{m} \Pr[\langle u_i - w_\ell, x \rangle > \tau \wedge \langle u'_i - w_\ell, x \rangle \leq \tau] \leq O\left(\eta m^2 / \tau\right),
\end{aligned}$$

where the first step follows by triangle inequality and (3.15), and the last step follows by Lemma 1.3.35.

To complete the proof, we now turn to proving that the quantities $g^{(i)}(x)$ and $g^{(i-1)}(x)$ are sandwiched between $\langle u'_i, x \rangle$ and $\langle u_i, x \rangle$, which will imply (3.15). Suppose that $g^{(i-1)}(x) = \langle u_j^{(i-1)}, x \rangle$ for some index j.

Case 1: $\langle u'_i, x \rangle \geq \langle u_j^{(i-1)}, x \rangle$.

In this case $\min_{b \in \mathcal{I}_a} \langle u_b^{(i-1)}, x \rangle \leq \langle u'_i, x \rangle$ for all a. If $\langle u_i, x \rangle \geq \langle u'_j, x \rangle$, then changing u'_i to

u_i will not change the values of any of the clauses. So suppose $\langle u_i, x\rangle < \langle u'_j, x\rangle$, in which case the value of the function cannot increase. Then if index i appears in any clause $\mathcal{I}_a$ for which $\min_{b\in\mathcal{I}_a}\langle u^{(i-1)}_b, x\rangle = \langle u^{(i-1)}_j, x\rangle$, then $g^{(i)}(x) \geq \langle u_i, x\rangle$. Otherwise, the value of the function stays the same. We conclude that the first inequality in (3.14) holds.

Case 2: $\langle u'_i, x\rangle < \langle u^{(i-1)}_j, x\rangle$.

In this case there is some $\mathcal{I}_a$ for which $\langle u^{(i-1)}_j, x\rangle = \min_{b\in\mathcal{I}_a}\langle u^{(i-1)}_b, x\rangle$ and in which index i does not appear. If $\langle u_i, x\rangle \leq \langle u'_i, x\rangle$, then changing u'_i to u_i will not change the value of this $\mathcal{I}_a$ clause, and the values of the other clauses will not increase, so the value of the function will not change. So suppose $\langle u_i, x\rangle > \langle u'_i, x\rangle$. Changing u'_i to u_i will not affect any clause $\mathcal{I}_a$ not containing i or for which $\min_{b\in\mathcal{I}_a}\langle u^{(i-1)}_b, x\rangle \leq u'_i$. For all other clauses, their value will either stay the same or increase to u_i, in which case $g^{(i)}(x) \leq \langle u_i, x\rangle$. We conclude that the second inequality in (3.14) holds. □

3.4.4 Netting Over Piecewise Linear Functions

Suppose we have recovered an ℓ-dimensional subspace $\widetilde{W}$ that approximately lies within V. In this section we show how to produce a finite list of candidate kickers with relevant subspace $\widetilde{W}$, one of which is guaranteed to approximate F restricted to some ℓ-dimensional subspace W. Ignoring the finiteness of this list for now, we first show that as long as $\widetilde{W}$ is sufficiently close to lying within V, there exists *some* kicker close to *some* restriction $F|_W$.

Lemma 3.4.7. *Let $\widetilde{w}_1, \ldots, \widetilde{w}_\ell$ be a frame ν-nearly within V, with span $\widetilde{W}$. There exist an ℓ-dimensional subspace $W \subset V$ and a Λ-Lipschitz kicker $\widetilde{F}^*$ with relevant subspace $\widetilde{W}$ which is $(M, 2\sqrt{\nu}\cdot\ell\Lambda)$-structurally-close to $F|_W$.*

Proof of Lemma 3.4.7. By Lemma 1.3.10, there exist orthonormal vectors $w_1, \ldots, w_\ell$ for which $\|w_i - \widetilde{w}_i\| \leq 2\sqrt{\nu\ell}$. Let W be their span.

The function $F|_W$ is a continuous piecewise-linear function with at most M pieces, so by Theorem 3.3.11 and Lemma 3.3.7, there exist vectors $u_1, \ldots, u_M \in W$ and subsets $\mathcal{I}_1, \ldots, \mathcal{I}_m \subseteq [M]$ for which $F(x) = \max_{j\in[m]}\min_{i\in\mathcal{I}_j}\langle u_i, x\rangle$ and $\|u_i\| \leq \Lambda$ for all i. For any $i \in [M]$, write $u_i = \sum_{i'\in[\ell]}\alpha_{i,i'}w_{i'}$. Define $\widetilde{u}^*_i \triangleq \sum_{i'\in[\ell]}\alpha_{i,i'}\widetilde{w}_{i'}$ and define the kicker $\widetilde{F}^*$ with relevant subspace $\widetilde{W}$ by $\widetilde{F}^*(x) \triangleq \max_{j\in[m]}\min_{i\in\mathcal{I}_j}\langle\widetilde{u}^*_i, x\rangle$.

Note that for any i,

$$\|\widetilde{u}^*_i - u_i\| = \sum_{i'\in[\ell]} \alpha_{i,i'}\|\widetilde{w}_{i'} - w_{i'}\| \le 2\sqrt{\nu\ell}\cdot\sum_{i'}|\alpha_{i,i'}| \le 2\sqrt{\nu}\cdot\ell\|u_i\| \le 2\sqrt{\nu}\cdot\ell\Lambda,$$

where the penultimate step is by Cauchy-Schwarz, so $\widetilde{F}^*$ is $(M, 2\sqrt{\nu}\cdot\ell\Lambda)$-structurally-close to $F|_W$ as claimed. Lastly, note that $\|\widetilde{u}^*_i\| = \|u_i\| \le \Lambda$, so $\widetilde{F}^*$ is indeed Λ-Lipschitz. □

We now show that the existential guarantee of Lemma 3.3.16 implies that if we enumerate over a fine enough net of kickers, then we can recover an approximation to $\widetilde{F}^*$ from Lemma 3.4.7 in time singly exponential in poly(M).

```
Algorithm 8: EnumerateKickers($\widetilde{W}$, $\varepsilon'$)
Input: Subspace $\widetilde{W}$ spanned by orthonormal vectors $\widetilde{w}_1,\ldots,\widetilde{w}_\ell$, granularity $\varepsilon' > 0$
Output: List of kickers $\widetilde{F}$ with relevant subspace $\widetilde{W}$
$\mathcal{L} \leftarrow \emptyset$.
Let $\mathcal{N}$ be an $\varepsilon'\Lambda$-net over the set of vectors in $\widetilde{W}$ with norm at most $\Lambda$.
for $\widetilde{u}_1,\ldots,\widetilde{u}_M \in \mathcal{N}$ do
    for functions $A : \Omega_M \to [M]$ do
        Let $\widetilde{F}$ be the kicker given by $\widetilde{F}(x) = \sum_{\omega\in\Omega_M} \mathbb{1}\left[\{\langle\widetilde{u}_i, x\rangle\}_{i\in[M]} \vdash \omega\right]\cdot\langle\widetilde{u}_{A(\omega)}, x\rangle$.
        Append $\widetilde{F}$ to $\mathcal{L}$.
return $\mathcal{L}$.
```

Lemma 3.4.8. *Take any $\varepsilon' > 0$. Given a frame $\widetilde{w}_1,\ldots,\widetilde{w}_\ell$ with span $\widetilde{W}$, for any Λ-Lipschitz kicker $\widetilde{F}^*$ with relevant subspace $\widetilde{W}$, there exists a kicker $\widetilde{F}$ with relevant subspace $\widetilde{W}$ in the output $\mathcal{L}$ of* EnumerateKickers*$(\widetilde{W}, \varepsilon')$ which is $(M, \varepsilon'\Lambda)$-structurally-close to $\widetilde{F}$. Furthermore, $|\mathcal{L}| \le M^{M^2}\cdot(1+2/\varepsilon')^\ell$.*

In particular, if $\widetilde{w}_1,\ldots,\widetilde{w}_\ell$ is a frame ν-nearly within V, then for $\varepsilon' = 2\sqrt{\nu}\cdot\ell$, $\mathcal{L}$ contains a kicker $\widetilde{F}$ which is $(M, C_{\mathsf{piecewise}}\sqrt{\nu})$-structurally-close to $F|_W$ for some ℓ-dimensional subspace $W \subseteq V$, where

$$C_{\mathsf{piecewise}} \triangleq 4k\Lambda.$$

Furthermore, $|\mathcal{L}| \le M^{M^2}O(1/\sqrt{\nu})^\ell$ in this case.

Proof. By Lemma 3.3.16, the function $\widetilde{F}^*$ in the hypothesis can be written in the form

$\widetilde{F}^*(x) = \sum_{\omega\in\Omega_M} \mathbb{1}\left[\{\langle \widetilde{u}^*_i, x\rangle\}_{i\in[M]} \vdash \omega\right] \cdot \langle \widetilde{u}^*_{A(\omega)}, x\rangle$ for some vectors $\{\widetilde{u}^*_i\}_{i\in[M]}$ and function $A : \Omega_M \to [M]$.

Because $\mathcal{N}$ in Step 2 of EnumerateKickers is an $\varepsilon'\Lambda$-net over the set of vectors in $\widetilde{W}$ with norm at most Λ, there exist vectors $\widetilde{u}_1, ..., \widetilde{u}_M \in \mathcal{N}$ for which $\|\widetilde{u}_i - \widetilde{u}^*_i\| \le \varepsilon'\Lambda$. If we define $\widetilde{F}$ by $\widetilde{F}(x) = \sum_{\omega\in\Omega_M} \mathbb{1}\left[\{\langle \widetilde{u}_i, x\rangle\}_{i\in[M]} \vdash \omega\right] \cdot \langle \widetilde{u}_{A(\omega)}, x\rangle$, then by design, $\widetilde{F}$ is $(M, \varepsilon'\Lambda)$-structurally-close to $\widetilde{F}$.

It remains to bound the size of $\mathcal{L}$. For any $\varepsilon' > 0$ there is an ε'-net $\mathcal{N}'_{\varepsilon'}$ for the L_2 unit ball in $\widetilde{W}$ of size at most $(1 + 2/\varepsilon')^\ell$. Define $\mathcal{N} \triangleq \Lambda \cdot \mathcal{N}'_{\varepsilon'}$. Furthermore, there are $|\Omega_M|^M \le M^{M^2}$ functions $A : \Omega_M \to [M]$. This yields the desired bound on $|\mathcal{L}|$.

The final part of the lemma follows by invoking Lemma 3.4.7 and noting that the lattice polynomial representation of $\widetilde{F}^*$ and that of $F|_W$ are identical in the proof of Lemma 3.4.7, so the structural closeness of $\widetilde{F}$ to $F|_W$ follows by triangle inequality. □

3.4.5 Netting Over Neural Networks

Enumerating over arbitrary kickers with M pieces requires runtime scaling exponentially in poly(M). For ReLU networks of size S, M could be as large as $\exp(S)$, so naively using EnumerateKickers in our application to learning ReLU networks would incur doubly exponential dependence on k in the runtime. In this section we show how to enumerate over ReLU networks more efficiently. We first prove the analogue of Lemma 3.4.7 for ReLU networks.

Lemma 3.4.9. *Suppose F additionally satisfies Assumption 4. Let $\widetilde{w}_1, \ldots, \widetilde{w}_\ell$ be a frame ν-nearly within V, with span $\widetilde{W}$. There exist an ℓ-dimensional subspace $W \subset V$ and weight matrix $\mathbf{W}^*_0 \in \mathbb{R}^{k_0\times d}$ with rows in $\widetilde{W}$ for which*

$$\|\mathbf{W}_0\Pi_W - \mathbf{W}^*_0\|_2 \le 2\sqrt{\nu} \cdot \ell\sqrt{k} \cdot B \tag{3.16}$$

and for which $\|\mathbf{W}^*_0\|_2 \le B$.

Proof. As in the proof of Lemma 3.4.7, Lemma 1.3.10 yields orthonormal vectors $w_1, \ldots, w_\ell$ for which $\|w_i - \widetilde{w}_i\| \le 2\sqrt{\nu\ell}$. Let W be their span.

If F has weight matrices $\mathbf{W}_0 \in \mathbb{R}^{k_0 \times d}, \mathbf{W}_1 \in \mathbb{R}^{k_1 \times k_0}, \ldots, \mathbf{W}_{L+1} \in \mathbb{R}^{1 \times k_L}$, then $F|_W$ is a ReLU network with weight matrices $\mathbf{W}_0 \Pi_W, \mathbf{W}_1, \ldots, \mathbf{W}_{L+1}$. Denoting the rows of $\mathbf{W}_0 \Pi_W \in \mathbb{R}^{k_0 \times d}$ as $u_1, \ldots, u_{k_0}$, we may write them as $u_i = \sum_{i' \in [\ell]} \alpha_{i,i'} w_{i'}$ for $i \in [k_0]$.

Define $\widetilde{u}_i^* \triangleq \sum_{i' \in [\ell]} \alpha_{i,i'} \widetilde{w}_{i'}$. As in the proof of Lemma 3.4.7, we have that

$$\|\widetilde{u}_i^* - u_i\| \leq 2\sqrt{\nu} \cdot \ell \|u_i\| \leq 2\sqrt{\nu} \cdot \ell B,$$

where in the last step we have used the fact that the maximum norm of any row of $\mathbf{W}_0 \Pi_W$ is at most the maximum norm of any row of $\mathbf{W}_0$, which is upper bounded by $\|\mathbf{W}_0\|_2 \leq B$.

Let $\widetilde{\mathbf{W}}_0^*$ denote the matrix whose rows consist of $\widetilde{u}_1^*, \ldots, \widetilde{u}_{k_0}^*$. We have that

$$\|\mathbf{W}_0 \Pi_W - \widetilde{\mathbf{W}}_0^*\|_2 \leq \|\mathbf{W}_0 \Pi_W - \widetilde{\mathbf{W}}_0^*\|_F \leq 2\sqrt{\nu} \cdot \ell \sqrt{k} \cdot B$$

as claimed. Finally, the bound on $\|\mathbf{W}_0^*\|_2$ follows from the fact that $\mathbf{W}_0^* = \mathbf{W}_0 \cdot \mathbf{O} \cdot \Pi_W$ for an orthogonal matrix $\mathbf{O}$ mapping the frame $\{w_1, ..., w_\ell\}$ to $\{\widetilde{w}_1, ..., \widetilde{w}_\ell\}$. □

Algorithm 9: EnumerateNetworks($\widetilde{W}$, ε')

```
Input: Subspace $\widetilde{W}$ spanned by orthonormal vectors $\widetilde{w}_1, \ldots, \widetilde{w}_\ell$, granularity $\varepsilon' > 0$
Output: List of size-$S$ ReLU networks $\widetilde{F}$ with relevant subspace $\widetilde{W}$
$\mathcal{L} \leftarrow \emptyset$.
for tuples $(\widetilde{k}_0, ..., \widetilde{k}_{L+1}) \in \mathbf{Z}_{>0}^{L+2}$ satisfying $\sum_{i=0}^{L+1} \widetilde{k}_i = S$ do
    For every $0 \leq i \leq L+1$, let $\mathcal{N}_i$ be an $\varepsilon'$-net (in operator norm) over the set of
      matrices in $\mathbb{R}^{\widetilde{k}_i \times \widetilde{k}_{i-1}}$ with operator norm at most $B + \varepsilon'$.
    for $\widetilde{\mathbf{W}}_0 \in \mathcal{N}_0, ..., \widetilde{\mathbf{W}}_{L+1} \in \mathcal{N}_{L+1}$ do
        Define the ReLU network $\widetilde{F}$ with weight matrices $\mathsf{clip}_{\varepsilon'}(\mathbf{W}_0), ..., \mathsf{clip}_{\varepsilon'}(\mathbf{W}_{L+1})$.
        Append $\widetilde{F}$ to $\mathcal{L}$.
return $\mathcal{L}$.
```

We can now show the analogue of Lemma 3.4.8 for ReLU networks.

Lemma 3.4.10. *Take any $0 < \varepsilon' \leq B$ and any frame $\widetilde{w}_1, \ldots, \widetilde{w}_\ell$ with span $\widetilde{W}$. For any ReLU network $\widetilde{F}^*$ of size S with relevant subspace $\widetilde{W}$ and depth $L+2$ whose weight matrices have operator norm at most B, there exists a ReLU network $\widetilde{F}$ with relevant subspace $\widetilde{W}$ in* *the output $\mathcal{L}$ of* ENUMERATENETWORKS*$(\widetilde{W},\varepsilon')$ which is $(2^S, 2^{O(L)}B^{L+1}\varepsilon')$-structurally-close (as a piecewise-linear function) to $\widetilde{F}$. Furthermore, $|\mathcal{L}| \le 2^{O(S)} \cdot (1+4B/\varepsilon')^{O(S^2)}$.*

In particular, if $\widetilde{w}_1, ..., \widetilde{w}_\ell$ is a frame ν-nearly within V, then for $\varepsilon' = 2\sqrt{\nu}\cdot\ell\sqrt{k}\cdot B$, $\mathcal{L}$ contains a ReLU network $\widetilde{F}$ which is $(M, C_{\mathsf{network}}\sqrt{\nu})$-structurally-close to $F|_W$ for some ℓ-dimensional subspace $W \subseteq V$, where

$$C_{\mathsf{network}} \triangleq 2^{O(L)}B^{L+2}k^{3/2}$$

Furthermore, $|\mathcal{L}| \le O(1/\sqrt{\nu})^{O(S^2)}$ in this case.

Proof. Let $\mathbf{W}'_0 \in \mathbb{R}^{k'_0\times d}, \dots, \mathbf{W}_{L+1} \in \mathbb{R}^{1\times k_L}$ denote the weight matrices of $\widetilde{F}^*$. Consider the iteration of the outer loop of ENUMERATENETWORKS in which the architecture of $\widetilde{F}^*$ is guessed correctly, that is, for which $\widetilde{k}_i = k'_i$ for all $0 \le i \le L+1$. By the choice of nets, there is some iteration of the inner loop of the algorithm for which the weight matrices $\{\widetilde{\mathbf{W}}_i\}$ satisfy

$$\|\mathbf{W}'_i - \widetilde{\mathbf{W}}_i\|_2 \le \varepsilon' \ \forall\ 0 \le i \le L+1. \tag{3.17}$$

Define the ReLU network $\widetilde{F}$ with relevant subspace $\widetilde{W}$ to have weight matrices $\widetilde{\mathbf{W}}_0, \widetilde{\mathbf{W}}_1, \dots, \widetilde{\mathbf{W}}_{L+1}$. By the fact that operator norm closeness implies entrywise closeness, together with Fact 3.3.1 and Theorem 3.3.17, there are lattice polynomial representations for $\widetilde{F}^*$ and $\widetilde{F}$ with identical clauses, and for which the vectors at the leaves consist of $\mathbf{W}'_{L+1}\mathbf{\Sigma}_L\mathbf{W}'_L\cdots\mathbf{\Sigma}_0\mathbf{W}'_0\Pi_W$ and $\widetilde{\mathbf{W}}_{L+1}\mathbf{\Sigma}_L\widetilde{\mathbf{W}}_L\cdots\mathbf{\Sigma}_0\widetilde{\mathbf{W}}_0$ respectively for all possible diagonal matrices $\mathbf{\Sigma}_i \in \{0,1\}^{k'_i\times k'_i}$. For any such choice of matrices $\{\mathbf{\Sigma}_i\}$, note that

$$\begin{aligned}
&\|\mathbf{W}'_{L+1}\mathbf{\Sigma}_L\mathbf{W}'_L\cdots\mathbf{W}'_0 - \widetilde{\mathbf{W}}_{L+1}\mathbf{\Sigma}_L\widetilde{\mathbf{W}}_L\cdots\widetilde{\mathbf{W}}_0\| \\
&\quad\le \|(\mathbf{W}'_{L+1} - \widetilde{\mathbf{W}}_{L+1})\mathbf{\Sigma}_L\mathbf{W}'_L\cdots\mathbf{W}'_0\| + \cdots + \|\widetilde{\mathbf{W}}_{L+1}\mathbf{\Sigma}_L\widetilde{\mathbf{W}}_L\cdots(\mathbf{W}'_0 - \widetilde{\mathbf{W}}_0)\| \\
&\quad\le \|\mathbf{W}'_{L+1} - \widetilde{\mathbf{W}}_{L+1}\|\prod_{i=0}^{L}\|\mathbf{W}'_i\|_2 + \cdots + \prod_{i=1}^{L+1}\|\widetilde{\mathbf{W}}_i\|_2\|\mathbf{W}'_0 - \widetilde{\mathbf{W}}_0\|_2 \\
&\quad\le (L+2)\cdot(B+\varepsilon')^{L+1}\cdot\varepsilon' \\
&\quad\le 2^{O(L)}B^{L+1}\cdot\varepsilon',
\end{aligned} \tag{3.18}$$

where in the last step we used the assumption that $\varepsilon' \le B$. This implies the claim about structural closeness.

We next bound the size of $|\mathcal{L}|$. For any choice of $\widetilde{k}_0, ..., \widetilde{k}_{L+1}$, note that by Corollary 1.3.29,

$$\left|\mathcal{N}_{\widetilde{k}_0} \times \cdots \times \mathcal{N}_{\widetilde{k}_{L+1}}\right| \le (1 + 4B/\varepsilon')^{L\widetilde{k}_0 + \widetilde{k}_0\widetilde{k}_1 + \cdots + \widetilde{k}_L\widetilde{k}_{L+1} + \widetilde{k}_{L+1}}$$
$$\le (1 + 4B/\varepsilon')^{O(S^2)}$$

where in the penultimate step we used that

$$L\widetilde{k}_0 + \widetilde{k}_0\widetilde{k}_1 + \cdots + \widetilde{k}_L\widetilde{k}_{L+1} + \widetilde{k}_{L+1} \le (L + \widetilde{k}_0 + \cdots + \widetilde{k}_{L+1})(\widetilde{k}_0 + \cdots + \widetilde{k}_{L+1} + 1) = (L+S)(S+1) \le O(S^2).$$

There are $\binom{S+L+1}{L+1} = 2^{O(S)}$ choices of $(\widetilde{k}_0, \ldots, \widetilde{k}_{L+1})$ in the outer loop of EnumerateNetworks, so $|\mathcal{L}| \le 2^{O(S)} \cdot (1 + 4B/\varepsilon')^{O(S^2)}$ as claimed.

Finally, to obtain the last part of the lemma, we can take $\widetilde{F}^*$ above to have the same weight matrices as F except for the input layer, which we will take to be $\mathbf{W}'_0 \triangleq \widetilde{\mathbf{W}}^*_0$ for the weight matrix guaranteed by Lemma 3.4.9. By (3.16), this choice of $\mathbf{W}'_0$ is close to $\mathbf{W}_0\Pi_W$ for some subspace $W \subseteq V$. Take $\varepsilon' = 2\sqrt{\nu} \cdot \ell\sqrt{k} \cdot B$. For $\{\widetilde{\mathbf{W}}_i\}$ satisfying (3.17), by triangle inequality (3.16) we get that

$$\|\mathbf{W}_0\Pi_W - \widetilde{\mathbf{W}}_0\|_2 \le \|\mathbf{W}_0\Pi_W - \mathbf{W}'_0\|_2 + \|\mathbf{W}'_0 - \widetilde{\mathbf{W}}_0\|_2 \le 2\varepsilon'.$$

Using this, by a calculation analogous to the one leading to (3.18), we find that $\widetilde{F}$ is $(2^S, 2^{O(L)}B^{L+1}\varepsilon')$-structurally-close to $F|_W$, from which the claim follows by our choice of $\varepsilon' = 2\sqrt{\nu} \cdot \ell\sqrt{k} \cdot B$. In this case, we get that $|\mathcal{L}| \le 2^{O(S)}(1 + 2/\sqrt{\nu})^{O(S^2)} \le O(1/\sqrt{\nu})^{O(S^2)}$ as claimed. □

With subroutines for enumerating over ReLU networks and kickers in hand, we can now formally state our algorithm, FilteredPCAv2 (see Algorithm 10 below). The algorithm as stated applies to the case where F is a neural network satisfying Assumptions 3 and 4, but we can easily modify the algorithm to work in the case where F is only a kicker satisfying Assumption 3 by replacing the call to EnumerateNetworks$(\widetilde{W}, 2\sqrt{\nu_0} \cdot \ell\sqrt{k} \cdot B)$ in Line 9 with a call to EnumerateKickers($\widetilde{W}, 2\sqrt{\nu_0}\cdot\ell$), the call to EnumerateNetworks($\widetilde{W}, B^{-L-1}2^{-\Omega(L)}\cdot\varepsilon/\sqrt{k}$) in Line 19 with a call to EnumerateKickers($\widetilde{W}, \varepsilon/(2\sqrt{k}\Lambda)$), and the assignment $N' \leftarrow \mathrm{poly}(B^{L+2}, k, 1/\varepsilon)\cdot\log(1/\delta)$ in Line 20 with the assignment $N' \leftarrow \mathrm{poly}(\Lambda, k, 1/\varepsilon)\cdot\log(1/\delta)$).

3.4.6 Perturbation Bounds

We now show how to leverage Lemma 3.4.6 to show that even with access to a subspace $\widetilde{W}$ which is only approximately within V as well as the restriction of F to that subspace, we can recover another vector orthogonal to $\widetilde{W}$ which mostly lies within V.

The first step is to show that in this approximate setting, the analogue of $\mathbf{M}^W$ from Section 3.4.2 is spectrally close to $\mathbf{M}^W$. It is in showing this perturbation bound that we invoke the stability result of Section 3.4.3.

Lemma 3.4.11. *Suppose F only satisfies Assumption 3 (resp. both Assumptions 3 and 4). Let $\widetilde{w}_1, \ldots, \widetilde{w}_\ell \in \mathbb{S}^{d-1}$ be a frame ν-nearly within V, with span $\widetilde{W}$. For $* \in \{\textsf{piecewise}, \textsf{network}\}$, define*

$$\xi_*(\nu) \triangleq O\left(k\left(\frac{C_*\sqrt{\nu}M^2}{c\sqrt{k}\Lambda}\right)^{1-1/k} \vee \sqrt{\nu k}\right) \tag{3.19}$$

and suppose $N \geq \Omega(\{d \vee \log(1/\delta)\}/\xi_*^2)$.

Given subspace $W \subseteq V$ and $\widetilde{F}$ for which $F|_W$ and $\widetilde{F}$ are $(M, C_{\textsf{piecewise}}\sqrt{\nu})$-structurally-close (resp. $(M, C_{\textsf{network}}\sqrt{\nu})$-structurally close), then we have that

$$\|\widetilde{\mathbf{M}}^{\widetilde{W}}_{\textsf{emp}} - \mathbf{M}^W\|_2 \leq 3\xi(\nu)$$

with probability at least $1-\delta$.

Proof. For convenience denote $\widetilde{\mathbf{M}}^{\widetilde{W}}_{\textsf{emp}}$ and $\mathbf{M}^W$ by $\widetilde{\mathbf{M}}_{\textsf{emp}}$ and $\mathbf{M}$ respectively. Also, depending on whether F only satisfies Assumption 3 or both Assumptions 3 and 4, define $C_* \triangleq C_{\textsf{piecewise}}$ or $C_* \triangleq C_{\textsf{network}}$ respectively. It will also be convenient to define

$$\mathbf{M}' \triangleq \Pi_{W^\perp} \mathop{\mathbb{E}}_{x,y}\left[\mathbb{1}[|y - F|_W(x)| > \tau]\cdot(xx^\top - \mathrm{Id})\right]\Pi_{W^\perp}$$

```
Algorithm 10: FilteredPCAv2($\mathcal{D}, \varepsilon, \delta$)
  Input: Sample access to $\mathcal{D}$, target error $\varepsilon$, failure probability $\delta$
  Output: Size-$S$ ReLU network $\widetilde{F} : \mathbb{R}^d \to \mathbb{R}$ for which $\|\widetilde{F} - F\| \le O(\varepsilon)$ with
          probability at least $1 - \delta$
  $\mathcal{W} \leftarrow \emptyset$.
  $\tau \leftarrow c\sqrt{k} \cdot \Lambda$ as in (3.10).
  $\nu_0 \leftarrow \mathrm{poly}(k^k, 1/\underline{\lambda}^k, M^k, \Lambda)^{-1}$, where $\underline{\lambda}$ is defined in (3.11).
  $\xi \leftarrow O\left(k\left(\sqrt{\nu_0 k} \cdot M^2/c\right)^{1-1/k}\right)$ as in (3.19).
  $N \leftarrow \Omega(\{d \vee \log(2k/\delta)\}/\xi^2)$.
  for $0 \le \ell \le k-1$ do
      Draw samples $(x_1, y_1), \ldots, (x_N, y_N) \sim \mathcal{D}$.
      If $\mathcal{W} = \{\widetilde{w}_1, \ldots, \widetilde{w}_\ell\}$, let $\widetilde{W}$ denote the span of these vectors.
      $\mathcal{L} \leftarrow$EnumerateNetworks($\widetilde{W}, 2\sqrt{\nu_0} \cdot \ell\sqrt{k} \cdot B$).
      for $\widetilde{F} \in \mathcal{L}$ do
          Form the matrix
            $\widetilde{\mathbf{M}}^{\widetilde{W}}_{\mathsf{emp}} \triangleq \Pi_{\widetilde{W}^\perp}\left(\sum_{i=1}^N \mathbb{1}\left[|y_i - \widetilde{F}(\Pi_{\widetilde{W}} x)| > \tau\right] \cdot (x_i x_i^\top - \mathrm{Id})\right)\Pi_{\widetilde{W}^\perp}$.
          Run ApproxBlockSVD($\widetilde{\mathbf{M}}^{\widetilde{W}}_{\mathsf{emp}}, \underline{\lambda}/1000, \delta/(2|\mathcal{L}|k)$) to obtain approximate
            top singular vector $\widetilde{w}^{\ell+1}$.
          $\lambda \leftarrow (\widetilde{w}^{\ell+1})^\top \widetilde{\mathbf{M}}^{\widetilde{W}}_{\mathsf{emp}} \widetilde{w}^{\ell+1}$.
          if $\lambda \ge 9\underline{\lambda}/16$ then
              Append $\widetilde{w}^{\ell+1}$ to $\mathcal{W}$ and exit out of this inner loop and increment $\ell$.
      if no $\widetilde{w}^{\ell+1}$ was appended to $\mathcal{W}$ then
          break
  Let $\widetilde{W}$ denote the span of the vectors in $\mathcal{W}$.
  $\mathcal{L} \leftarrow$EnumerateNetworks($\widetilde{W}, B^{-L-1}2^{-\Omega(L)} \cdot \varepsilon/\sqrt{k}$).
  $N' \leftarrow \mathrm{poly}(B^{L+2}, k, 1/\varepsilon) \cdot \log(1/\delta)$.
  for $\widetilde{F} \in \mathcal{L}$ do
      Form an empirical estimate $\hat{\varepsilon}$ for $\|\widetilde{F} - F\|$ by drawing $N'$ samples.
      if $\hat{\varepsilon} \le 3\varepsilon$ then
          return $\widetilde{F}$.
```

as well as the population version of $\widetilde{\mathbf{M}}_{\mathsf{emp}}$, that is, $\widetilde{\mathbf{M}} \triangleq \mathbb{E}_{(x_1,y_1),\ldots,(x_N,y_N)}[\widetilde{\mathbf{M}}_{\mathsf{emp}}]$.

We will upper bound

$$\|\widetilde{\mathbf{M}}_{\mathsf{emp}} - \mathbf{M}\|_2 \le \|\widetilde{\mathbf{M}}_{\mathsf{emp}} - \widetilde{\mathbf{M}}\|_2 + \|\widetilde{\mathbf{M}} - \mathbf{M}'\|_2 + \|\mathbf{M}' - \mathbf{M}\|_2.$$

by upper bounding each of the summands on the right-hand side by ξ_*.

By Lemma 1.3.32 and our choice of N, $\|\widetilde{\mathbf{M}}_{\mathsf{emp}} - \widetilde{\mathbf{M}}\|_2 \le \xi_*$ with probability at least $1-\delta$.

To upper bound $\|\widetilde{\mathbf{M}} - \mathbf{M}'\|_2$, we can naively upper bound

$$\left\|\mathop{\mathbb{E}}_{x,y}\left[\mathbb{1}[|y - F|_W(x)| > \tau] \cdot (xx^\top - \mathrm{Id})\right]\right\| \le 2,$$

so by Claim 1.3.11 and Lemma 1.3.9 we have

$$\|\widetilde{\mathbf{M}} - \mathbf{M}'\|_2 \le 2\sqrt{2} \cdot d_C(\widetilde{W}, W) \le 4\sqrt{\nu \cdot k} \le \xi_*$$

Finally, we upper bound $\|\mathbf{M}' - \mathbf{M}\|_2$. For any test vector $v \in \mathbb{S}^{d-1}$ orthogonal to W,

$$\begin{aligned} v^\top(\mathbf{M} - \mathbf{M}')v &= \mathop{\mathbb{E}}_{x}\left[\left(\mathbb{1}[|y - F|_W(x)| > \tau] - \mathbb{1}[|y - \widetilde{F}(\Pi_{\widetilde{W}} x)| > \tau]\right) \cdot (\langle v, x\rangle^2 - 1)\right] \\ &\le \Pr_x\left[\mathrm{sgn}(|y - F|_W(x)| - \tau) \neq \mathrm{sgn}(|y - \widetilde{F}(\Pi_{\widetilde{W}} x)| - \tau)\right]^{1-1/k} \cdot O(k) \\ &\le O\left(k\left(\frac{C_*\sqrt{\nu}M^2}{\tau}\right)^{1-1/k}\right) = O\left(k\left(\frac{C_*\sqrt{\nu}M^2}{c\sqrt{k}\Lambda}\right)^{1-1/k}\right) \le \xi_* \end{aligned}$$

where the second step follows by Holder's and the fact that $\mathbb{E}_{g\sim\mathcal{N}(0,1)}[(g^2-1)^k]^{1/k} \le O(k)$, and the third step follows by Lemma 3.4.6, which we may apply because $\widetilde{F}$ and $F|_W$ are $(M, 4\sqrt{\nu}\cdot \ell\Lambda)$-structurally-close. □

Finally, we use the above perturbation bound to show that in a single iteration of the main outer loop of FilteredPCAv2, if there is some variance unexplained by the subspace $\widetilde{W}$ found so far (see (3.20)), then we will find another "good" direction orthogonal to $\widetilde{W}$ which is also approximately within the span of V. Note that this claim has two components:

completeness, i.e. in the list of candidate functions we have enumerated, there is *some* function for which the top singular vector of $\widetilde{\mathbf{M}}^{\widetilde{W}}_{\mathsf{emp}}$ in Step (11) is a good direction, and *soundness*, i.e. whatever direction is ultimately chosen in Step 15 of FilteredPCAv2 is a good direction.

Lemma 3.4.12. *Suppose F only satisfies Assumption 3 (resp. both Assumptions 3 and 4). Suppose $\nu \le \varepsilon^2/(4kC^2_{\mathsf{piecewise}})$ (resp. $\nu \le \varepsilon^2/(4kC^2_{\mathsf{network}})$). For $0 \le \ell < k$, let $\widetilde{w}_1,\dots,\widetilde{w}_\ell$ be a frame ν-nearly within V, with span $\widetilde{W}$. Define $\xi = \xi_{\mathsf{piecewise}}(\nu)$ (resp. $\xi = \xi_{\mathsf{network}}(\nu)$) according to (3.19), and suppose $N \ge \Omega(\{d \vee \log(1/\delta)\}/\xi^2)$ and $\tau = c\sqrt{k}\cdot\Lambda$.*

Suppose $\xi \le \underline{\lambda}/6$, and suppose

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,Id)}[(F(x) - F(\Pi_{\widetilde{W}}x))^2] \ge \varepsilon^2. \tag{3.20}$$

Let $\mathcal{L}$ be the output of EnumerateKickers*$(\widetilde{W}, 2\sqrt{\nu}\cdot\ell)$ (resp.* EnumerateNetworks*$(\widetilde{W}, 2\sqrt{\nu}\cdot \ell\sqrt{k}\cdot B)$). With probability at least $1 - |\mathcal{L}|\cdot\delta$ over the randomness of the N samples, the following hold:*

1. ***Completeness**: There exists some $\widetilde{F} \in \mathcal{L}$ such that, if $\widetilde{\mathbf{M}}^{\widetilde{W}}_{\mathsf{emp}}$ is defined according to Step 11 of* FilteredPCAv2*, its top singular value is at least $\underline{\lambda} - 3\xi$.*

2. ***Soundness**: For any $\widetilde{F} \in \mathcal{L}$ for which $\|\widetilde{\mathbf{M}}^{\widetilde{W}}_{\mathsf{emp}}\|_2 \ge \underline{\lambda} - 3\xi$, the top singular vector w satisfies $\|\Pi_V w\| \ge 1 - c'\xi^2/\underline{\lambda}^2$ for some absolute constant $c' > 0$ and is orthogonal to $\widetilde{W}$.*

Proof. When the choice of $\widetilde{F}$ is clear from context, for convenience we will denote $\mathbf{M}^W$ and $\widetilde{\mathbf{M}}^{\widetilde{W}}_{\mathsf{emp}}$ by $\mathbf{M}$ and $\widetilde{\mathbf{M}}_{\mathsf{emp}}$ respectively.

By Lemma 3.4.8 (resp. Lemma 3.4.10) and our assumed bound on ν, there exists $\widetilde{F}$ in the output of EnumerateKickers (resp. EnumerateNetworks) which is $(M, \varepsilon/2k)$-structurally-close to $F|_W$ for some ℓ-dimensional subspace $W \subsetneq V$.

By triangle inequality, Lemma 3.3.13, and (3.20), and our assumed bounds on ν, we have that $\|F - F|_W\| \ge \varepsilon/2$. So by Lemma 3.4.5 and (3.11), we know $\|\mathbf{M}\| \ge \underline{\lambda}$.

Because this $\widetilde{F}$ is $(M, C_{\mathsf{piecewise}}\sqrt{\nu})$-structurally-close (resp. $(M, C_{\mathsf{network}}\sqrt{\nu})$-structurally close) to $F|_W$, Lemma 3.4.11 implies that with probability $1-\delta$, $\|\mathbf{M} - \widetilde{\mathbf{M}}_{\mathsf{emp}}\|_2 \le 3\xi$, so

$\widetilde{\mathbf{M}}_{\mathsf{emp}}$ has top singular value at least $\underline{\lambda} - 3\xi$. This proves completeness.

Now take any $\widetilde{F}$ for which $\|\widetilde{\mathbf{M}}_{\mathsf{emp}}\|_2 \geq \underline{\lambda} - 3\xi$. The fact that the top singular vector w is orthogonal to $\widetilde{W}$ is immediate. And by Lemma 3.4.11, with probability $1 - \delta$ over the samples, $\|\mathbf{M} - \widetilde{\mathbf{M}}_{\mathsf{emp}}\|_2 \leq 3\xi$. So if we take $\lambda, \varepsilon, \mathbf{A}, \widehat{\mathbf{A}}$ in Corollary 1.3.8 to be $\underline{\lambda}$, 3ξ, $\mathbf{M}$, and $\widetilde{\mathbf{M}}_{\mathsf{emp}}$ respectively, then because $\xi \leq \underline{\lambda}/6$, we get that the top singular vector w of $\widetilde{\mathbf{M}}_{\mathsf{emp}}$ satisfies $\|\Pi_V w\| \geq 1 - O(\xi^2/\underline{\lambda}^2)$. This proves soundness, upon union bounding over all $\widetilde{F} \in \mathcal{L}$. □

3.4.7 Putting Everything Together

To conclude the proof of Theorems 3.4.1 and 3.4.2, we first show that for the subspace $\widetilde{W}$ formed in Step 18, if $\widetilde{W}$ is sufficiently close to the true relevant subspace V or if (3.20) is violated, then one can run EnumerateKickers (resp. EnumerateNetworks) one more time to produce a function with small squared error relative to F.

Lemma 3.4.13. *Suppose F only satisfies Assumption 3 (resp. both Assumptions 3 and 4). Define*

$$\varepsilon^* \triangleq \varepsilon/(2\sqrt{k}\Lambda) \qquad \textit{(resp. } \varepsilon^* \triangleq B^{-L-1}2^{-\Omega(L)} \cdot \varepsilon/\sqrt{k}\textit{)}$$

Let $\widetilde{w}_1, ..., \widetilde{w}_\ell$ be a frame with span $\widetilde{W}$. If either 1) $\ell = k$ and this frame is $\varepsilon^2/4kC^2_{\mathsf{piecewise}}$-nearly (resp. $\varepsilon^2/4kC^2_{\mathsf{network}}$-nearly) within V, or 2) inequality (3.20) is violated. Then the output $\mathcal{L}$ of EnumerateKickers*$(\widetilde{W}, \varepsilon^*)$ (resp.* EnumerateNetworks*$(\widetilde{W}, \varepsilon^*)$) contains a function $\widetilde{F}$ for which $\|F - \widetilde{F}\| \leq O(\varepsilon)$. Furthermore, $|\mathcal{L}| \leq M^{M^2} \cdot O(\Lambda/\varepsilon)^k$ (resp. $|\mathcal{L}| \leq O(B^{L+2}2^{O(L)}/\varepsilon)^{O(S^2)}$).*

In particular, if 1) or 2) holds for the subspace $\widetilde{W}$ at the end of running FilteredPCAv2*, then the output $\widetilde{F}$ of* FilteredPCAv2 *satisfies $\|F - \widetilde{F}\| \leq O(\varepsilon)$.*

Proof. We first show that if either 1) or 2) holds, then there exists $\widetilde{F}$ in $\mathcal{L}$ for which $\|\widetilde{F} - F\| \leq O(\varepsilon)$.

Suppose 1) holds. If F only satisfies Assumption 3 (resp. Assumptions 3 and 4), then by the final part of Lemma 3.4.8 (resp. Lemma 3.4.10), there is a function $\widetilde{F}$ in $\mathcal{L}$ which is $(M, \varepsilon/2k)$-structurally-close (resp. $(2^S, \varepsilon/2k)$-structurally-close) to $F|_W$ for ℓ-dimensional subspace $W \subseteq V$. Because $\ell = k$ when 1) holds, this subspace must be V, so in fact $F|_W = F$ and therefore $\widetilde{F}$ is structurally-close to F. By Lemma 3.3.13, we conclude that $\|\widetilde{F} - F\| \le \varepsilon$.

Suppose 2) holds. If F only satisfies Assumption 3 (resp. Assumptions 3 and 4), then we can take $\widetilde{F}^*$ in the first part of Lemma 3.4.8 (resp. Lemma 3.4.10) to be the function $x \mapsto F(\Pi_{\widetilde{W}} x)$, which is clearly also a Λ-Lipschitz kicker (resp. ReLU network of size S whose weight matrices have operator norm at most B) with relevant subspace $\widetilde{W}$. It follows that $\mathcal{L}$ contains some function $\widetilde{F}$ which is $(M, \varepsilon/(2\sqrt{k}))$- (resp. $(2^S, \varepsilon/(2\sqrt{k}))$-structurally-close to $\widetilde{F}^*$. By Lemma 3.3.13, we conclude that $\|\widetilde{F} - F\| \le 3\varepsilon/2$.

For the last part of the lemma, note that by Lemma 3.5.1 in Appendix 3.5.1 that for any function $\widetilde{F}$ for which $\|\widetilde{F} - F\|^2 \le \mu$, we can estimate $\|\widetilde{F} - F\|^2$ to error $O(\varepsilon^2)$ from $O((\mu + \Lambda^2 k)\log(1/\delta)/\varepsilon^4)$ samples (resp. $O((\mu + B^{2L+4}k)\log(1/\delta)/\varepsilon^2))$. Note that for any $\widetilde{F} \in \mathcal{L}$, by the second part of Lemma 3.3.13 we have that $\|\widetilde{F} - F\| \le O(\Lambda\sqrt{k})$ (resp. $\|\widetilde{F} - F\| \le O(B^{L+2}\sqrt{k})$). □

We can now conclude the proof of correctness for FilteredPCAv2.

Proof of Theorem 3.4.1. First note that the only randomness in FilteredPCAv2 comes from calling ApproxBlockSVD and drawing samples, so henceforth we will condition on the event that the former always succeeds and on the success of Lemma 1.3.32 for every batch of samples drawn in Step 7 of FilteredPCAv2. By our choice of parameters in FilteredPCAv2 and a union bound, this event happens with probability at least $1 - \delta$.

If F satisfies Assumption 3 only (resp. both Assumptions 3 and 4), let $\xi(\nu) = \xi_{\mathsf{piecewise}}(\nu)$ and $C_* = C_{\mathsf{piecewise}}$ (resp. $\xi(\nu) = \xi_{\mathsf{network}}(\nu)$ and $C_* = C_{\mathsf{network}}$), recalling the definition from (3.19).

Call $\nu \ge 0$ *admissible* if $\nu \le \varepsilon^2/4kC_*^2$ and $\xi(\nu) \le \underline{\lambda}/6$. Let $\iota : \mathbb{R} \to \mathbb{R}$ be the function given by $\iota(\nu) = c'\xi(\nu)^2/\underline{\lambda}^2$, where c' is the absolute constant in Lemma 3.4.12. Note that if we define

$$\beta \triangleq (c'/\underline{\lambda}^2) \cdot O\left(k^2 \cdot \left(\frac{C_*^2 M^4}{c^2 k \Lambda^2}\right)^{1-1/k}\right),$$

then $\iota(\nu) = (\beta \cdot \nu^{1-1/k}) \vee (k\nu)$.

Because we are conditioning on every invocation of ApproxBlockSVD succeeding, the quantity λ computed in Step 13 is certainly $\xi(\nu)/2$-close to the true top singular value of

$\widetilde{\mathbf{M}}^{\widetilde{W}}$. So Lemma 3.4.12 tells us that in any iteration ℓ of the main loop in FilteredPCAv2, if $\{\widetilde{w}_1, \ldots, \widetilde{w}_\ell\}$ is a frame ν-nearly within V for admissible ν, then either 1) we reach Line 15 in the inner loop and append some $\widetilde{w}_{\ell+1}$ for which $\{\widetilde{w}_1, \ldots, \widetilde{w}_{\ell+1}\}$ is a frame $\iota(\nu)$-nearly within V, or 2) (3.20) is violated, in which case condition 2) of Lemma 3.4.13 implies that FilteredPCAv2 would output a function $\widetilde{F}$ for which $\|F - \widetilde{F}\| \le O(\varepsilon)$.

So all we need to verify is that there is a choice of ν_0 for which the k numbers

$$\nu_0, \iota(\nu_0), \ldots, \underbrace{\iota(\iota(\cdots \iota}_{k-1}(\nu_0)\cdots)) \tag{3.21}$$

are all admissible, after which we can invoke condition 1) of Lemma 3.4.13 to conclude that FilteredPCAv2 outputs a function $\widetilde{F}$ for which $\|F - \widetilde{F}\| \le O(\varepsilon)$. It is clear that for ν sufficiently small, ι is increasing in ν. So it suffices to choose ν_0 sufficiently small that the last number in the sequence (3.21) is admissible.

Then the last number in (3.21) is at most

$$\left(\beta^{\sum_{j=0}^{k-1}(1-1/k)^j} \cdot \nu_0^{(1-1/k)^k}\right) \vee (k^k \nu_0) \le \left(\beta^k \cdot \nu_0^{1/e}\right) \vee (k^k \nu_0).$$

If F satisfies Assumption 3 only and we take $C_* = C_{\mathsf{piecewise}}$, then

$$\beta^k = (c'/\underline{\lambda}^2)^k \cdot O\left(k^{2k} \cdot (kM^4/c^2)^{k-1}\right),$$

so for

$$\nu_0 \triangleq \mathrm{poly}(k^k, 1/\underline{\lambda}^k, M^k, \Lambda/\varepsilon)^{-1} = \mathrm{poly}(e^{k^3\Lambda^2/\varepsilon^2}, M^k)^{-1}$$

sufficiently small, we have that $(\beta^k \cdot \nu_0^{1/e}) \vee (k^k \nu_0)$ is admissible.

And because in each of the at most k iterations of the main loop of FilteredPCAv2,

$$N = O(\{d \vee \log(Mk/\delta)\}/\xi(\nu_0)^2) \le d\log(1/\delta)\mathrm{poly}(e^{k^3\Lambda^2/\varepsilon^2}, M^k)$$

samples are drawn, the final sample complexity is $d\log(1/\delta)\mathrm{poly}(e^{k^3\Lambda^2/\varepsilon^2}, M^k)$ as claimed. The runtime is dominated by the at most $M^{M^2}O(1/\sqrt{\nu_0})^\ell = M^{M^2} \cdot \mathrm{poly}(e^{k^4\Lambda^2/\varepsilon^2}, M^{k^2})$ calls to ApproxBlockSVD, one for each element of $\mathcal{L}$ output by EnumerateKickers, (note that the runtime and sample complexity cost of running EnumerateKickers at the very end is of much lower order). As there is a matrix-vector oracle for the matrices on which we run ApproxBlockSVD which takes time $O(d^2)$, by Fact 1.3.6 each of these calls takes, up to lower order factors that will be absorbed elsewhere, $\widetilde{O}(d^2 \log(1/\delta))$ time, so we conclude that FilteredPCAv2 runs in time

$$\widetilde{O}(d^2 \log(1/\delta)) \cdot M^{M^2} \cdot \mathrm{poly}(e^{k^4\Lambda^2/\varepsilon^2}, M^{k^2})$$

as claimed.

If F satisfies Assumptions 3 and 4 and we take $C_* = C_{\mathsf{network}}$, then

$$\beta^k = (c'/\underline{\lambda}^2)^k \cdot O\left(k^{2k}\left(\frac{2^{O(L)}B^{2L+4}k2^{4S}}{c^2\Lambda^2}\right)^{k-1}\right),$$

where we have used that $M \le 2^S$ for size-S ReLU networks. So for

$$\nu_0 \triangleq \mathrm{poly}(k^k, 1/\underline{\lambda}^k, 2^{kS}, (B^{L+2}/\Lambda)^k, \Lambda/\varepsilon)^{-1} = \mathrm{poly}(e^{k^3\Lambda^2/\varepsilon^2}, 2^{kS}, (B^{L+2}/\Lambda)^k)$$

sufficiently small, we have that $(\beta^k \cdot \nu_0^{1/e}) \vee (k^k \nu_0)$ is admissible.

And because in each of the at most k iteration of the main loop of FilteredPCAv2,

$$N = O(\{d \vee \log(2^S k/\delta)\}/\xi(\nu_0)^2) \le d\log(1/\delta)\mathrm{poly}(e^{k^3\Lambda^2/\varepsilon^2}, 2^{kS}, B^{(L+2)k}/\Lambda^k)$$

samples are drawn, the final sample complexity is $d\log(1/\delta)\mathrm{poly}(e^{k^3\Lambda^2/\varepsilon^2}, 2^{kS}, B^{(L+2)k}/\Lambda^k)$ as claimed. The runtime is dominated by the at most $O(1/\sqrt{\nu_0})^{O(S^2)} = \mathrm{poly}(e^{k^3S^2\Lambda^2/\varepsilon^2}, 2^{kS^3}, B^{(L+2)kS^2}/\Lambda^{kS^2})$ calls to ApproxBlockSVD, one for each element of $\mathcal{L}$ output by EnumerateNetworks (note that the runtime and sample complexity cost of running EnumerateNetworks at the very end is of much lower order). Each of these calls takes, up to lower order factors that will be absorbed elsewhere, $\widetilde{O}(d^2 \log(1/\delta))$ time, so we conclude that FilteredPCAv2 runs in time

$$\widetilde{O}(d^2 \log(1/\delta)) \cdot \mathrm{poly}(e^{k^3S^2\Lambda^2/\varepsilon^2}, 2^{kS^3}, (B^{L+2}/\Lambda)^{kS^2})$$

as claimed. □

Remark 3.4.14 (Comparison to FILTEREDPCAV2 from the previous chapter)**.** *Here we briefly discuss what goes wrong if one simply tries mimicking the approach of [CM20]. Provided one has already recovered some (orthonormal) directions $w_1, ..., w_\ell$ spanning a subspace $W \subset V$, one would consider the matrix*

$$\mathbf{M}^W_{\mathsf{CM}} \triangleq \Pi_{W^\perp} \mathop{\mathbb{E}}_{x,y}\left[\mathbb{1}[|y| > \tau \wedge \|\Pi_W x\|^2 \le \alpha] \cdot (xx^\top - Id)\right]\Pi_{W^\perp}$$

for some $\alpha, \tau > 0$. The motivation for conditioning on $\|\Pi_W x\|^2 \le \alpha$ is that we now have

$$\langle \Pi_{V\backslash W}, \mathbf{M}^W_{\mathsf{CM}}\rangle = \mathop{\mathbb{E}}_{x,y}\left[\mathbb{1}[|y| > \tau \wedge \|\Pi_W x\|^2 \le \alpha] \cdot (\|\Pi_{V\backslash W} x\|^2 - (k-\ell))\right],$$

and if one could choose τ strictly greater than the supremum of $|F(x)|$ over all x for which $\|\Pi_W x\|^2 \le \alpha$ and $\|\Pi_{V\backslash W} x\|^2 \le 2(k-\ell)$, then we would conclude that

$$\langle \Pi_{V\backslash W}, \mathbf{M}^W_{\mathsf{CM}}\rangle \ge (k-\ell)\cdot \Pr[|y| > \tau \wedge \|\Pi_W x\| \le \alpha] \tag{3.22}$$

and it would suffice to lower bound the probability on the right-hand side of (3.22). This is precisely the route taken by [CM20] for learning low-degree polynomials, but in the case of ReLU networks, it is not hard to devise functions F for which the probability on the right-hand side of (3.22) is zero for such choices of τ, e.g. if $d = k = 2$, $\ell = 1$, $v_1 = e_1$, and

$$F(x) \triangleq \phi(x/\alpha + y) - \phi(-x/\alpha + y).$$

3.5 Appendix: Deferred Proofs

3.5.1 Concentration for Piecewise Linear Functions

Lemma 3.5.1. *For any $\delta > 0$ and any $t \le \Lambda^2 k$, the following holds. Let $F : \mathbb{R}^d \to \mathbb{R}$ be a Λ-Lipschitz kicker with relevant subspace V of dimension k. Then for samples $x_1, ..., x_N \sim \mathcal{N}(0, Id)$, where $N = \Theta\left((\mu + \Lambda^2 k)^2 \log(1/\delta)/t^2\right)$, the empirical estimate $\widehat{\sigma}^2 \triangleq \frac{1}{N}\sum_i F(x_i)^2$*

satisfies

$$\left|\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,Id)}[F(x)^2]-\widehat{\sigma}^2\right|\le t$$

with probability at least $1-\delta$.

Proof. As F is Λ-Lipschitz and continuous piecewise-linear, by Theorem 3.3.11 and Lemma 3.3.7 it has a lattice polynomial representation $\max_{j\in[m]}\min_{i\in\mathcal{I}_j}\langle u_i,\cdot\rangle$ for some clauses $\{\mathcal{I}_j\}$ and vectors $\{u_i\}$ for which $\|u_i\|\le\Lambda$. In particular, by Cauchy-Schwarz, $|F(x)|\le\Lambda\|x\|$ for all x. Now define the function $G(x)\triangleq F(x)^2-\mu$ where $\mu\triangleq\mathbb{E}_{x\sim\mathcal{N}(0,\mathrm{Id})}[F(x)^2]$. We can therefore naively upper bound the moments of G by

$$\mathbb{E}[|G|^t]^{1/t}\le\mu+\mathbb{E}[F^{2t}]^{1/t}\le\mu+\Lambda^2\cdot\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,\Pi_V)}[\|x\|^{2t}]^{1/t}\le\mu+O(\Lambda^2k)\cdot(t-1)$$

for all $t\ge 2$, where the last step follows by Corollary 1.3.17. Furthermore, $\mathbb{E}[|G|]\le 2\mu$. For $x\sim\mathcal{N}(0,\mathrm{Id})$, $G(x)$ is therefore a sub-exponential, mean-zero random variable with sub-exponential norm $K\triangleq O(\mu+\Lambda^2k)$, so by Fact 1.3.24 and the bound on t in the hypothesis, for $N=\Theta(K^2\log(1/\delta)/t^2)$, the claim follows. □

3.5.2 Representing Boolean Functions as ReLU Networks

Lemma 3.5.2. *For any function* $F:\{\pm1\}^n\to\{\pm1\}$, *there exists a set of weight matrices* $\mathbf{W}_0,\ldots,\mathbf{W}_{n-1}$ *for which* $F(x)=\mathbf{W}_{n-1}\phi(\mathbf{W}_{n-2}\phi(\cdots\phi(\mathbf{W}_0x)\cdots))$ *for all* $x\in\{\pm1\}^n$.

Proof. From the Fourier expansion of F as $F(x)=\sum_S\widehat{F}[S]\prod_{i\in S}x_i$, we see that it suffices to show how to represent any Fourier basis function $\prod_{i\in S}x_i$ with a ReLU network with depth n. We first show how to represent the function x_1x_2. Observe that for any $x_1,x_2\in\{\pm1\}$, we have that

$$x_1\cdot x_2=\phi(x_1+x_2)+\phi(-x_1-x_2)-\phi(x_2)-\phi(-x_2),\tag{3.23}$$

which is a two-layer neural network. Suppose inductively that for some $1\le m<n$, there exist weight matrices $\mathbf{W}'_0,\ldots,\mathbf{W}'_{m-1}$ for which $\prod_{i=1}^m x_i=\mathbf{W}'_{m-1}\phi(\mathbf{W}'_{m-2}\phi(\cdots\phi(\mathbf{W}'_0x)\cdots))$

for all $x \in \{\pm 1\}^n$. Then to compute $\prod_{i=1}^{m+1} x_i$, we can use (3.23) to conclude that

$$\prod_{i=1}^{m+1} x_i = \phi\left(\prod_{i=1}^{m} x_i + x_{m+1}\right) + \phi\left(-\prod_{i=1}^{m} x_i - x_{m+1}\right) - \phi(x_{m+1}) - \phi(-x_{m+1}).$$

It is clear that this can be represented as a ReLU network with depth $m+1$. □

Part II

Learning from Adversarially Corrupted Data

Chapter 4

Learning From Untrusted Batches With Sum-of-Squares

4.1 Introduction

In this chapter and the next, we consider the problem of learning from untrusted batches, originally introduced by Qiao and Valiant [QV17] and summarized in our Definition 1.2.8. Recall that the problem goes as follows:

(a) We are given m batches, consisting of k samples each. Furthermore the samples come from a discrete domain of size n. Each uncorrupted batch has the property that its samples were drawn i.i.d. from some distribution μ_i that is ω-close in total variation distance[1] to a distribution μ that is common to all the batches. Moreover a $1-\varepsilon$ fraction of the batches are uncorrupted.

(b) The remaining ε fraction of the batches are arbitrarily corrupted. In fact, an adversary is allowed to choose the contents of the corrupted batches after observing all of the uncorrupted batches.

The basic question is: *How well can we estimate μ in total variation distance?* As discussed in Section 1.2.2, the key features of this problem are designed to model some of the main

[1] The total variation distance between distributions p, q over a domain D is defined to be $\max_{S\subseteq D} p(S) - q(S)$

challenges in federated learning. In particular, we get batches of data from different users, but no batch is large enough by itself to learn an accurate model. In fact, the batches are generated from heterogenous sources because the ideal model for one user is often different than the ideal model for another. Additionally some of the batches are arbitrarily corrupted by an adversary who wishes to game our learning algorithm. In many applications, a non-trivial fraction of the data is supplied by malicious users. The meta question is: *Can we leverage information across the batches to learn an accurate model?*

In fact, the setup of learning with untrusted batches seems to model many other scenarios of interest. Our main focus will be settings where we have some additional structure or prior knowledge about the distributions we would like to learn. For example, suppose we want to estimate the demand curve across heterogenous groups. In particular, let $q_1 < q_2 < \cdots < q_n$ be a collection of increasing prices. Then set $\mu_{i,j}$ to be the probability that a random individual from group i would buy the product when offered a price q_j but not at the price q_{j+1}. We may not have enough data from each group to accurately estimate μ_i. Nevertheless we can hope to leverage data across the groups to estimate an aggregate curve μ that is a good approximation to each μ_i. Interestingly, the goal of being robust to an ε-fraction of the batches being corrupted now takes on a different meaning in this setting: We are asking whether we can estimate μ from data collected across the various groups in such a way that no ε-fraction of the groups can bias our estimates too much.

Qiao and Valiant [QV17] showed that it is possible to estimate μ within

$$O\left(\frac{\varepsilon}{\sqrt{k}} + \omega\right)$$

in total variation distance, from untrusted batches. Moreover they showed that this is the best possible up to constant factors. The somewhat surprising aspect of their bound is that it improves with larger k. This is a consequence of the “tensorization" property of the total variation distance which roughly says that the total variation distance between two distributions grows by at least a $\Omega(\sqrt{k})$ factor when we take k repetitions.

However, Qiao and Valiant [QV17] were only able to give an exponential time algorithm. Their approach was to estimate μ by estimating the total probability it assigns to every subset of the domain. Each of these subproblems is again a problem of learning with untrusted batches, but one on a discrete domain with just two elements. Qiao and Valiant [QV17] gave another algorithm, but one that requires $\omega = 0$ – i.e. each of the uncorrupted batches must be generated from the same underlying distribution. Their second algorithm was based on low-rank tensor approximation. They wrote down an order k tensor whose entries represent the probability of seeing any particular k tuple of samples as a batch, and showed that some slice of this tensor is an accurate estimate of μ. This algorithm also has the drawback that in order to estimate the entries of the tensor, you need n^k samples. In most applications, it would be infeasible to have so much data that you see essentially every possible batch. Their work left open the problem of getting efficient algorithms for learning with untrusted batches.

4.1.1 Our Results– Sum of Squares

In this chapter, we use the sum-of-squares hierarchy to design new algorithms for the problem of learning from untrusted batches (in the next chapter, we show how to get improved runtime and sample complexity guarantees using alternating minimization). An important feature of our approach in this chapter is that it is easy to incorporate additional prior information about the shape of the distribution into our sum-of-squares framework. But first, as a warm up, we will study the original learning with untrusted batches problem. We give a sequence of polynomial time algorithms whose estimation error approaches the information-theoretically optimal bound:

Theorem 4.1.1 (See Theorem 4.4.1 for formal statement)**.** *Fix any integer $t \geq 4$. There is a polynomial time algorithm to estimate μ to within*

$$O\left(\frac{\varepsilon^{1-1/t}}{\sqrt{k/t}} + \omega\right)$$

in total variation distance from m ε-corrupted batches, each of size k. Moreover the number of batches we need is polynomial in n.

This result improves over the 2^n time algorithm of Qiao and Valiant [QV17]. Note that the other algorithm of Qiao and Valiant [QV17] runs in time n^k but only works in the special case where $\omega = 0$ – i.e. all the uncorrupted batches come from the same underlying distribution. Moreover, in the above result, if we set $t = \log 1/\varepsilon$ then we get within a polylogarithmic factor of the optimal estimation error, but at the expense of running in quasipolynomial time:

Corollary 4.1.2. *There is an algorithm to estimate μ to within*

$$O\left(\frac{\varepsilon\sqrt{\log 1/\varepsilon}}{\sqrt{k}} + \omega\right)$$

in total variation distance from m ε-corrupted batches, each of size k. Moreover the running time and the number of batches we need are polynomial in $n^{\log 1/\varepsilon}$.

We note that in independent and concurrent work, [JO19] obtained a better guarantee for this same problem in that they managed to achieve polynomial rather than quasipolynomial time and sample complexity using an algorithm based on alternating minimization.

Finally, we come to what we believe to be our main contribution. In many applications, getting samples is expensive and we might only be able to afford a number of samples that is sublinear in the size of the domain. In such cases, it is important to utilize additional information such as prior knowledge about the shape of the distribution. Indeed, this is the case in the example we discussed earlier, where we often know that the distribution μ satisfies the monotone hazard rate condition. It is known that such distributions can be well-approximated by piecewise polynomial functions [CDSS13, CDSS14b, ADLS17].

In fact, the idea of imposing structure on the underlying distribution has a long and storied history in statistics and machine learning where it leads to better estimation rates and algorithms that use fewer samples [Bru55, Hil54, Weg70]. We ask: *Can prior information about the shape of a distribution be leveraged to get better algorithms for learning from untrusted batches?* Our main result is:

Theorem 4.1.3 (See Theorem 4.5.1 for formal statement)**.** *Fix any integer $t \geq 4$. If μ is approximated by an s-part piecewise polynomial function with degree at most d, there is a* *polynomial time algorithm to estimate μ to within*

$$O\left(\frac{\varepsilon^{1-1/t}}{\sqrt{k/t}} + \omega\right)$$

in total variation distance from m ε-corrupted batches, each of size k. Moreover the number of batches we need is polylogarithmic in n and polynomial in s and d.

While the problem of learning a piecewise polynomial distribution may not seem natural in applications, previous work of [CDSS13, CDSS14b, ADLS17] has demonstrated that this can be combined with results from approximation theory [Tim14] to achieve strong density estimation results for a large class of distribution families such as log-concave distributions, Gaussians, monotone distributions, monotone hazard rate distributions, Binomial distributions, Poisson distributions, and mixtures thereof [ADLS17].

In the next subsection, we describe our main techniques at a high level. The main takeaway is that the sum-of-squares hierarchy gives a seamless way to incorporate prior information about the structure into the estimation problem, which can lead to much better algorithms (in our case we are able to get sublinear sample complexity).

4.1.2 Our Techniques

Recently, there has been a flurry of progress in high-dimensional robust estimation [DKK+19a, LRV16, CSV17, DKK+17]. While the techniques seem to be quite different from each other – some relying on iterative filtering algorithms to remove outliers, and others relying on sum-of-squares proofs of identifiability – at their heart, they are about finding ways to reweight the empirical distribution on the observed samples in such a way that it has bounded moments along any one-dimensional projection [HL18, KSS18, DKS18b].

Our main observation is that algorithms for learning from untrusted batches can also be derived from this framework, but by working with a different family of test functions. When we consider moments of a one-dimensional projection, we are looking at test functions that are unit vectors (or tensor powers of them) in the ℓ_2-norm. In comparison, the exponential time algorithm of Qiao and Valiant [QV17] tries all ways of partitioning the domain into two sets. We can equivalently think about it as choosing a test vector (or tensor power of one) that has unit ℓ_∞-norm. In this way, we study the families of distributions for which we can find a sum-of-squares certificate that they have bounded moments with respect to unit ℓ_∞ test functions. We show that the multinomial distribution has this property, and using the proofs-to-algorithms methodology [HL18, KSS18], this gives our improved algorithm for the general problem of learning with untrusted batches.

The beauty of this common abstraction is that it flexibly allows us to build in other problem specific constraints, like shape constraints on μ. Here, classical results from VC theory [VC74, DL01] say that it suffices to learn the distribution in a weaker norm (see Definition 4.5.2) than total variation distance, which has fewer degrees of freedom. From our perspective, the change is that, in this case, instead of allowing all unit ℓ_∞ test functions, we only have to consider those which come from tensor powers of a vector that has a bounded number of sign changes. However, encoding this constraint in the sum-of-squares hierarchy is quite non-trivial, as it is not clear how to encode this combinatorial constraint within the algebraic language of the sum-of-squares proof system. To get around this, we demonstrate that we can relax the combinatorial constraint into a linear algebraic one, namely, sparsity in the *Haar wavelet basis*. We then exploit properties of the Haar wavelet basis to encode this constraint into our relaxation. *The main open question of our work is to push this philosophy further, and explore what other sorts of provably robust algorithms can be built out of different choices of test functions.*

4.1.3 Related Work

The problem of learning from untrusted batches was introduced by [QV17], and is motivated by problems in reliable distributed learning such as *federated learning* [MMR+17, KMY+16]. In the TCS community, the problem of learning from batches has been considered in a number of settings [LRR13, TKV17], but these results cannot tolerate noise in the data.

More generally, the question of univariate density estimation, and specifically, density estimation of structured distributions, has a vast literature and we cannot hope to fully survey it here. See [BBBB72] for a survey of classical results in the area. Many different natural structural assumptions have been considered in the statistics and learning theory communities, such as monotonicity [Gre56, Gro85, Bir87a, Bir87b, JW09], monotone hazard rate [CDSS13, CR14, HMR18], unimodality [Rao69, Weg70, Fou97], convexity and concavity [HP76, KM10], log-concavity [BRW09, DR09, Wal09], k-modality [CT04, BW07, GW09, BW10], smoothness [Bru58, KP92, DJKP95, KPT96, DJKP96, DJ98], and mixtures of structured distributions [RW84, TSM85, Lin95, Das99, DS00, AK01, VW02, FOS05, AM05, KMV10, MV10, DDS12b, DDS12a, DDO+13, DKS16a, DDKT16, DKS16b, DKS16c]. The reader is referred to [O'B16, Dia16] for a more extensive review of this vast literature. Recently it has been demonstrated that the classical piecewise polynomial (or spline) methods, see e.g. [WW83, Sto94, SHKT97, WN07], can be adapted to obtain general estimators for almost all of these problems with nearly-optimal sample complexity and runtime [CDSS13, CDSS14b, CDSS14a, ADH+15, ADLS17]. While these estimators are typically tolerant of worst-case noise, it is unclear how to adapt them to the batch setting, to obtain improved statistical rates.

Finally, our work is also related to a recent line of work on robust statistics [DKK+19a, LRV16, CSV17, DKK+17, HL18, KSS18], a classical problem dating back to the 60s and 70s [Ans60, Tuk60, Hub92, Tuk75]. See [Li18b, Ste18] for a more comprehensive survey of this line of work. We remark that the majority of this work focuses on estimation in ℓ_2-norm or Frobenius norm, with two notable exceptions: [BDLS17] uses learning in a sparsity-inducing norm to improve the sample complexity for sparse mean estimation, and [SCV18] gives an information-theoretic characterization of when mean estimation in general norms is possible, but they do not give efficient algorithms. Our techniques are most closely related to the sum-of-squares based algorithms of [HL18, KSS18], and this general technique has also found application in other robust learning problems such as robust regression [KKM18] and list-decodable regression [KKK19, RY19].

4.1.4 Organization

In Section 4.2, we provide a high-level overview of our techniques. In Section 4.3, we give notation, a formal description of the generative model, a recap of the key SoS tools needed, and show a sum-of-squares proof that multinomial distributions have bounded moments. In Section 4.4 we give a proof of Theorem 4.4.1. In Section 4.5, we give a proof of Theorem 4.5.1. The technical heart of this chapter is Section 4.6, where we fill in the details on how to efficiently encode key constraints from our SoS relaxations using matrix SoS. In Appendix 4.7, we provide proofs deferred from earlier sections.

4.2 High-Level Argument

In this section we give an overview of how we prove Theorems 4.4.1 and 4.5.1. The ideas required for the latter are a strict subset of those for the former, so we first describe the aspects common to both proofs before elaborating in Section 4.2.4 and 4.2.5 on techniques specific to Theorem 4.5.1, which we view as the main contribution of this chapter. As these latter sections are somewhat technical, readers new to the use of sum-of-squares for robust mean estimation may feel free to skip them on first reading, as the other sections will be sufficient for understanding the proof of Theorem 4.4.1.

4.2.1 Robust Mean Estimation

We first recast the problem of learning from untrusted batches as a generalization of the problem of robustly estimating the mean of a multinomial distribution in L_1 distance.

To the i-th batch of k samples $Y_i = (Y_i^1, ..., Y_i^k)$ from $[n]$ we may associate the vector of frequencies $X_i \in \Delta^n$ (where $\Delta^n \subset \mathbb{R}^n$ is the probability simplex) given by

$$(X_i)_j = \frac{1}{k}\sum_{\nu=1}^{k} \mathbb{1}[(Y_i)_\nu = j] \ \forall j \in [n].$$

If $Y_1, ..., Y_N$ are independent batches of k iid draws from $\mu_1, ..., \mu_N$ respectively, then $X_1, ..., X_N$ are independent draws from $\mathrm{Mul}_k(\mu_1), ..., \mathrm{Mul}_k(\mu_N)$ respectively, where $\mathrm{Mul}_k(\mu_i)$ is defined to be the normalized multinomial distribution given by k draws from μ_i. We can think of the learning algorithm as taking in vectors $X_1, ..., X_N \in \Delta^n$, such that a $(1-\varepsilon)N$-sized subset of them, indexed by $S_G \subset [N]$, are independent draws from $\mathrm{Mul}_k(\mu_j)$ for $j \in S_G$, and the remaining points are arbitrary vectors in Δ^n. The goal of the learning algorithm is to learn μ in L_1 distance. Note that when $\mu_i = \mu$ for all $i \in S_G$, this is precisely the problem of robustly estimating the mean μ of a (normalized) multinomial distribution.

For simplicity, we will assume that $\omega = 0$ for the rest of this subsection, i.e. that

$\mu_1 = \cdots = \mu_N$. Indeed, one appealing feature of our techniques is the ease with which one can extend the techniques we describe below to handle the case of nonzero ω.

4.2.2 Searching for a Moment-Bounded Subset

A recurring theme in the robust learning literature [DKK+19a,LRV16,HL18,KSS18,DKS18b] is that one can detect corruptions in the data by looking for anomalies in the empirical moments. In our setting, one useful feature of multinomial distributions $\text{Mul}_k(\mu)$ is that their moments up to degree k satisfy sub-Gaussian-type bounds.

Theorem 4.2.1 ([Lat97]). *For a (normalized) binomial random variable $Z \sim \frac{1}{k} \cdot Bin(k, p)$,*

$$\mathbb{E}[(Z - p)^t]^{1/t} \lesssim \sqrt{t/k}$$

for any even $t \leq k$.

Multinomial distributions inherit these same properties:

Lemma 4.2.2. *For any discrete distribution μ and any vector $v \in \{\pm 1\}^n$, if $X \sim Mul_k(\mu)$, then*

$$\mathbb{E}[\langle X - \mu, v\rangle^t]^{1/t} \lesssim \sqrt{t/k}$$

for any even $t \leq k$.

At a high level, our algorithms will search for a $(1 - \varepsilon)N$-sized subset S of the samples whose empirical moments satisfy these bounds, namely

$$\frac{1}{|S|} \sum_{i \in S} \langle X_i - \hat{\mu}, v\rangle^t \leq (8t/k)^{t/2} \quad \forall v \in \{\pm 1\}^n, \tag{4.1}$$

where $\hat{\mu} = \frac{1}{|S|} \sum_{i \in S} X_i$ is the empirical mean of S. This search problem can be reformulated as solving some system $\mathcal{P}$ of polynomial equalities and inequalities (see Section 4.4 for a formal specification). So if we could solve this system and argue that the empirical mean of *any* subset $S \subset [N]$ which satisfies the system is $O(\varepsilon/\sqrt{k})$-close in L_1 to μ, then we'd be done.

There are two complications to this approach:

(A) The problem of solving polynomial systems is NP-hard in general.

(B) Constraint (4.1) is a collection of exponentially many constraints.

By now it is well-understood how to circumvent issues like (A): use the sum-of-squares (SoS) hierarchy to relax the problem of searching for a single solution to $\mathcal{P}$, or even a *distribution* over solutions, to the problem of searching for a *pseudodistribution* over solutions. We will give formal definitions in Section 1.3.8, but roughly speaking, a pseudodistribution satisfying $\mathcal{P}$ is a linear functional that is indistinguishable from a distribution when evaluated on low-degree polynomials arising from the polynomials in $\mathcal{P}$.

The key point then is that if one can write down a "simple" proof that any solution to $\mathcal{P}$ has empirical mean close to μ, i.e. a proof using only low-degree polynomials arising from the polynomials in $\mathcal{P}$,[2] then the following learning algorithm will succeed:

(1) Solve an SDP to find a pseudodistribution $\tilde{\mathbb{E}}$ satisfying $\mathcal{P}$ in polynomial time.

(2) Extract from $\tilde{\mathbb{E}}$ an estimate for μ.[3]

We remark that this methodology of extracting SoS algorithms from simple proofs of identifiability has been used extensively in many recent works; we refer the reader to [RSS18] for a comprehensive overview.

4.2.3 Quantifying over $\{\pm 1\}^n$ via Matrix SoS

We now show how to address issue (B) above. The key is to design a smaller system of polynomial constraints which imply each of the exponentially many constraints in (4.1) under the SoS proof system, that is to say, we should be able to derive all of the constraints in (4.1) from the constraints in the smaller system, using only "low-degree" steps like Cauchy-Schwarz and Holder's. We remark that although the trick we will describe for doing this has appeared previously in the literature under the name of "matrix SoS proofs" [HL18],

[2] Practically speaking, for a proof to be "simple" in the above sense effectively means that the steps in the proof involve nothing more than applications of Cauchy-Schwarz and Holder's inequalities and avoid use of concentration and union bounds.

[3] We are glossing over this second step, but it turns out that a naive rounding scheme suffices (see Section 4.4.4).

we believe a complete but informal treatment of this technique will help the reader better appreciate the subtleties in how we extend this approach to obtain Theorem 4.5.1.

To describe the trick, we first abstract out the more problem-specific details of the polynomial systems we will consider. Say we wish to encode the following exponentially large program with a smaller polynomial system.

Program $\mathcal{Q}$. *The variables consist of $\{Z_{\alpha,\beta}\}$ for all multisets $\alpha, \beta \subseteq [n]$ of size $t/2$, as well as some other variables $x_1, ..., x_M$. The constraints include $\{p_1(x, Z) \geq 0, ..., p_m(x, Z) \geq 0, q_1(x, Z) = 0, ..., q_m(x, Z) = 0\}$ as well as the constraint*

$$\langle Z, v^{\otimes t/2}(v^{\otimes t/2})\rangle \leq 1 \quad \forall v \in \{\pm 1\}^n. \tag{4.2}$$

Suppose we know that Program $\mathcal{Q}$ has a satisfying assignment (Z^*, x^*) to its variables—in the systems we will actually work with, the existence of a satisfying assignment will be immediate, e.g. the set of all uncorrupted points is a satisfying assignment to the program sketched in Section 4.2.2.

Remark 4.2.3. *While the meaning of Z will be irrelevant to the proceeding discussion, the reader might find it helpful to think of Z^*, up to scaling, as the matrix $\mathbf{Z}[S_G]$ defined by:*

$$\mathbf{Z}[S] \triangleq \frac{1}{|S|}\sum_{i\in S}\left[(X_i - \mu_i)^{\otimes t/2}\right]^\top\left[(X_i - \mu_i)^{\otimes t/2}\right] - \frac{1}{|S|}\sum_{i\in S}\mathbb{E}_{X\sim\mathcal{D}_i}\left[(X - \mu_i)^{\otimes t/2}\right]^\top\left[(X - \mu_i)^{\otimes t/2}\right]. \tag{4.3}$$

The reason is that via the identity

$$\langle \mathbf{V}[S], v^{\otimes t/2}(v^{\otimes t/2})^\top = \frac{1}{|S|}\sum_{i\in S}\langle X_i - \mu_i, v\rangle^t - \frac{1}{|S|}\sum_{i\in S}\mathbb{E}_{X\sim\mathcal{D}_i}\langle X - \mu_i, v\rangle^t,$$

$\mathbf{Z}[S_G]$ gives a succinct way of describing the deviation of the empirical moments of the subset S from the true moments.

Returning to the task at hand, we would like to write down an auxiliary program $\widehat{\mathcal{Q}}$ which satisfies three criteria, namely that $\widehat{\mathcal{Q}}$

(a) has polynomially many variables and constraints

(b) implies Program $\mathcal{Q}$ under the SoS proof system, and

(c) is satisfiable.

In this case, we would be done: we could simply solve an SDP to find a pseudodistribution $\tilde{\mathbb{E}}$ satisfying $\widehat{\mathcal{Q}}$ and round it. Because of (c) we know our SDP solver will return something, because of (a) we know it will do so in polynomial time, and because of (b) $\tilde{\mathbb{E}}$ enjoys all the same properties that a pseudodistribution satisfying Program $\mathcal{Q}$ would.

To see how to design such an auxiliary program $\widehat{\mathcal{Q}}$, let us suppose further that the satisfying assignment (Z^*, x^*) for Program $\mathcal{Q}$ satisfies the property that (4.2) holds *as a polynomial inequality in* v. Specifically, if we had formal variables $v_1, ..., v_n$, suppose that one knew the existence of a proof, starting with just the polynomial equations $\{v_1^2 = 1, ..., v_n^2 = 1\}$ cutting out the Boolean hypercube, that the inequality $\langle Z^*, v^{\otimes t/2}(v^{\otimes t/2})^\top\rangle \le 1$ held, where we now view this inequality as a polynomial equation solely in the variables $v_1, ..., v_n$, with coefficients specified by the fixed choice of Z^*.

Showing this last assumption holds in the settings we consider will be nontrivial, but assuming for now that it does, the final idea needed to write down $\hat{\mathcal{Q}}$ is the following. Instead of searching for Z^* satisfying the exponentially large collection of constraints (4.2), we can search for Z^* for which the abovementioned SoS proof of $\langle Z^*, v^{\otimes t/2}(v^{\otimes t/2})^\top\rangle \le 1$ exists. The key point is that this search problem can be encoded in a much smaller polynomial system.

In particular, as will be evident once we give formal definitions of SoS proofs, the existence of such an SoS proof is equivalent to satisfiability of some new polynomial constraints in Z^* and some auxiliary variables corresponding to the steps of the SoS proof. To form $\hat{\mathcal{Q}}$, we will introduce these auxiliary variables and replace constraint (4.2) with these new polynomial constraints. The reason this general approach is called "matrix SoS" is that these new variables will be matrix-valued, and these new constraints will be inequalities between matrix-valued polynomials. The full details of this approach are provided in Section 4.6.1.

4.2.4 VC Meets Sum-of-Squares

Next, we describe the ideas that go into proving Theorem 4.5.1. The first is that when μ is (η, s)-piecewise degree-d, to learn μ in total variation distance, it is enough to learn μ in a much weaker norm which we will denote by $\|\cdot\|_{\mathcal{A}_K}$, where K is a parameter that depends on s and d. This insight was the workhorse behind state-of-the-art density estimation algorithms for various structured univariate distribution classes [Dia16, ADLS17, LS17]. In our setting, the main point is that if we have an estimate $\tilde{\mu}$ for μ for which $\|\tilde{\mu} - \mu\|_{\mathcal{A}_K} \le \zeta$, then by a result of [ADLS17], we can refine $\tilde{\mu}$ to get an estimate μ^* for which $d_{\mathrm{TV}}(\mu, \mu^*) \le O(\zeta + \eta)$ efficiently. We review the details for this in a self-contained manner in Section 4.5.1.

The algorithm of [ADLS17] will form an important part of the boilerplate for our learning algorithm, but the key difficulty will be to actually find $\tilde{\mu}$ which is close to μ in this weaker norm. We defer definitions to Section 4.5.1, but informally, $\|\mu - \tilde{\mu}\|_{\mathcal{A}_K}$ is small if and only if $\langle \mu - \tilde{\mu}, v\rangle$ is small for all $v \in \mathcal{V}^n_K \subset \{\pm 1\}^n$, where $\mathcal{V}^n_K$ is the set of all $v \in \{\pm 1\}^n$ with at most K sign changes when read as a vector from left to right (for example, $(1, 1, -1, -1, 1, 1, 1) \in \mathcal{V}^7_2$).

The natural approach to do this would be to search for a $(1 - \varepsilon)N$-sized subset S of the samples whose empirical moments satisfy

$$\frac{1}{|S|}\sum_{i \in S} \langle X_i - \hat{\mu}, v\rangle^t \le (8t/k)^{t/2} \quad \forall v \in \mathcal{V}^n_K. \tag{4.4}$$

Roughly, the sample complexity savings would then come from the fact that the empirical moments will concentrate in much fewer samples because the set of directions we need to union bound over is much smaller.

Of course, if $K = O(1)$, we could afford to simply write down all $\mathrm{poly}(n)$ constraints in (4.4). For typical applications of piecewise polynomial approximations though, K has a logarithmic dependence on n, so our main challenge is to obtain runtimes that do not depend exponentially on K. In particular, just as we will use matrix SoS to succinctly encode (4.1) for Theorem 4.4.1, we will use matrix SoS to succinctly encode (4.4) for Theorem 4.5.1. Next, we discuss some of the subtleties that arise in this encoding.

4.2.5 Quantifying over $\mathcal{V}^n_K$

As in Section 4.2.3, we will abstract out the problem-specific details and focus on finding an encoding for the following program:

Program $\mathcal{Q}'$. *The variables consist of $\{Z_{\alpha,\beta}\}$ for all multisets $\alpha, \beta \subseteq [n]$ of size $t/2$, as well as some other variables $x_1, ..., x_M$. The constraints include $\{p_1(x, Z) \geq 0, ..., p_m(x, Z) \geq 0, q_1(x, Z) = 0, ..., q_m(x, Z) = 0\}$ as well as the constraint*

$$\langle Z, v^{\otimes t/2}(v^{\otimes t/2})^\top \rangle \leq 1 \quad \forall v \in \mathcal{V}_K^n.$$

The primary stumbling block is that, unlike the Boolean hypercube, $\mathcal{V}_K^n$ is not cut out by a small number of polynomial relations. Indeed, conventional wisdom says that the sum-of-squares hierarchy is ill-suited to capturing combinatorial constraints like the ones defining $\mathcal{V}_K^n$.

The first observation is that there is an alternative orthonormal basis, the *Haar wavelet basis*, under which we can express any $v \in \mathcal{V}_K^n$ as a vector with a small number $s = \tilde{O}(K)$ of nonzero entries. One issue with this is that L_0 sparsity cannot be captured by a small number of polynomial constraints, but we could try relaxing this to L_1 sparsity and attempt to derive an SoS proof of (4.4) out of the L_1 constraint.

Specifically, one could try to argue that any pseudodistribution $\tilde{\mathbb{E}}$ over the formal variables $v_1, ..., v_n, \mathbf{W}_1, ..., \mathbf{W}_n$ satisfying the inequalities

(a) $v_i^2 = 1$ for all $i \in [n]$.

(b) $-\mathbf{W}_i \leq (Hv)_i \leq \mathbf{W}_i$ for all $i \in [n]$.

(c) $\sum_i \mathbf{W}_i \leq s$.

must satisfy

$$\tilde{\mathbb{E}}\left[\langle Z^*, v^{\otimes t/2}(v^{\otimes t/2})^\top \rangle\right] \leq 1, \tag{4.5}$$

where Z^* is a constant, fixed to a satisfying assignment to $\mathcal{Q}$. Note that (4.5) can be rewritten as

$$\left\langle Z^*, \tilde{\mathbb{E}}\left[v^{\otimes t/2}(v^{\otimes t/2})^\top\right] \right\rangle \leq 1,$$

and one can check (Lemma 4.7.1) that the set of all $n^{t/2} \times n^{t/2}$ matrices of the form $\tilde{\mathbb{E}}\left[v^{\otimes t/2}(v^{\otimes t/2})^\top\right]$ for $\tilde{\mathbb{E}}$ satisfying the three inequalities above is contained in the convex set $\mathcal{K}$ of all matrices whose Haar transforms are $L_{1,1}$-norm bounded[4] by s^t and Frobenius norm bounded by $n^{t/2}$.

At this point it will be useful to instantiate all of this in the setting of this chapter. Thinking of Z^*, up to scaling, as $\mathbf{Z}[S_G]$ as defined in (4.3), we need to ensure that its inner product with any matrix from $\mathcal{K}$ is at most one. The matrix $\mathbf{Z}[S_G]$ depends on the uncorrupted samples N, so at this point we are merely tasked with proving some large deviation bound (where "proof" now is in the literal, non-SoS sense).

We expect this to hold with high probability for N sublinear in n because the covering number of $\mathcal{K}$ should be much smaller than that of the set of all matrices with Frobenius norm bounded by n^t. As covering number bounds can be quite subtle, we opt instead for a shelling argument. Specifically, we can show that any element M with bounded $L_{1,1}$ and Frobenius norms can be written as a sum of s^t-sparse matrices whose Frobenius norms sum to at most $\|M\|_F$ (see Lemma 4.6.8 and its consequences in Section 4.6.3 and Appendix 4.7), reducing the task of building a net over $\mathcal{K}$ to building a net $\mathcal{N}$ over s^t-sparse matrices of Frobenius norm bounded by $n^{t/2}$.

The final and perhaps most important subtlety that arises is that as stated, this argument cannot achieve sublinear sample complexity because *the inverse Haar transform of an s^t-sparse matrix with Frobenius norm $n^{t/2}$ may have large max-norm*, which would preclude the sorts of univariate concentration bounds one would hope to apply on each direction in $\mathcal{N}$. More concretely, the issue is that ultimately, the net $\mathcal{N}$ over s^t-sparse matrices of bounded Frobenius norm corresponds to a net $\mathcal{N}'$ over $\mathcal{K}$ given by the inverse Haar transform of all elements of $\mathcal{N}$. And we would need to show that for any given $M \in \mathcal{N}'$, $\langle \mathbf{Z}[S_G], M\rangle$ is at most one with high probability. But if we have no control over the scaling of the max-norm of these M's, this is evidently impossible.

The workaround for this subtlety requires modifying the three inequalities used above, as well as the definition of $\mathcal{K}$, by incorporating properties of the Haar wavelet basis beyond just the fact that vectors from $\mathcal{V}^n_K$ are sparse in this basis. Roughly speaking, the key is to exploit the inherent multi-scale nature of the Haar wavelet basis.

This is best understood with an example. Instead of matrices, we will work with vectors

[4] The $L_{1,1}$ norm of a matrix is defined to be the sum of the absolute values of its entries.

(the reader can think of this as the "$t = 1$" case). In the following example, we will first try to convey 1) that there exist sparse vectors with L_2 norm $\sqrt{n}$ but whose inverse Haar transforms are as large as $\sqrt{n/2}$ in L_∞ norm. To reiterate, this is an issue because any $w \in \mathbb{R}^n$ which is a Haar transform of some vector $v \in \{\pm 1\}^n$ with few sign changes is sparse and has L_2 norm $\sqrt{n}$, yet the inverse Haar transform of w, i.e. v itself, has L_∞ norm 1. In other words, simply relaxing the set of $v \in \{\pm 1\}^n$ to the set of all vectors whose Haar transforms are sparse introduces problematic new vectors with substantially different properties than the vectors v. We will then 2) give a flavor of how we circumvent this crucial subtlety.

Example 4.2.4. *Let $n = 2^m$. The Haar wavelet basis for $\mathbb{R}^n$ contains the vector*

$$\psi_\ell \triangleq \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 0, ..., 0\right).$$

Say this is the ℓ-th vector in the basis. Then the vector w which has ℓ-th entry equal to $\sqrt{n}$ and all other entries 0 is clearly sparse and has L_2 norm $\sqrt{n}$. But its inverse Haar transform is

$$\left(\sqrt{n/2}, -\sqrt{n/2}, 0, 0, ..., 0\right),$$

which has largest entry $\sqrt{n/2}$, whereas obviously any $v \in \{\pm 1\}^n$ has largest entry 1.

One reason this example is not so bad is that if we express any $v \in \{\pm 1\}^n$ as a linear combination of Haar wavelets, the coefficient for the ℓ-th Haar wavelet, by orthonormality of the Haar wavelet basis, is $\langle v, \psi_\ell \rangle \le \sqrt{2}$. That is, the Haar transform of any such v has ℓ-th entry at most $\sqrt{2}$. So if we added to the collection of constraints defining $\mathcal{K}$ this additional constraint, we would already get rid of some problematic vectors like w.

More generally, problematic vectors like w in Example 4.2.4 exist at every "level" of the Haar wavelet basis, and it will be necessary to handle each of these levels appropriately. We defer the details to Lemma 4.6.6 and its consequences in Sections 4.6.3 and Appendix 4.7.

4.3 Technical Preliminaries

4.3.1 Miscellaneous Notation

- Given polynomials $p, q_1, ..., q_m$ in formal variables $x_1, ..., x_n$, we say that p is in the ideal generated by $q_1, ..., q_m$ at degree d if there exist polynomials $\{s_i\}_{i\in[m]}$ for which $q(x) = \sum_{i=1}^m s_i(x)q_i(x)$ where each $s_i(x)q_i(x)$ is of degree at most d.

- Recall the definition of the flattened tensor from (4.3). For any $S \subseteq [N]$,

$$\mathbf{Z}[S] \triangleq \frac{1}{|S|}\sum_{i\in S}\left[(X_i-\mu_i)^{\otimes t/2}\right]^\top\left[(X_i-\mu_i)^{\otimes t/2}\right] - \frac{1}{|S|}\sum_{i\in S}\mathbb{E}_{X\sim\mathcal{D}_i}\left[(X-\mu_i)^{\otimes t/2}\right]^\top\left[(X-\mu_i)^{\otimes t/2}\right].$$

- Given matrix $\mathbf{M}$, denote by $\|\mathbf{M}\|_{1,1}$ the sum of the absolute values of its entries.

- Given $p \in [0,1]$, let $\mathrm{Bin}(k,p)$ denote the *normalized* binomial distribution, which takes values in $\{0, 1/k, \cdots, 1\}$ rather than $\{0, 1, \cdots, k\}$.

- Given $\mu \in \Delta^n$, let $\mathrm{Mul}_k(\mu)$ denote the distribution over Δ^n given by sampling a frequency vector from the multinomial distribution arising from k draws from the distribution over $[n]$ specified by μ, and dividing by k. For example, when $n = 2$ and $\mu = (p, 1-p)$, $\mathrm{Mul}_k(\mu) = \mathrm{Bin}(k,p)$.

4.3.2 The Generative Model

Throughout the rest of the paper, let $\varepsilon, \omega > 0$, $n, k, N \in \mathbb{N}$, and let $\mu \in \Delta^n$ be some probability distribution over $[n]$. We restate the formal setting for learning from untrusted batches, originally introduced in Definition 1.2.8, with slightly different notation and terminology.

Definition 4.3.1. *We say $Y_1, ..., Y_N$ is an* ε-corrupted ω-diverse set of N batches of size k from μ *if they are generated via the following process:*

- *For every $i \in [(1-\varepsilon)N]$, $\tilde{Y}_i = (\tilde{Y}_i^1, ..., \tilde{Y}_i^k)$ is a set of k iid draws from μ_i, where $\mu_i \in \Delta^n$ is some probability distribution over $[n]$ for which $d_{TV}(\mu, \mu_i) \le \omega$.*

- *A computationally unbounded adversary inspects $\tilde{Y}_1, ..., \tilde{Y}_{(1-\varepsilon)N}$ and adds εN arbitrarily chosen tuples $\tilde{Y}_{(1-\varepsilon)N+1}, ..., \tilde{Y}_N \in [n]^k$, and returns the entire collection of tuples in any arbitrary order as $Y_1, ..., Y_N$.*

Let $S_G, S_B \subset [N]$ denote the indices of the uncorrupted (good) and corrupted (bad) batches.

It turns out that we might as well treat each Y_i as an unordered tuple. That is, for any Y_i, define $X_i \in \Delta^n$ to be the vector of frequencies whose a-th entry is $\frac{1}{k}\sum_{j=1}^k \mathbb{1}[Y_i^j = a]$ for all $a \in [n]$. Then for each, $i \in S_G$, X_i is an independent draw from $\mathrm{Mul}_k(\mu_i)$. Henceforth, we will work solely with this frequency vector perspective.

4.3.3 Certifiably Bounded Distributions

Recall from Section 4.2.3 that a prerequisite for the "matrix SoS" approach to work is that the exponentially large program from Section 4.2.2 must have a satisfying assignment for which there exists an SoS proof of the requisite empirical moment bounds (4.1) using the axioms $\{v_i^2 = 1 \ \forall i \in [n]\}$. A necessary condition for this to hold is for there to be an SoS proof from these axioms that the true moments of μ itself satisfy these same bounds. Again, we emphasize that these bounds should be regarded as polynomial inequalities solely in the variables $v_1, ..., v_n$.

Here we formalize what we mean by the existence of such a proof.

Definition 4.3.2. *A distribution $\mathcal{D}$ over $\mathbb{R}^d$ with mean μ is (t, ∞)-explicitly bounded with variance proxy σ if for every even $2 \le s \le t$:*

$$\{v_i^2 = 1 \ \forall i \in [n]\} \vdash_s \mathbb{E}_{X\sim\mathcal{D}}[\langle X - \mu, v\rangle^s] \le (\sigma s)^{s/2} \tag{4.6}$$

We remark that while a consequence of Theorem 4.2.1, due to [Lat97], is that the moments of any multinomial distribution satisfy these bounds, the proof in that work uses exponentials and is thus not an SoS proof without additional modifications to the argument. Here we give an SoS proof, at the cost of less desirable constants than those of [Lat97]. To our knowledge, this SoS proof is new.

Lemma 4.3.3. *Let $\mathcal{D} = Mul_k(\mu)$ for any $\mu \in \Delta^n$. Then $\mathcal{D}$ is (k, ∞)-explicitly bounded with variance proxy $8/k$.*

Proof. It is enough to show (4.6) for v for which $\|v\|_\infty = 1$. By definition $\mu = \mathbb{E}_{X\sim\mathcal{D}}[X]$, so we may symmetrize as follows:

$$\begin{aligned}\vdash_s \mathbb{E}_{X\sim\mathcal{D}}[\langle X - \mu, v\rangle^s] &= \mathbb{E}_{X\sim\mathcal{D}}[\langle X - \mathbb{E}_{X'\sim\mathcal{D}}[X'], v\rangle^s] \\ &\le \mathbb{E}_{X,X'\sim\mathcal{D}}[\langle X - X', v\rangle^s],\end{aligned}$$

where the inequality follows from SoS Cauchy-Schwarz. But note that the random variable $\langle X, v\rangle$ is the average of k independent copies of the random variable which takes on value v_i with probability μ_i for every $i \in [n]$. So define Z to be the symmetric random variable which takes on value $(v_i - v_{i'})$ with probability $\mu_i\mu_{i'}$ for every $(i, i') \in [n] \times [n]$. Then for $Z_1, ..., Z_k$ independent copies of Z,

$$\langle X - X', v\rangle \overset{d}{=} \frac{1}{k}\sum Z_i$$

We conclude that for any $1 \le s \le k$,

$$\begin{aligned}\vdash_s \mathbb{E}_{X\sim\mathcal{D}}[\langle X - \mu, v\rangle^s] &\le \frac{1}{k^s}\mathbb{E}[(Z_1 + \cdots + Z_k)^s] \\ &= \frac{1}{k^s}\sum_{\beta:|\beta|=s}\binom{s}{\beta_1, ..., \beta_k}\mathbb{E}[Z_\beta] && (4.7) \\ &= \frac{1}{k^s}\sum_{\substack{\beta:|\beta|=s \\ \beta_i \text{ even } \forall 1\le i\le k}}\binom{s}{\beta_1, ..., \beta_k}\mathbb{E}[Z_\beta] && (4.8) \\ &\le \frac{1}{k^s}(2sk)^{s/2}\cdot\max_\beta \mathbb{E}[Z_\beta] && (4.9) \\ &\le (2s/k)^{s/2}\cdot\max_\beta\prod_{i=1}^k \mathbb{E}[Z_i^{\beta_i}] && (4.10) \\ &\le (8s/k)^{s/2}, && (4.11)\end{aligned}$$

where the sum in (4.7) ranges over all monomials β of total degree s, that is, all tuples $\beta \in [s]^k$ for which $\sum_{i=1}^k \beta_i = s$. Equation (4.8) follows from the fact that $\mathbb{E}[Z_\beta] = \prod_{i=1}^k \mathbb{E}[Z_i^{\beta_i}]$ by independence, and $\mathbb{E}[Z_i^d] = 0$ for any odd d because Z is symmetric. For equation (4.9), note that by balls-and-bins, there are $\binom{s/2+k-1}{s/2} \leq \left(\frac{3ek}{s}\right)^{s/2}$ choices of β, and $\binom{s}{\beta_1,...,\beta_k} \leq s! \leq s^{s+1/2}e^{-s+1}$, and we may crudely bound the product of these quantities as

$$(3ek/s)^{s/2} \cdot s^{s+1/2}e^{-s+1} \leq (2sk)^{s/2}.$$

Equation (4.10) follows by independence, and for (4.11) we need that for every even $2 \leq d \leq s$, there is a degree-s SoS proof that $\mathbb{E}[Z^d] \leq 2^d$. But by Fact 1.3.44, $\{v_i^2 = 1 \ \forall i \in [n]\} \vdash_2 -2 \leq v_i - v_{i'} \leq 2$, from which there is a degree-d proof that $(v_i - v_{i'})^d \leq 2^d$. So

$$\{v_i^2 = 1 \ \forall i \in [n]\} \vdash_d \mathbb{E}[Z^d] = \sum_{i,i'} p_i p_{i'} (v_i - v_{i'})^d \leq 2^d \sum_{i,i'} p_i p_{i'} = 2^d$$

as claimed. □

4.4 Efficiently Learning from Untrusted Batches

In this section we prove our result on the general problem of learning from untrusted batches.

Theorem 4.4.1. *Let $t \geq 4$ be any integer. There is an algorithm that draws an ε-corrupted set of N ω-diverse batches of size k from μ for $N \geq \omega^{-2}\varepsilon^{-2}n^{O(t)} \cdot k^t/t^{t-1}$, runs in time $\omega^{-2t}\varepsilon^{-2t}n^{O(t^2)} \cdot k^{t^2}/t^{t(t-1)}$, and with probability $1 - 1/\mathrm{poly}(n)$ outputs a distribution $\hat{\mu}$ for which $d_{TV}(\mu, \hat{\mu}) \leq O(\omega + \varepsilon^{1-1/t}\sqrt{t/k})$.*

We will describe our polynomial system and algorithm, list deterministic conditions under which our algorithm will succeed, give an SoS proof of identifiability, and conclude the proof of Theorem 4.4.1 by analyzing the rounding step of our algorithm. We will defer technical details for how to encode some of the constraints of our polynomial system to Section 4.6.

4.4.1 An SoS Relaxation

Let t be a power of two, to be chosen later. For $\mu \in \Delta^n$, let $\mathcal{D} = \mathrm{Mul}_k(\mu)$. Let $X_1, ..., X_N \in \Delta^n$ be the set of iid samples from $\mathcal{D}_1, ..., \mathcal{D}_N$ respectively, where for each $i \in [N]$ we have $\mathcal{D}_i = \mathrm{Mul}_k(\mu_i)$ for some $\mu_i \in \Delta^n$ satisfying $d_{\mathrm{TV}}(\mu_i, \mu) \leq \omega$. Let $\{X_i\}_{i\in[N]} \in \Delta^n$ be those samples after an ε-fraction have been corrupted.

Program $\mathcal{P}$. *The variables are $\{w_i\}_{i\in[N]}$, $\{\hat{\mu}_i\}_{i\in[N]}$, and $\hat{\mu}$, and the constraints are*

1. *$w_i^2 = w_i$ for all $i \in [N]$.*
2. *$\sum w_i = (1-\varepsilon)N$.*
3. *For every $v \in \{\pm 1\}^n$ and every $i \in [N]$, $\langle \hat{\mu}_i - \hat{\mu}, v\rangle \leq 5\omega$.*
4. *$\sum_{i\in[N]} w_i X_i = \hat{\mu} \sum_{i\in[N]} w_i$.*
5. *For every $v \in \{\pm 1\}^n$*
$$\sum_{i\in[N]} w_i \langle X_i - \hat{\mu}_i, v\rangle^t \leq (8t/k)^{t/2} \cdot \sum_{i\in[N]} w_i \tag{4.12}$$
6. *$\hat{\mu}_i \geq 0$ for all $i \in [n]$ and $\sum_i \hat{\mu}_i = 1$.*

Note that constraints (3) and (5) are quantified over all $v \in \{\pm 1\}^n$, so as stated, Program $\mathcal{P}$ is a system of exponentially many polynomial constraints. In Section 4.6, we will explain how to encode these constraints as a small system of polynomial constraints. For now, we state the following without proof.

Lemma 4.4.2. *There is a system $\widehat{\mathcal{P}}$ of degree-$O(t)$ polynomial equations and inequalities in the variables $\{w_i\}$, $\{\hat{\mu}_i\}$, $\hat{\mu}$, and $n^{O(t)}$ other variables, whose coefficients depend on $\varepsilon, t, X_1, ..., X_n$ such that*

1. *(Satisfiability) With probability at least $1 - 1/\mathrm{poly}(n)$, $\widehat{\mathcal{P}}$ has a solution in which $\hat{\mu} = \mu$ and for each $i \in [N]$, $\hat{\mu}_i = \mu_i$ and w_i is the indicator for whether X_i is an uncorrupted point.*
2. *(Encodes Moment Bounds) $\widehat{\mathcal{P}} \vdash_{O(t)} \mathcal{P}$.*
3. *(Solvability) If Program $\widehat{\mathcal{P}}$ is satisfied, then for every integer $C > 0$, there is an $n^{O(Ct)}$-time algorithm which outputs a degree-Ct pseudodistribution which satisfies $\widehat{\mathcal{P}}$ up to additive error 2^{-n}.*

This suggests the following algorithm for learning from untrusted batches: use semidefinite programming to efficiently obtain a pseudodistribution over solutions to Program $\widehat{\mathcal{P}}$, and round this pseudodistribution to an estimate for μ by computing the pseudoexpectation of the $\hat{\mu}$ variable. A formal specification of this algorithm, which we call LearnFromUntrusted, is given in Algorithm 11 below.

Algorithm 11: LearnFromUntrusted($\varepsilon, \omega, n, k, \{X_i\}, t$)

```
Input: Corruption parameter $\varepsilon$, diversity parameter $\omega$, support size $n$, batch size $k$,
       samples $\{X_i\}_{i\in[N]}$, degree $t$
Output: Estimate $\hat{\mu}$
Run SDP solver to find a pseudodistribution $\tilde{\mathbb{E}}$ of degree $O(t)$ satisfying the
  constraints of Program $\widehat{\mathcal{P}}$.
return $\tilde{\mathbb{E}}[\hat{\mu}]$.
```

Remark 4.4.3. *Here we clarify some points regarding numerical accuracy of* LearnFromUntrusted *and the other algorithms presented in this chapter. Formally, the pseudodistribution computed by* LearnFromUntrusted *satisfies the constraints of Program $\widehat{\mathcal{P}}$ to precision 2^{-n} in the sense that for any sum-of-squares q and constraint polynomials $f_1, ..., f_\ell \in \widehat{\mathcal{P}}$ for which $\deg(q \cdot \prod_{i\in[\ell]} f_i) \le O(t)$, we have that $\tilde{\mathbb{E}}\left[q \cdot \prod_{i\in[\ell]} f_i\right] \ge -2^{-n}\|q\|_2$, where $\|q\|_2$ denotes the L_2 norm of the vector of coefficients of q. On the other hand, in our analysis, we show that $\widehat{\mathcal{P}} \vdash_{O(t)} \mathcal{P}$ and then argue using the constraints of $\mathcal{P}$ instead. But because the coefficients in the SoS proof that $\widehat{\mathcal{P}} \vdash_{O(t)} \mathcal{P}$ are polynomially bounded, the pseudodistribution computed by* LearnFromUntrusted *also satisfies the constraints of Program $\mathcal{P}$ to precision $2^{-\Omega(n)}$, which will be sufficient for the simple rounding we analyze in Section 4.4.4.*

4.4.2 Deterministic Conditions

We will condition on the following deterministic conditions holding simultaneously:

(I) The "Satisfiability" condition of Lemma 4.4.2 holds.

(II) The mean of the uncorrupted points concentrates:

$$\|\frac{1}{N}\sum_{i\in S_G}(X_i - \mu_i)\|_1 \le O(\omega + \varepsilon^{1-1/t}\sqrt{t/k})$$

(III) The empirical t-th moments concentrate:

$$\{v_i^2 = 1 \ \forall i \in [n]\} \vdash_t \frac{1}{N} \sum_{i \in [N]} \langle X_i - \mu_i, v \rangle^t - \frac{1}{N} \sum_{i \in [N]} \mathbb{E}_{X_i \sim \mathcal{D}_i} \langle X_i - \mu_i, v \rangle^t \leq (8t/k)^{t/2}$$

Lemma 4.4.4. *Conditions (I), (II), (III) hold simultaneously with probability* $1 - 1/\mathrm{poly}(n)$.

We first need the following elementary concentration inequalities.

Fact 4.4.5. *If* $X_1, ..., X_N$ *are drawn from* $Mul_k(\mu_1), ..., Mul_k(\mu_N)$ *respectively, then*

$$\Pr\left[\| \frac{1}{N} \sum_{i \in [N]} X_i - \frac{1}{N} \sum_{i \in [N]} \mu_i \|_1 > \varepsilon \right] \leq n \cdot e^{-2\varepsilon^2 N / n^2}$$

Proof. Note that for each $i \in [N]$, $j \in [n]$, $(X_i)_j$ is distributed as $\mathrm{Ber}((\mu_i)_j)$. So by Hoeffding's inequality,

$$\Pr\left[\left| \frac{1}{N} \sum_{i \in [N]} (X_i)_j - \frac{1}{N} \sum_{i \in [N]} (\mu_i)_j \right| \geq \eta \right] \leq e^{-2N\eta^2}.$$

The claim follows by taking $\eta = \varepsilon / n$ and union bounding over j. □

Fact 4.4.6. *Let* $X_1, ..., X_N$ *be independent samples from* $\mathcal{D}_1, ..., \mathcal{D}_N$. *For every* $i \in [N]$, *define* $Z_i = X_i - \mu_i$. *If* $N \geq \Omega(t \cdot (k/8t)^t \cdot n^{2t} \log^2(n))$, *then with probability* $1 - 1/\mathrm{poly}(n)$ *the following holds: for every multi-index* $\theta \in [t]^n$ *for which* $\sum \theta_i = t$ *we have that*

$$\left| \frac{1}{N} \sum_{i \in [N]} Z_i^\theta - \frac{1}{N} \sum_{i \in [N]} \mathbb{E}_{Z \sim \mathcal{D}_i - \mu_i} [Z^\theta] \right| \leq n^{-t} \cdot (8t/k)^{t/2}.$$

Proof. Note that because $X_i, \mu_i \in [0, 1]$, the random variables Z_i^θ only take values within $[-1, 1]$. By Hoeffding's inequality,

$$\Pr\left[\left| \frac{1}{N} \sum_{i \in [N]} Z_i^\theta - \frac{1}{N} \sum_{i \in [N]} \mathbb{E}_{Z \sim \mathcal{D}_i - \mu_i} [Z^\theta] \right| \geq \eta \right] \leq 2e^{-N\eta^2/2},$$

so the lemma follows by by taking $\eta = n^{-t} \cdot (8t/k)^{t/2}$ and union-bounding over all n^t choices of θ. □

Proof of Lemma 4.4.4. (I) holds with probability at least $1-1/\mathrm{poly}(n)$ according to Lemma 4.4.2.

Because $\{X_i\}_{i\in S_G}$ are independent draws from $\{\mu_i\}_{i\in S_G}$, (II) holds with probability at least $1-1/\mathrm{poly}(n)$ provided $N \ge \Omega((k/t)n^2\log^2 n \cdot \omega^{-2}\varepsilon^{-2})$, according to Fact 4.4.5.

Finally, we verify (III) holds with high probability. For every $i \in [N]$ define $Z_i = X_i - \mu_i$. The inequality we would like to exhibit an SoS proof for is equivalent to the inequality

$$\left|\sum_{\theta,\theta':|\theta|=|\theta'|=t/2} v_\theta v_{\theta'}\left(\frac{1}{N}\sum_{i\in[N]} Z_i^\theta Z_i^{\theta'} - \frac{1}{N}\sum_{i\in[N]} \mathbb{E}_{Z\sim\mathcal{D}_i-\mu_i}\left[Z^\theta Z^{\theta'}\right]\right)\right| \le (8t/k)^{t/2}, \tag{4.13}$$

where $v_\theta \triangleq \prod_{i\in\theta} v_i$. Note that

$$\{v_i^2 = 1\ \forall i \in [n]\} \vdash_t -1 \le v_\theta v_{\theta'} \le 1,$$

If the outcome of Fact 4.4.6 holds for all n^t monomials of the form $\theta \cup \theta'$, then there is a degree-t proof, using the axioms $\{v_i^2 = 1\ \forall i \in [n]\} \vdash_t$, that (4.13) holds. We conclude that (III) holds with probability $1-1/\mathrm{poly}(n)$.

By a union bound over all events upon which we conditioned, we conclude that (I), (II), (III) are simultaneously satisfied with probability $1-1/\mathrm{poly}(n)$. □

4.4.3 Identifiability

The key step is to give an SoS proof of identifiability. In other words, we must demonstrate in the SoS proof system that the constraints of Program $\mathcal{P}$ imply that $\hat{\mu}$ is sufficiently close to μ. The main claim in this section is the following.

Lemma 4.4.7. *Suppose Conditions (I)-(III) hold. Then for any $v \in \{\pm 1\}^n$, we have that*

$$\mathcal{P} \vdash_{O(t)} \langle \hat{\mu} - \mu, v\rangle^t \le O\left(\omega^t + \varepsilon^{t-1}(t/k)^{t/2}\right).$$

First note that for any $i \in [N]$,

$$\sum_{i\in[N]} w_i\langle\hat{\mu}-\mu, v\rangle = \sum_{i\in[N]} w_i\langle\hat{\mu}-\mu_i, v\rangle + \sum_{i\in[N]} w_i\langle\mu_i-\mu, v\rangle$$

$$\leq \sum_{i\in[N]} w_i\langle \hat{\mu} - \mu_i, v\rangle + 2N\omega, \tag{4.14}$$

where the inequality follows from the assumption that $d_{\mathrm{TV}}(\mu, \mu_i) \leq \omega$. We bound the former term in (4.14):

$$\begin{aligned}
\sum_{i\in[N]} w_i\langle \hat{\mu} - \mu_i, v\rangle &= \sum_{i\in[N]} w_i\langle X_i - \mu_i, v\rangle \\
&= \sum_{i\in S_G}\langle X_i - \mu_i, v\rangle + \sum_{i\in S_G}(w_i - 1)\langle X_i - \mu_i, v\rangle + \sum_{i\in S_B} w_i\langle X_i - \mu_i, v\rangle \\
&= \sum_{i\in S_G}\langle X_i - \mu_i, v\rangle + \sum_{i\in S_G}(w_i - 1)\langle X_i - \mu_i, v\rangle + \\
&\qquad \sum_{i\in S_B} w_i\langle X_i - \hat{\mu}_i, v\rangle + \sum_{i\in S_B} w_i\langle \hat{\mu} - \mu_i, v\rangle + \sum_{i\in S_B} w_i\langle \hat{\mu}_i - \hat{\mu}, v\rangle \\
&\leq \sum_{i\in S_G}\langle X_i - \mu_i, v\rangle + \sum_{i\in S_G}(w_i - 1)\langle X_i - \mu_i, v\rangle + \\
&\qquad \sum_{i\in S_B} w_i\langle X_i - \hat{\mu}_i, v\rangle + \sum_{i\in S_B} w_i\langle \hat{\mu} - \mu_i, v\rangle + 5N\varepsilon\omega
\end{aligned} \tag{4.15}$$

where the inequality follows from Constraint 3 of Program $\mathcal{P}$. This rearranges to

$$\sum_{i\in S_G} w_i\langle \hat{\mu} - \mu_i, v\rangle \leq 5N\varepsilon\omega + \sum_{i\in S_G}\langle X_i - \mu_i, v\rangle + \sum_{i\in S_G}(w_i - 1)\langle X_i - \mu_i, v\rangle + \sum_{i\in S_B} w_i\langle X_i - \hat{\mu}_i, v\rangle.$$

Taking the t-th power of both sides of (4.14) and invoking (4.15) and the inequality $\vdash_t (a+b+c+d+e)^t \leq \exp(t)(a^t + b^t + c^t + d^t + e^t)$, we conclude that

$$\begin{aligned}
\mathcal{P} \vdash_t \left(\sum_{i\in S_G} w_i\right)^t \langle \hat{\mu} - \mu, v\rangle^t \leq \exp(t) \Bigg[&(N(2+5\varepsilon)\omega)^t + \underbrace{\left(\sum_{i\in S_G}\langle X_i - \mu_i, v\rangle\right)^t}_{\text{Lemma 4.4.8}} \\
&+ \underbrace{\left(\sum_{i\in S_G}(w_i - 1)\langle X_i - \mu_i, v\rangle\right)^t}_{\text{Lemma 4.4.10}} + \underbrace{\left(\sum_{i\in S_B} w_i\langle X_i - \hat{\mu}_i, v\rangle\right)^t}_{\text{Lemma 4.4.11}} \Bigg],
\end{aligned} \tag{4.16}$$

which we bound using Lemmas 4.4.8, 4.4.10, and 4.4.11 below. Intuitively, the term for Lemma 4.4.8 corresponds to sampling error from uncorrupted samples from $\mathcal{D}$, the term for Lemma 4.4.10 corresponds to the possible failure of the subset selected by w_i to capture some small fraction of the uncorrupted samples, and the term for Lemma 4.4.11 corresponds to the error contributed by the adversarially chosen vectors.

Lemma 4.4.8. *Suppose Conditions (I)-(III) hold. Then for any $v \in \{\pm 1\}^n$, we have that*

$$\mathcal{P} \vdash_{O(t)} \left(\sum_{i \in S_G} \langle X_i - \mu_i, v \rangle \right)^t \le O(N)^t \cdot (\omega^t + \varepsilon^{t-1} \cdot (t/k)^{t/2}).$$

Proof. By SoS Holder's, we have that

$$\begin{aligned} \mathcal{P} \vdash_{O(t)} \left(\sum_{i \in S_G} \langle X_i - \mu_i, v \rangle \right)^t &= \left\langle \sum_{i \in S_G} (X_i - \mu_i), v \right\rangle^t \\ &\le \| \sum_{i \in S_G} (X_i - \mu_i) \|_1^t \\ &\le \left(N \cdot O(\omega + \varepsilon^{1-1/t} \cdot \sqrt{t/k}) \right)^t \\ &\le O(N)^t \cdot (\omega^t + \varepsilon^{t-1} \cdot (t/k)^{t/2}) \end{aligned}$$

as claimed, where the penultimate step follows by (II) and the last step follows by (scalar) Holder's. □

For Lemma 4.4.10, we will use the following helper lemma.

Lemma 4.4.9. *Suppose Condition (III) holds. Then for any $v \in \{\pm 1\}^n$, we have that*

$$\mathcal{P} \vdash_{O(t)} \sum_{i \in [N]} \langle X_i - \mu_i, v \rangle^t \le 2N(8t/k)^{t/2}.$$

Proof. By Lemma 4.4.4 and Lemma 4.3.3, we have that

$$\sum_{i \in [N]} \langle X_i - \mu_i, v \rangle^t \le N \cdot (8t/k)^{t/2} + \sum_{i \in [N]} \mathbb{E}_{X_i \sim \mathcal{D}_i} \left[(X_i - \mu_i)^{\otimes t/2} \right] \left[(X_i - \mu_i)^{\otimes t/2} \right]^\top \le 2N(8t/k)^{t/2}.$$

□

Lemma 4.4.10. *Suppose Conditions (I)-(III) hold. Then for any $v \in \{\pm 1\}^n$, we have that*

$$\mathcal{P} \vdash_{O(t)} \left(\sum_{i \in S_G} (w_i - 1) \langle X_i - \mu_i, v \rangle \right)^t \leq 2\varepsilon^{t-1} \cdot N^t \cdot (8t/k)^{t/2}.$$

Proof. By SoS Holder's, we have that

$$\begin{aligned} \mathcal{P} \vdash_{O(t)} \left(\sum_{i \in S_G} (w_i - 1) \langle X_i - \mu_i, v \rangle \right)^t &= \left(\sum_{i \in S_G} (1 - w_i) \langle X_i - \mu_i, v \rangle \right)^t \\ &\leq \left(\sum_{i \in S_G} (1 - w_i) \right)^{t-1} \left(\sum_{i \in S_G} \langle X_i - \mu_i, v \rangle^t \right) \\ &\leq (\varepsilon N)^{t-1} \cdot \sum_{i \in [N]} \langle X_i - \mu_i, v \rangle^t \\ &\leq (\varepsilon N)^{t-1} \cdot 2N (8t/k)^{t/2} \end{aligned}$$

where the third step follows from the fact that $\vdash_2 \sum_{i \in S_G} (1 - w_i) \leq \sum_{i \in [N]} (1 - w_i) = \varepsilon N$, and the fourth step follows from Lemma 4.4.9. □

Lemma 4.4.11. *Suppose Conditions (I)-(III) hold. Then for any $v \in \{\pm 1\}^n$, we have that*

$$\mathcal{P} \vdash_{O(t)} \left(\sum_{i \in S_B} w_i \langle X_i - \hat{\mu}_i, v \rangle \right)^t \leq 2\varepsilon^{t-1} N^t (8t/k)^{t/2}.$$

Proof. We have that

$$\begin{aligned} \mathcal{P} \vdash_{O(t)} \left(\sum_{i \in S_B} w_i \langle X_i - \hat{\mu}_i, v \rangle \right)^t &= \left(\sum_{i \in S_B} w_i^2 \langle X_i - \hat{\mu}_i, v \rangle \right)^t && (4.17) \\ &\leq \left(\sum_{i \in S_B} w_i \right)^{t-1} \cdot \left(\sum_{i \in S_B} w_i \langle X_i - \hat{\mu}_i, v \rangle^t \right) && (4.18) \\ &\leq \left(\sum_{i \in S_B} w_i \right)^{t-1} \cdot \left(\sum_{i \in [N]} w_i \langle X_i - \hat{\mu}_i, v \rangle^t \right) && (4.19) \\ &\leq |S_B|^{t-1} \cdot 2(8t/k)^{t/2} \sum_{i \in [N]} w_i && (4.20) \end{aligned}$$

$$
\begin{aligned}
&= 2(\varepsilon N)^{t-1}(8t/k)^{t/2} \cdot N \\
&= 2\varepsilon^{t-1} N^t (8t/k)^{t/2},
\end{aligned}
$$

where (4.17) follows from the Booleanity constraints, (4.18) follows from SoS Holder's, (4.19) follows from even-ness of t, (4.20) follows from the definition of $|S_B|$ and from the moment bound (4.12). □

We can now finish the proof of Lemma 4.4.7.

Proof of Lemma 4.4.7. By (4.16) and Lemmas 4.4.8, 4.4.10, and 4.4.11, we have that

$$
\mathcal{P} \vdash_{O(t)} \left(\sum_{i \in S_G} w_i \right)^t \langle \hat{\mu} - \mu, v \rangle^t \leq O(N)^t \left(\omega^t + \varepsilon^{t-1} (t/k)^{t/2} \right).
$$

Since $\mathcal{P} \vdash_2 \sum_{i \in S_G} w_i \geq (1 - 2\varepsilon)N$, we conclude that

$$
\mathcal{P} \vdash_{O(t)} \langle \hat{\mu} - \mu, v \rangle^t \leq O\left(\omega^t + \varepsilon^{t-1} (t/k)^{t/2} \right)
$$

as claimed. □

4.4.4 Rounding

We are now ready to complete the proof of Theorem 4.4.1 by specifying how to round a pseudodistribution satisfying Program $\widehat{\mathcal{P}}$.

Lemma 4.4.12. *Let $\tilde{\mathbb{E}}$ be a degree-$O(t)$ pseudodistribution satisfying $\widehat{\mathcal{P}}$. Then $\tilde{\mathbb{E}}[\hat{\mu}] \in \Delta^n$ and $d_{TV}(\tilde{\mathbb{E}}[\hat{\mu}], \mu) \leq O(\omega + \varepsilon^{1-1/t}\sqrt{t/k})$.*

Proof. The fact that $\tilde{\mathbb{E}}[\hat{\mu}]$ follows from the fact that $\tilde{\mathbb{E}}$ satisfies Constraints 6 of Program $\widehat{\mathcal{P}}$. For the second part of the lemma, note that by the dual characterization of L_1 distance, it suffices to show that for any $v \in \{\pm 1\}^n$,

$$
\langle \tilde{\mathbb{E}}[\hat{\mu}] - \mu, v \rangle \leq O\left(\omega + \varepsilon^{1-1/t}\sqrt{t/k} \right)
$$

By Lemma 4.4.2, $\widehat{\mathcal{P}} \vdash_{O(t)} \mathcal{P}$. Furthermore, by Lemma 4.4.7, for any $v \in \{\pm 1\}^n$,

$$\tilde{\mathbb{E}}[\langle \hat{\mu} - \mu, v\rangle^t] \leq O\left(\omega^t + \varepsilon^{t-1}(t/k)^{t/2}\right),$$

so we get that

$$\begin{aligned}\langle \tilde{\mathbb{E}}[\hat{\mu}] - \mu, v\rangle^t &\leq \tilde{\mathbb{E}}\left[\langle \hat{\mu} - \mu, v\rangle^t\right] \\ &\leq O\left(\omega^t + \varepsilon^{t-1}(t/k)^{t/2}\right),\end{aligned}$$

where the first step is a consequence of Fact 1.3.41. Now by the fact that $(a+b)^{1/t} \leq a^{1/t}+b^{1/t}$ for positive scalars a, b, we conclude. □

We can now complete the proof of Theorem 4.4.1.

Proof of Theorem 4.4.1. The output of our algorithm will be $\textsc{Round}[\tilde{\mathbb{E}}]$ for $\tilde{\mathbb{E}}$ satisfying Program $\widehat{\mathcal{P}}$ and therefore Program $\mathcal{P}$, so $\textsc{Round}$ produces a hypothesis h for which $d_{\mathrm{TV}}(h, \mu) \leq O(\omega + \varepsilon^{1-1/t} \cdot \sqrt{t/k})$, as claimed. □

4.5 Improved Sample Complexity Under Shape Constraints

In this section we prove the following, which says that the algorithmic framework of the preceding sections can be leveraged to learn *shape-constrained distributions* from untrusted batches with sample complexity *sublinear* in the domain size n.

Theorem 4.5.1. *Let $t \geq 4$ be any integer, and let $\eta > 0$. If μ is (η, s)-piecewise degree-d, then there is an algorithm that draws an ε-corrupted set of N ω-diverse batches of size k from μ for $N = \omega^{-2}\varepsilon^{-2}(sd\log n)^{O(t)} \cdot k^t/t^{t-1}$, runs in time $\omega^{-t}\varepsilon^{-t}(sdn)^{O(t)} \cdot k^{t^2}/t^{t(t-1)}$, and with probability $1-1/\mathrm{poly}(n)$ outputs a distribution $\hat{\mu}$ for which $d_{TV}(p, \hat{\mu}) \leq O(\eta+\omega+\varepsilon^{1-1/t}\sqrt{t/k})$.*

Importantly, by combining this result with known approximation theoretic results, we are able to obtain sample complexities that are either independent of the domain size or depend at most polylogarithmically on it, for a large class of natrual distributions, such as monotone distributions, monotone hazard rate distributions, log-concave distributions, discrete Guassians, Poisson Binomial distributions, and mixtures thereof, see e.g. [ADLS17] for more details. After giving the basic ingredients from VC complexity for how to learn shape-constrained distributions in sublinear sample complexity in a classical sense, we describe and analyze the polynomial system Program $\mathcal{P}'$, deferring technical details for how to encode some of the constraints of this program to Section 4.6 and Appendix 4.7.

4.5.1 $\mathcal{A}_K$ Norms and VC Complexity

Definition 4.5.2 ($\mathcal{A}_K$ norms, see e.g. [DL01])**.** *For positive integers $K \leq n$, define $\mathcal{A}_K$ to be the set of all unions of at most K disjoint intervals over $[n]$, where an interval is any subset of $[n]$ of the form $\{a, a+1, ..., b-1, b\}$. The $\mathcal{A}_K$ distance between two distributions p, q over $[n]$ is*

$$\|p - q\|_{\mathcal{A}_K} = \max_{S \in \mathcal{A}_K} |p(S) - q(S)|.$$

Equivalently, say that $v \in \{\pm 1\}^n$ has $2K$ sign changes *if there are exactly $2K$ indices $i \in [n-1]$ for which $v_{i+1} \neq v_i$. Then if $\mathcal{V}^n_{2K}$ denotes the set of all such v, we have*

$$\|p - q\|_{\mathcal{A}_K} = \frac{1}{2} \max_{v \in \mathcal{V}^n_{2K}} \langle p - q, v \rangle.$$

Note that

$$\|\cdot\|_{\mathcal{A}_1} \leq \|\cdot\|_{\mathcal{A}_2} \leq \cdots \leq \|\cdot\|_{\mathcal{A}_{n/2}} = \|\cdot\|_{TV}.$$

Definition 4.5.3. *We say that a distribution over $[n]$ is (η, s)*-piecewise degree-d *if there is a partition of $[n]$ into t disjoint intervals $\{[a_i, b_i]\}_{1 \leq i \leq s}$, together with univariate degree-d polynomials $r_1, ..., r_s$ and a distribution $\mathbf{q}$ on $[n]$, such that $d_{TV}(\mathbf{p}, \mathbf{q}) \leq \eta$ and such that for all $i \in [s]$, $\mathbf{q}(x) = r_i(x)$ for all $x \in [n]$ in $[a_i, b_i]$.*

Lemma 4.5.4. *Let $K = s(d+1)$. If μ is (η, s)-piecewise degree-d and $\|\mu - \hat{\mu}\|_{\mathcal{A}_K} \leq \zeta$, then there is an algorithm which, given the vector $\hat{\mu}$, outputs a distribution μ^* for which $d_{TV}(\mu, \mu^*) \leq 2\zeta + 4\eta$ in time* poly$(s, d, 1/\varepsilon)$.

Proof. Let μ' be a $(0, s)$-piecewise degree-d distribution for which $d_{\mathrm{TV}}(\mu, \mu') = \eta$. By

Theorem 4.5.5 below, one can produce an s-piecewise degree-d distribution μ^* minimizing $\|\hat{\mu}-\mu^*\|_{\mathcal{A}_K}$ to within additive error η in time $\mathrm{poly}(s,d,1/\eta)$. We already know by triangle inequality that

$$\|\hat{vp}-\mu'\|_{\mathcal{A}_K} \le \|\hat{vp}-\mu\|_{\mathcal{A}_K} + \|\mu-\mu'\|_{\mathcal{A}_K} \le \zeta+\eta,$$

so by η-approximate minimality we know $\|\hat{\mu}-\mu^*\|_{\mathcal{A}_K} \le \zeta + 2\eta$. By another application of triangle inequality, we conclude that $\|\mu'-\mu^*\|_{\mathcal{A}_K} \le 2\zeta+3\eta$. Because μ' and μ^* are both s-piecewise degree-d, the vector $\mu'-\mu^*$ has at most $2s(d+1)$ sign changes. Indeed, the common refinement of the intervals defining the two piecewise polynomials is at most $2s$ intervals, and the difference between two degree-d polynomials over any of these intervals is degree-d (the additional $+1$ comes from the endpoints of each of the intervals). So we get that $d_{\mathrm{TV}}(\hat{\mu}-\mu^* = \|\hat{\mu}-\mu^*\|_{\mathcal{A}_K} \le 2\zeta+3\eta$, and one final application of triangle inequality allows us to conclude that $d_{\mathrm{TV}}(\mu-\mu^*) = 2\zeta+4\eta$. □

Theorem 4.5.5 ([ADLS17]). *There is an algorithm which, given a vector $\mu\in\Delta^n$, computes an s-piecewise degree-d hypothesis h which minimizes $\|h-\mu\|_{\mathcal{A}_{s(d+1)}}$ to within additive error γ in time $n\cdot\mathrm{poly}(d,1/\gamma)$.*

Henceforth, we will focus solely on the problem of learning in $\mathcal{A}_\ell$ norm, where

$$\ell = 2K \triangleq 2s(d+1). \tag{4.21}$$

4.5.2 Another SoS Relaxation

To prove Theorem 4.5.1, by Lemma 4.5.4 it suffices to learn μ in $\mathcal{A}$ distance, that is, we wish to produce a hypothesis $\hat{\mu}$ for which $\frac{1}{2}\max_{v\in\mathcal{V}_\ell^n}\langle\mu-\hat{\mu},v\rangle$ is small.

Program $\mathcal{P}'$. *The variables are $\{w_i\}_{i\in[N]}$, $\{\hat{\mu}_i\}_{i\in[N]}$, and $\hat{\mu}$, and the constraints are*

1. *$w_i^2 = w_i$ for all $i\in[N]$.*

2. *$\sum w_i = (1-\varepsilon)N$.*

3. *For every $v\in\{\pm1\}^n$ with at most ℓ sign changes and every $i\in[N]$, $\langle\hat{\mu}_i-\hat{\mu},v\rangle\le 5\omega$.*

4. *$\sum_{i\in[N]} w_iX_i = \hat{\mu}\sum_{i\in[N]} w_i$.*

5. *For every $v \in \{\pm 1\}^n$ with at most ℓ sign changes,*

$$\sum_{i\in[N]} w_i \langle X_i - \hat{\mu}_i, v\rangle^t \le (8t/k)^{t/2} \cdot \sum_{i\in[N]} w_i.$$

6. *$\hat{\mu}_i \ge 0$ for all $i \in [n]$ and $\sum_i \hat{\mu}_i = 1$.*

Lemma 4.5.6. *There is a system $\widehat{\mathcal{P}}'$ of degree-$O(t)$ polynomial equations and inequalities in the variables $\{w_i\}$, $\{\hat{\mu}_i\}$, $\hat{\mu}$, and $n^{O(t)}$ other variables, whose coefficients depend on $\varepsilon, t, X_1, ..., X_n$ such that*

1. *(Satisfiability) With probability at least $1 - 1/\mathrm{poly}(n)$, Program $\widehat{\mathcal{P}}'$ has a solution in which $\hat{\mu} = \mu$ and for each $i \in [N]$, $\hat{\mu}_i = \mu_i$ and w_i is the indicator for whether X_i is an uncorrupted point.*

2. *(Encodes Moment Bounds) $\widehat{\mathcal{P}}' \vdash_{O(t)} \mathcal{P}'$.*

3. *(Solvability) If Program $\widehat{\mathcal{P}}'$ is satisfied, then for every integer $C > 0$, there is an $n^{O(Ct)}$-time algorithm which outputs a degree-Ct pseudodistribution which satisfies Program $\widehat{\mathcal{P}}$ up to additive error 2^{-n}.*

Together with Lemma 4.5.4, this suggests the following algorithm for learning from untrusted batches when μ is (η, s)-piecewise degree-d: use semidefinite programming to efficiently obtain a pseudodistribution over solutions to Program $\widehat{\mathcal{P}}$, round this pseudodistribution to an estimate for μ by computing the pseudoexpectation of the $\hat{\mu}$ variable, and then refine this by computing the best piecewise polynomial approximation to this estimate. The only difference between this algorithm and LearnFromUntrusted is the the third step.

A formal specification of this algorithm, which we call PiecewiseLearn, is given in Algorithm 12 below.

4.5.3 Deterministic Conditions and Identifiability

We will condition on the following deterministic conditions holding simultaneously:

```
Algorithm 12: PiecewiseLearn($\varepsilon, \omega, n, k, \{X_i\}, t, (\eta, s, d))$)
  Input: Corruption parameter $\varepsilon$, diversity parameter $\omega$, support size $n$, batch size $k$,
         samples $\{X_i\}_{i\in[N]}$, degree $t$, $(\eta, s, d)$ for which $\mu$ is $(\eta, s)$-piecewise degree-$d$
  Output: Estimate $\mu^*$
  Run SDP solver to find a pseudodistribution $\tilde{\mathbb{E}}$ of degree $O(t)$ satisfying the
    constraints of Program $\widehat{\mathcal{P}}$.
  $\tilde{\mu} \leftarrow \tilde{\mathbb{E}}[\hat{\mu}]$.
  $K \leftarrow s(d+1)$.
  Using the algorithm of [ADLS17], form the $s$-piecewise degree-$d$ distribution $\mu^*$ that
    minimizes $\|\tilde{\mu} - \mu^*\|_{\mathcal{A}_K}$ (up to additive error $\eta$).
  return $\mu^*$.
```

(I) The “Satisfiability” condition of Lemma 4.5.6 holds.

(II) The mean of the uncorrupted points concentrates in $\mathcal{A}_\ell$ norm:

$$\|\frac{1}{N}\sum_{i\in S_G}(X_i - \mu_i)\|_{\mathcal{A}_\ell} \le O(\varepsilon^{1-1/t}\sqrt{t/k})$$

(III) For every $v \in \{\pm 1\}^n$ with at most ℓ sign changes,

$$\left|\frac{1}{N}\sum_{i\in[N]}\langle X_i - \mu_i, v\rangle^t - \frac{1}{N}\sum_{i\in[N]}\mathbb{E}_{X_i\sim\mathcal{D}_i}\langle X_i - \mu_i, v\rangle^t\right| \le (8t/k)^{t/2}$$

Lemma 4.5.7. *Conditions (I), (II), (III) hold simultaneously with probability* $1 - 1/\mathrm{poly}(n)$.

Proof. (I) holds with probability at least $1 - 1/\mathrm{poly}(n)$ according to Lemma 4.5.6.

For (II), we will apply Lemma 4.6.12 with $\mathcal{N}$ taken to be the collection of all $v \in \{\pm 1\}^n$ with at most ℓ sign changes. $|\mathcal{N}| = n^{O(\ell)}$, so provided $|S_G| \ge \Omega((k/t)\varepsilon^{-2}\cdot \ell \log^2 n)$, we get that (II) holds with probability $1 - 1/\mathrm{poly}(n)$.

For (III), we will apply Lemma 4.7.3 with $\mathcal{N}$ taken to be the collection of all $v^{\otimes t/2}(v^{\otimes t/2})^\top$ for which $v \in \{\pm 1\}^n$ has at most ℓ sign
ges. $|\mathcal{N}| = n^{O(\ell)}$, so when $N \ge \Omega((k/8t)^t \cdot \ell \log n)$, (III) holds with probability $1 - 1/\mathrm{poly}(n)$. □

The SoS proof of identifiability given Program $\mathcal{P}'$ is identical to the proof of identifiability given Program $\mathcal{P}$ in Section 4.4.3, the only difference being that all intermediate steps in the proof are quantified over $v \in \{\pm 1\}^n$ with at most ℓ sign changes, rather than over all $v \in \{\pm 1\}^n$. This yields the following:

Lemma 4.5.8. *Suppose Conditions (I)-(III) hold. Then for any $v \in \{\pm 1\}^n$ with at most ℓ sign changes, we have that*

$$\mathcal{P}' \vdash_{O(t)} \langle \hat{\mu} - \mu, v \rangle^t \leq O\left(\omega^t + \varepsilon^{t-1}(t/k)^{t/2}\right).$$

4.5.4 Rounding

Once we have Lemma 4.5.8, the rounding step can be analyzed in essentially the same way as Lemma 4.4.12. We include a proof for completeness.

Lemma 4.5.9. *Let $\tilde{\mathbb{E}}$ be a pseudoexpectation satisfying Program $\widehat{\mathcal{P}}'$. Then $\tilde{\mathbb{E}}[\hat{\mu}] \in \Delta^n$ and $\|\tilde{\mathbb{E}}[\hat{\mu}] - \mu\|_{\mathcal{A}_\ell} \leq O(\omega + \varepsilon^{1-1/t} \cdot \sqrt{t/k})$.*

Proof. $\tilde{\mathbb{E}}[\hat{\mu}] \in \Delta^n$ because $\tilde{\mathbb{E}}$ satisfies Constraint 6 of Program $\mathcal{P}'$. For the second part of the lemma, by definition of $\mathcal{A}_\ell$ distance, it suffices to show that for any $v \in \{\pm 1\}^n$ with at most ℓ sign changes,

$$\langle \tilde{\mathbb{E}}[\hat{\mu}] - \mu, v \rangle \leq O\left(\omega + \varepsilon^{1-1/t}\sqrt{t/k}\right).$$

By Lemma 4.5.6, $\widehat{\mathcal{P}}' \vdash_{O(t)} \mathcal{P}'$. Furthermore, by Lemma 4.5.8, for any $v \in \{\pm 1\}^n$ with at most ℓ sign changes,

$$\tilde{\mathbb{E}}[\langle \hat{\mu} - \mu, v \rangle^t] \leq O\left(\omega^t + \varepsilon^{t-1}(t/k)^{t/2}\right),$$

so we get that

$$\begin{aligned}\langle \tilde{\mathbb{E}}[\hat{\mu}] - \mu, v \rangle^t &\leq \tilde{\mathbb{E}}[\langle \hat{\mu} - \mu, v \rangle^t] \\ &\leq O\left(\omega^t + \varepsilon^{t-1}(t/k)^{t/2}\right),\end{aligned}$$

where the first step is a consequence of Fact 1.3.41. Now by the fact that $(a+b)^{1/t} \leq a^{1/t} + b^{1/t}$ for positive scalars a, b, the lemma follows. □

We can now complete the proof of Theorem 4.5.1.

Proof of Theorem 4.5.1. The output of our algorithm will be $\textsc{Round}[\tilde{\mathbb{E}}]$ for $\tilde{\mathbb{E}}$ satisfying Program $\mathcal{P}'$. Because $\tilde{\mathbb{E}}[\hat{\mu}] \in \Delta^n$ satisfies $\|\tilde{\mathbb{E}}[\hat{\mu}] - \mu\|_{\mathcal{A}_\ell} \le O(\omega + \varepsilon^{1-1/t} \cdot \sqrt{t/k})$, we conclude that by Lemma 4.5.4, the assumption that μ is (η, s)-piecewise degree-d, and the fact that $\ell = 2s(d+1)$, $\textsc{Round}$ produces a hypothesis h for which $d_{\mathrm{TV}}(h, \mu) \le O(\eta + \omega + \varepsilon^{1-1/t} \cdot \sqrt{t/k})$, as claimed. □

4.6 Encoding Moment Constraints

In this section we will prove Lemmas 4.4.2 and 4.5.6. The programs $\widehat{\mathcal{P}}$ and $\widehat{\mathcal{P}}'$ referenced in those Lemmas will involve systems of inequalities among matrix-valued polynomials. We begin by giving an overview of how such inequalities fit into the SoS proof system.

4.6.1 Matrix SoS Proofs

Let $x_1, ..., x_n$ be formal variables. In this subsection we show how the SoS proof system can reason about constraints of the form $M(x) \succeq 0$, where $M(x)$ is some symmetric matrix whose entries are polynomials in x.

Let $M_1(x), ..., M_m(x)$ be symmetric matrix-valued polynomials of x of various sizes (1×1 matrix-valued polynomials are simply scalar polynomials), and let $q_1(x), ..., q_m(x)$ be scalar polynomials. The expression

$$\{M_1 \succeq 0, ..., M_m \succeq 0, q_1(x) = 0, ..., q_m(x) = 0\} \vdash_d p(x) \ge 0$$

means that there exists a vector u, a matrix $Q(x)$ whose entries are polynomials in the ideal generated by $q_1, ..., q_m$, and vector-valued polynomials $\{r_S^j\}_{j \le N, S \subseteq [m]}$ (where S's are multisets) for which

$$p(x) = Q(x) + u^\top \left[\sum_{S \subseteq [m]} \left(\sum_j (r_S^j(x))(r_S^j(x))^\top \right) \otimes [\otimes_{i \in S} M_i(x)] \right] u \tag{4.22}$$

and $Q(x)$ and the entries of each summand in (4.22) are all polynomials of degree at most d.

A pseudodistribution $\tilde{\mathbb{E}}$ of degree $2d$ is said to satisfy $\{M_1(x) \succeq 0, ..., M_m(x) \succeq 0\}$ if for every multiset $S \subseteq [m]$ and polynomial $p(x)$ for which the entries of $p(x)^2 \cdot (\otimes_{i \in S} M_i(x))$ are degree at most $2d$, we have

$$\tilde{\mathbb{E}}[p(x)^2 \cdot (\otimes_{i \in S} M_i(x))] \succeq 0.$$

Such pseudodistributions can still be found efficiently via semidefinite programming.

Proofs of the following basic lemmas about matrix SoS can be found in [HL18].

Lemma 4.6.1 ([HL18], Lemma 7.1)**.** *If $\tilde{\mathbb{E}}$ is a degree-2d pseudodistribution satisfying $\{M_1 \succeq 0, ..., M_m \succeq 0\}$ and furthermore*

$$\{M_1 \succeq 0, ..., M_m \succeq 0\} \vdash_{2d} M \succeq 0,$$

then $\tilde{\mathbb{E}}$ also satisfies $\{M_1 \succeq 0, ..., M_m \succeq 0, M \succeq 0\}$.

Lemma 4.6.2 ([HL18], Lemma 7.2)**.** *If $f(x)$ is a degree-d vector-valued polynomial of dimension s and $M(x)$ is an $s \times s$ symmetric matrix-valued polynomial of degree d', then*

$$\{M \succeq 0\} \vdash_{dd'} \langle f(x), M(x) f(x) \rangle \geq 0.$$

4.6.2 Moment Constraints for Program $\mathcal{P}$

We first show how to encode Constraint 3 of Program $\mathcal{P}$, namely that for each $i \in [N]$

$$\{v_i^2 = 1 \ \forall \ 1 \leq i \leq n\} \vdash_2 \langle \hat{\mu}_i - \hat{\mu}, v \rangle \leq 5\omega. \tag{4.23}$$

This would hold if there existed sum-of-squares polynomials $q_S(v, \hat{p}_i, \hat{p})$ for which $5\omega - \langle \hat{\mu}_i - \hat{\mu}, v \rangle = \sum_S \prod_{i \in S} (1 - v_i^2) \cdot q_S(v, \hat{p}_i, \hat{p})$ such that each summand on the right-hand side is of degree at most 2. So let Q^S be an $n \times n$ matrix of indeterminates, with entries indexed by $i, j \in [n]$, which will correspond to the matrix of coefficients of $q(v, \hat{p}_i, \hat{p})$ as a quadratic polynomial in v.

Next we show how to encode Constraint 4.12 of Program $\mathcal{P}$. For every $S \subset [n]$ of size at most $O(t)$, let M^S be an $n^{t/2} \times n^{t/2}$ matrix of indeterminates, one for each pair of multi-indices γ, ρ over $[n]$ both of degree at most $t/2$. We would like to impose constraints on the entries $M^S_{\gamma,\rho}$ so that psd-ness of the matrices in $\{M^S : S \subseteq [n]\}$ encodes the fact that

$$\{v_i^2 = 1 \ \forall 1 \le i \le n\} \vdash_{O(t)} \sum_{i\in[N]} w_i \langle X_i - \hat{\mu}_i, v\rangle^t \le 2 \cdot (8t/k)^{t/2} \sum_{i\in[N]} w_i \tag{4.24}$$

Recall that the condition (4.24) means that there exist polynomials p_S for which

$$2 \cdot (8t/k)^{t/2} \sum_{i\in[N]} w_i - \sum_{i\in[N]} w_i \langle X_i - \hat{\mu}_i, v\rangle^t = \sum_{S:|S|\le O(t)} p_S(v, \{w_i\}, \{\hat{\mu}_i\}, \hat{\mu}) \cdot \prod_{i\in S}(1 - v_i^2),$$

where each p_S is a sum-of-squares polynomial such that $p_S(v, \{w_i\}, \{\hat{\mu}_i\}, \hat{\mu}) \cdot \prod_{i\in S}(1 - v_i^2)$ is degree $O(t)$. M^S will correspond to the matrix of coefficients of $p_S(v, \{w_i\}, \{\hat{\mu}_i\}, \hat{\mu})$ as a degree-t polynomial in v. Specifically, we will consider the following program.

Program $\hat{\mathcal{P}}$. *The variables are $\{w_i\}_{i\in[N]}$, $\hat{\mu}$, $\{\hat{\mu}_i\}_{i\in[N]}$, $\{Q^S_{i,j}\}$, and $\{M^S_{\gamma,\rho}\}$ and the constraints are*

1. $w_i^2 = w_i$ *for all* $i \in [N]$.

2. $\sum w_i = (1 - \varepsilon)N$.

3. $5\omega - \langle \hat{\mu}_i - \hat{\mu}, v\rangle = \sum_S \prod_{i\in S}(1 - v_i^2) \cdot \langle v, Q^S v\rangle$

4. $\sum_{i\in[N]} w_i X_i = \hat{\mu} \cdot \sum_{i\in[N]} w_i$

5. $$2 \cdot (8t/k)^{t/2} - \frac{1}{(1-\varepsilon)N} \sum_{i\in[N]} w_i \langle X_i - \hat{\mu}, v\rangle^t = \sum_{S:|S|\le O(t)} \prod_{i\in S}(1 - v_i^2) \cdot \langle v^{\otimes t/2}, M^S v^{\otimes t/2}\rangle$$

6. $Q^S \succeq 0$ *for all* $S \subset [n]$ *for which* $|S| \le 2$

7. $M^S \succeq 0$ *for all* $S \subset [n]$ *for which* $|S| \le O(t)$.

8. $\hat{\mu}_i \ge 0$ *for all* $i \in [n]$ *and* $\sum_i \hat{\mu}_i = 1$.

Definition 4.6.3. *Define the* canonical assignment *to the variables $\{w_i\}_{i\in[N]}$, $\hat{\mu}$, and $\{\hat{\mu}_i\}_{i\in[N]}$ to be as follows: for each $i \in [N]$, $w_i = \mathbb{1}[X_i$ is uncorrupted$]$, $\hat{\mu}_i = \mu_i$, and $\hat{\mu} = \frac{1}{(1-\varepsilon)N}\sum_i w_i X_i$.*

Proof of Lemma 4.4.2. The fact that $\widehat{\mathcal{P}} \vdash_{O(t)} \mathcal{P}$ follows by Lemma 4.6.2, and solvability follows from the fact that the problem of outputting a degree-$O(t)$ pseudodistribution satisfying a system of degree-$O(t)$ polyomial constraints can be encoded as a semidefinite program of size $n^{O(t)}$.

It remains to show satisfiability of Program $\widehat{\mathcal{P}}$. Constraints 1, 2, and 4 are clearly satisfied by the canonical assignment.

For Constraints 3 and 6, we want to show that for each $i \in [N]$, the SoS proof (4.23) exists as a polynomial inequality only in the variable v, with $\{\hat{\mu}_i\}$ and $\hat{\mu}$ now fixed. Fix any $i \in [N]$ and for convenience define $\alpha_j = (\hat{\mu}_i - \hat{\mu})_j$. From Fact 1.3.44, we get that

$$\{v_i^2 = 1 \ \forall\ 1 \le i \le n\} \vdash_2 \langle \hat{\mu}_i - \hat{\mu}, v\rangle = \sum_{j=1}^n \alpha_j v_j \le \sum_{j=1}^n |\alpha_j| = \|\hat{\mu}_i - \hat{\mu}\|_1.$$

By triangle inequality and the fact that $d_{\mathrm{TV}}(\mu_i, \mu_j) \le 2\omega$ for all $j \in [N]$,

$$\begin{aligned} \|\hat{\mu}_i - \hat{\mu}\|_1 &\le \|\frac{1}{(1-\varepsilon)N} \sum_{j \in S_G : j \neq i} (\mu_i - \mu_j)\|_1 + \|\frac{1}{(1-\varepsilon)N} \sum_{j \in S_G} (X_j - \mu_j)\|_1 \\ &\le 4\omega + \|\frac{1}{(1-\varepsilon)N} \sum_{j \in S_G} (X_j - \mu_j)\|_1 \end{aligned}$$

By Fact 4.4.5 and the fact that $\{X_j\}_{j \in S_G}$ is a collection of independent draws from $\{\mathrm{Mul}_k(\mu_j)\}_{j \in S_G}$ respectively, we know that

$$\|\frac{1}{(1-\varepsilon)N} \sum_{j \in S_G} (X_j - \mu_j)\|_1 \le \omega$$

with probability at least $1 - n \cdot e^{-2\omega^2 N/n^2}$, from which (4.23) follows.

Finally, for Constraints 5 and 7, suppose the following SoS proof exists:

$$\{v_i^2 = 1 \ \forall\ 1 \le i \le n\} \vdash_{O(t)} \frac{1}{(1-\varepsilon)N} \sum_{i \in S_G} \langle X_i - \hat{\mu}_i, v\rangle^t \le 2 \cdot (8t/k)^{t/2}, \tag{4.25}$$

where v is the only variable and $\{w_i\}$, $\{\hat{\mu}_i\}$, and μ have all been fixed. By definition, this means that there exist sum-of-squares polynomials $p_S(v)$ for every $S \subset [n]$ of size at most

$O(t)$ such that $p_S(v) \cdot \prod_{i \in S}(1 - v_i^2)$ is degree $O(t)$ and

$$2 \cdot (8t/k)^{t/2} - \frac{1}{(1-\varepsilon)N} \sum_{i \in S_G} w_i \langle X_i - \hat{\mu}_i, v \rangle^t = \sum_{S : |S| \le O(t)} p_S(v) \cdot \prod_{i \in S} (1 - v_i^2).$$

By taking M^S to be the matrix of coefficients for which $\langle v^{\otimes t/2}, M^S v^{\otimes t/2} \rangle = p_S(v)$ and noting that $M^S \succeq 0$ because p_S is an SoS, we satisfy the remaining Constraints 5 and 7 of Program $\widehat{\mathcal{P}}$.

It remains to verify that the SoS proof (4.25) exists with high probability. Because $\hat{\mu}_i = \mu_i$, it is enough to show that the SoS proof

$$\{v_i^2 = 1 \ \forall 1 \le i \le n\} \vdash_{O(t)} \frac{1}{(1-\varepsilon)N} \sum_{i \in S_G} \langle X_i - \mu_i, v \rangle^t \le 2 \cdot (8t/k)^{t/2},$$

exists. It is enough to bound the quantity

$$b(v) \triangleq \frac{1}{(1-\varepsilon)N} \sum_{i \in S_G} \langle X_i - \mu_i, v \rangle^t - \frac{1}{(1-\varepsilon)N} \sum_{i \in S_G} \mathbb{E}_{X \sim \mathcal{D}_i} \langle X - \mu_i, v \rangle^t$$

by $b(v) \le (8t/k)^{t/2}$. Together with Lemma 4.3.3, this will conclude the proof. But the desired bound on $b(v)$ follows by condition (III) in Lemma 4.4.4, with probability $1 - 1/\mathrm{poly}(n)$. □

4.6.3 Moment Constraints for Program $\mathcal{P}'$

The only changes in going from Program $\mathcal{P}$ to Program $\mathcal{P}'$ are Constraints 3 and 5. In this section, we explain how to succinctly quantify over all $v \in \{\pm 1\}^n$ with at most ℓ sign changes. To describe this encoding, we first recall some basic facts about the (discretized) Haar wavelet basis.

Haar Wavelets

Definition 4.6.4. *Let m be a positive integer and let $n = 2^m$. The* Haar wavelet basis *is an orthonormal basis over $\mathbb{R}^n$ consisting of the* father wavelet $\psi_{0_{father},0} = n^{-1/2} \cdot \mathbf{1}$*, the* mother wavelet $\psi_{0_{mother},0} = n^{-1/2} \cdot (1, ..., 1, -1, ..., -1)$ *(where $(1, ..., 1, -1, ..., -1)$ contains $n/2$ 1's and $n/2$ -1's), and for every i, j for which $1 \le i < m$ and $0 \le j < 2^i$, the wavelet $\psi_{i,j}$ whose*

$2^{m-i} \cdot j+1, ..., 2^{m-i} \cdot j+2^{m-i-1}$-th coordinates are $2^{-(m-i)/2}$ and whose $2^{m-i} \cdot j + (2^{m-i-1} + 1), ..., 2^{m-i} \cdot j + 2^{m-i}$-th coordinates are $-2^{-(m-i)/2}$, and whose remaining coordinates are 0.

Let H_m denote the $n \times n$ matrix whose rows consist of the vectors of the Haar wavelet basis for $\mathbb{R}^n$. When the context is clear, we will omit the subscript and refer to this matrix as H.

Example 4.6.5. *The Haar wavelet basis for $\mathbb{R}^8$ consists of the vectors*

$$\begin{aligned}
\psi_{0_{father},0} &= 2^{-3/2}(1,1,1,1,1,1,1,1)\\
\psi_{0_{mother},0} &= 2^{-3/2}(1,1,1,1,-1,-1,-1,-1)\\
\psi_{1,0} &= 2^{-1}(1,1,-1,-1,0,0,0,0)\\
\psi_{1,1} &= 2^{-1}(0,0,0,0,1,1,-1,-1)\\
\psi_{2,0} &= 2^{-1/2}(1,-1,0,0,0,0,0,0)\\
\psi_{2,1} &= 2^{-1/2}(0,0,1,-1,0,0,0,0)\\
\psi_{2,2} &= 2^{-1/2}(0,0,0,0,1,-1,0,0)\\
\psi_{2,3} &= 2^{-1/2}(0,0,0,0,0,0,1,-1)
\end{aligned}$$

The key observation is that there is an orthonormal basis under which any $v \in \{\pm 1\}^n$ with at most ℓ sign changes has an $(\ell \log n + 1)$-sparse representation.

Define $\mathcal{T} \triangleq \{0_{\text{father}}, 0_{\text{mother}}, 1, ..., m-1\}$. By abuse of notation, we will sometimes identify the indices 0_{father} and 0_{mother} with their numerical value of 0.

Lemma 4.6.6. *Let $v \in \{\pm 1\}^n$ have at most ℓ sign changes. Then*

$$\sum_{i \in \mathcal{T}} \sum_{j=0}^{2^i - 1} 2^{-(m-i)/2} |\langle \psi_{i,j}, v\rangle| \leq \ell \log n + 1. \tag{4.26}$$

Proof. We first show that Hv has at most $\ell \log n + 1$ nonzero entries. For any $\psi_{i,j}$ with nonzero entries at indices $[a,b] \subset [n]$ and such that $i \neq 0_{\text{father}}$, if v has no sign change in the interval $[a,b]$, then $\langle \psi_{i,j}, v\rangle = 0$. For every index $\nu \in [n]$ at which v has a sign change, there are at most $m = \log n$ choices of i,j for which $\psi_{i,j}$ has a nonzero entry at index ν, from which the claim follows by a union bound over all ℓ choices of ν, together with the fact that

$\langle \psi_{0_{\text{father}},0}, v\rangle$ may be nonzero.

Now for each (i,j) for which $\langle \psi_{i,j}, v\rangle \neq 0$, note that

$$2^{-(m-i)/2} \cdot |\langle \psi_{i,j}, v\rangle| \leq 2^{-(m-i)/2} \cdot \left(2^{-(m-i)/2} \cdot 2^{m-i}\right) = 1,$$

from which (4.26) follows. □

For notational simplicity in the arguments below, for $\nu \in [n]$, if the ν-th element of the Haar wavelet basis for $\mathbb{R}^n$ is some $\psi_{i,j}$, then let $\mu^{(\nu)}$ denote the weight $2^{-(m-i)/2}$. Also, for any $i \in \mathcal{T}$, let $T_i \subset [n]$ denote the set of all indices ν for which the ν-th Haar wavelet is of the form $\psi_{i,j}$ for some j.

The Matrix SoS Encoding By Lemma 4.6.6, instead of quantifying over all $v \in \{\pm 1\}^n$ with at most ℓ sign changes in Constraints 3 and 5 of Program $\mathcal{P}'$, we can quantify over all $v \in \mathbb{R}^n$ with Frobenius norm at most n and for which (4.26) is satisfied. Specifically, we can ask for an SoS proof of

$$\langle \hat{\mu}_i - \hat{\mu}, v\rangle \leq 5\omega \tag{4.27}$$

using Axioms 1.

Axioms 1 (Axioms for Constraint 3)**.** *Let $\mathbf{W}_1, ..., \mathbf{W}_n$ be auxiliary scalar variables.*

1. $v_i^2 = 1$ *for all* $i \in [n]$

2. $-\mathbf{W}_i \leq (Hv)_i \leq \mathbf{W}_i$ *for all* $i \in [n]$

3. $\sum_i \mu^{(i)} \cdot \mathbf{W}_i \leq \ell \log n + 1,$

Likewise, we can ask for an SoS proof of

$$\frac{1}{(1-\varepsilon)N} \sum_{i \in [N]} w_i \langle X_i - \hat{\mu}_i, v\rangle^t \leq 2 \cdot (8t/k)^{t/2}, \tag{4.28}$$

using Axioms 2.

Axioms 2 (Axioms for Constraint 5)**.** *Let $\{\mathbf{U}_\alpha\}$, where α ranges over all monomials in the indices $[n]$ of degree $t/2$.*

1. $v_i^2 = 1$ *for all* $i \in [n]$

2. $-\mathbf{U}_\alpha \le (H^{\otimes t/2} v^{\otimes t/2})_\alpha \le \mathbf{U}_\alpha$ *for all monomials* α *of degree* $t/2$

3. $\sum_\alpha \mu^{(\alpha)} \mathbf{U}_\alpha \le (\ell \log n + 1)^{t/2}$,

where $\mu^{(\alpha)} \triangleq \prod_{i\in\alpha} \mu^{(i)}$.

As in the proof of Lemma 4.4.2, the values of $\{\hat{\mu}_i\}$ and $\{w_i\}$ will be given by the canonical assignment, so the only variables in the SoS proofs of (4.27) and (4.28) will be $v_1, ..., v_n$ and, respectively, $\{\mathbf{W}_i\}_{i\in[n]}$ and $\{\mathbf{U}_\alpha\}_{|\alpha|\le t/2}$.

By definition, the existence of a degree-d SoS proof for (4.27) using Axioms 1 is equivalent to the existence of polynomials $f_J^{K_1,K_2}(v, \mathbf{W}, \{\hat{\mu}_i\}, \hat{\mu})$ and $g_J^{K_1,K_2}(v, \mathbf{W}, \{\hat{\mu}_i\}, \hat{\mu})$ for $J, K_1, K_2 \subset [n]$ for which

$$5\omega - \langle \hat{\mu}_i - \hat{\mu}, v\rangle = \sum_{J,K_1,K_2} f_J^{K_1,K_2} h_J^{K_1,K_2} + \left(\ell \log n + 1 - \sum_{i\in[n]} \mu^{(i)} \mathbf{W}_i\right) \sum_{J,K_1,K_2} g_J^{K_1,K_2} \cdot h_J^{K_1,K_2},$$

where

$$h_J^{K_1,K_2} \triangleq \prod_{i\in J} (1 - v_i^2) \cdot \prod_{k_1\in K_1} (\mathbf{W}_{k_1} - (Hv)_{k_1}) \cdot \prod_{k_2\in K_2} (\mathbf{W}_{k_2} + (H_v)_{k_2}),$$

and where each $f_J^{K_1,K_2}$ and $g_J^{T_1,T_2}$ is a sum-of-squares polynomial such that $f_J^{K_1,K_2} \cdot h_J^{K_1,K_2}$ and $(\mathbf{W}_j - (Hv)_j) \cdot g_J^{K_1,K_2} \cdot h_J^{K_1,K_2}$ is degree d. We will take this degree to be $d = O(1)$.

Completely analogously, the existence of a degree-d SoS proof for (4.28) using Axioms 2 is equivalent to the existence of polynomials $p_S^{T_1,T_2}(v, \mathbf{U}, \{w_i\}, \{\hat{\mu}_i\}, \hat{\mu})$ and $q_S^{T_1,T_2}(v, U, \{w_i\}, \{\hat{\mu}_i\}, \hat{\mu})$ for $S \subset [n]$, $T_1, T_2 \subseteq \{\alpha : |\alpha| \le t/2\}$ for which

$$2 \cdot (8t/k)^{t/2} \sum_{i\in[N]} w_i - \sum_{i\in[N]} w_i \langle X_i - \hat{\mu}_i, v\rangle^t = \sum_{S,T_1,T_2} p_S^{T_1,T_2} r_S^{T_1,T_2} + \left((\ell \log n + 1)^{t/2} - \sum_\alpha \mu^{(\alpha)} \cdot \mathbf{U}_\alpha\right) \sum_{S,T_1,T_2} \cdot q_S^{T_1,T_2} \cdot r_S^{T_1,T_2}$$

where

$$r_S^{T_1,T_2} \triangleq \prod_{i\in S}\left(1-v_i^2\right)\cdot\prod_{\alpha\in T_1}\left(\mathbf{U}_\alpha-(H^{\otimes t/2}v^{\otimes t/2})_\alpha\right)\cdot\prod_{\beta\in T_2}\left(\mathbf{U}_\beta+(H^{\otimes t/2}v^{\otimes t/2})_\beta\right)$$

and where each $p_S^{T_1,T_2}$ and $q_S^{T_1,T_2}$ is a sum-of-squares polynomial such that $p_S^{T_1,T_2}\cdot r_S^{T_1,T_2}$ and $\left(\mathbf{U}_\alpha-(H^{\otimes t/2}v^{\otimes t/2})_\alpha\right)\cdot q_S^{T_1,T_2}\cdot r_S^{T_1,T_2}$ is degree d. We will take this degree to be $d=O(t)$.

Let $F_J^{K_1,K_2}$ and $G_J^{K_1,K_2}$ respectively denote the matrices of coefficients of $f_J^{K_1,K_2}$ and $g_J^{K_1,K_2}$ as degree-$O(1)$ polynomials solely in the variables $\{v_i\}$ and $\{\mathbf{W}_i\}$, with entries denoted by $(F_J^{K_1,K_2})_{\gamma,\rho}$ and $(G_J^{K_1,K_2})_{\gamma,\rho}$. Likewise, let $P_S^{T_1,T_2}$ and $Q_S^{T_1,T_2}$ respectively denote the matrices of coefficients of $p_S^{T_1,T_2}$ and $q_S^{T_1,T_2}$ as degree-$O(t)$ polynomials solely in the variables $\{v_i\}$ and $\{\mathbf{U}_\alpha\}$, with entries denoted by $(P_S^{T_1,T_2})_{\gamma,\rho}$ and $(Q_S^{T_1,T_2})_{\gamma,\rho}$.

Remark 4.6.7. *As we will demonstrate in the course of our analysis, we only need consider K_1, K_2 of size at most 1, and T_1, T_2 of size at most 2, so the total number of constraints in the overall program will only be singly-exponential in t.*

We will consider the following program.

Program $\hat{\mathcal{P}}'$. *The variables are $\{w_i\}_{i\in[N]}$, $\hat{\mu}$, $\{\hat{\mu}_i\}_{i\in[N]}$, $\{Q_{ij}\}$, $\{(P_S^{T_1,T_2})_{\gamma,\rho}\}$, $\{(Q_S^{T_1,T_2})_{\gamma,\rho}\}$, and the constraints are*

1. $w_i^2 = w_i$ *for all* $i\in[N]$

2. $(1-\varepsilon)N \le \sum w_i \le (1-\varepsilon)N$

3. $$5\omega-\langle\hat{\mu}_i-\hat{\mu},v\rangle = \sum_{J,K_1,K_2} h_J^{K_1,K_2}\cdot\langle(v,\mathbf{W})^{\otimes t/2},F_K^{J_1,J_2}(v,\mathbf{W})^{\otimes t/2}\rangle$$
$$+\left(\ell\log n+1-\sum_i\mu^{(i)}\mathbf{W}_i\right)\sum_{J,K_1,K_2} h_J^{K_1,K_2}\cdot\langle(v,\mathbf{W})^{\otimes t/2},G_K^{J_1,J_2}(v,\mathbf{W})^{\otimes t/2}\rangle$$

4. $\sum_{i\in[N]} w_iX_i = \hat{\mu}\cdot\sum_{i\in[N]} w_i$

5.

$$2 \cdot (8t/k)^{t/2} \sum_{i\in[N]} w_i - \sum_{i\in[N]} w_i \langle X_i - \hat{\mu}_i, v\rangle^t = \sum_{S,T_1,T_2} r_S^{T_1,T_2} \cdot \langle (v, \mathbf{U})^{\otimes t/2}, P_S^{T_1,T_2}(v,\mathbf{U})^{\otimes t/2}\rangle$$
$$+ \left((\ell \log n + 1)^{t/2} - \sum_\alpha \mu^{(\alpha)} \cdot \mathbf{U}_\alpha \right) \sum_{S,T_1,T_2} r_S^{T_1,T_2} \cdot \langle (v,\mathbf{U})^{\otimes t/2}, Q_S^{T_1,T_2}(v,\mathbf{U})^{\otimes t/2}\rangle$$

6. *$F_S^{T_1,T_2}, G_S^{T_1,T_2} \succeq 0$ for all $T_1, T_2, S \subset [n]$ for which $|T_1|, |T_2|, |S| \le O(t)$..*

7. *$P_S^{T_1,T_2}, Q_S^{T_1,T_2} \succeq 0$ for all $T_1, T_2, S \subset [n]$ for which $|T_1|, |T_2|, |S| \le O(t)$.*

8. *$\hat{\mu}_i \ge 0$ for all $i \in [n]$ and $\sum_i \hat{\mu}_i = 1$.*

Proof of Lemma 4.5.6. As before, solvability follows from the fact that the problem of outputting a degree-$O(t)$ pseudodistribution satisfying a system of degree-$O(t)$ polynomial constraints can be encoded a a semidefinite program of size $n^{O(t)}$.

The fact that $\widehat{\mathcal{P}}' \vdash_{O(t)} \mathcal{P}'$ follows by definition and by Lemma 4.26.

Finally, we verify that under the canonical assignment, with high probability over $X_1, ..., X_N$ there exists a satisfying assignment to the remaining variables of Program $\widehat{\mathcal{P}}'$. As in the proof of Lemma 4.4.2, the canonical assignment clearly satisfies Constraints 1, 2, and 4.

We prove that Constraints 3 and 6 are satisfiable with high probability in Lemma 4.6.9, and we prove that Constraints 5 and 7 are satisfiable with high probability in Lemma 4.6.13. □

The following fact will be useful in the proofs of Lemma 4.6.9 and 4.6.13.

Lemma 4.6.8 ("Shelling trick"). *If $v \in \mathbb{R}^m$ satisfies $\|v\|_2 \le C$ and $\|v\|_1 = C \cdot \sqrt{k}$, then there exist k-sparse vectors $v_1, ..., v_{m/k}$ with disjoint supports for which $v = \sum_{i=1}^{m/k} v_i$ and $\sum_{i=1}^{m/k} \|v_i\|_2 \le 2C$.*

Proof. We may assume without loss of generality that $C = 1$. Let $B_1 \subset [m]$ be the indices of the k largest entries of v, B_2 be those of the next k largest, and so on, so we may write $[m]$ as the disjoint union $B_1 \cup \cdots \cup B_{m/k}$. For $i \in [m/k]$, define $v_i \in \mathbb{R}^m$ to be the restriction of v to the coordinates indexed by B_i. For any i, note that for any $j \in B_i$, $|v_j| \le \frac{1}{k}\|v_{j-1}\|_1$, so

$$\|v_i\|_2^2 = \sum_{j \in B_i} v_j^2 \leq k \cdot \frac{1}{k^2} \cdot \|v_{i-1}\|_1^2 = \frac{1}{k}\|v_{i-1}\|_1^2.$$

So $\|v_i\|_2 \leq \|v_{i-1}\|_1/\sqrt{k}$ and thus

$$\sum_{i=1}^{m/k} \|v_i\|_2 \leq \|v_1\|_2 + \frac{1}{\sqrt{k}}\|v\|_1 \leq 2$$

as desired. □

Lemma 4.6.9. *Under the canonical assignment, with high probability there is some choice of $\{(F_K^{J_1,J_2})_{\gamma,\rho}\}$ and $\{(G_K^{J_1,J_2})_{\gamma,\rho}\}$ for which Constraints 3 and 6 are satisfied.*

Proof. We first write

$$\langle \hat{\mu}_i - \hat{\mu}, v\rangle = \frac{1}{m}\sum_{j \neq i} \langle \mu_i - \mu_j, v\rangle + \frac{1}{m}\sum_{j \in S_G} \langle X_j - \mu_j, v\rangle.$$

Note that by Fact 1.3.44,

$$\{v_i^2 = 1 \ \forall\ 1 \leq i \leq n\} \vdash_2 \langle \mu_i - \mu_j, v\rangle \leq \|\mu_i - \mu_j\|_1 \leq \|\mu_i - \mu\|_1 + \|\mu_j - \mu\|_1 \leq 4\omega.$$

It remains to show that with high probability, there is a degree-$O(t)$ proof that Axioms 1 imply $\frac{1}{m}\sum_{j \in S_G} \langle X_j - \mu_j, v\rangle \leq \omega$.

Equivalently, we must show that for any degree-t pseudodistribution $\tilde{\mathbb{E}}$ over the variables v and $\mathbf{U}$ which satisfies Axioms 1, we have that

$$\frac{1}{m}\sum_{j \in S_G} \langle X_j - \mu_j, \tilde{\mathbb{E}}[v]\rangle \leq \omega. \tag{4.29}$$

The set of vectors $\tilde{\mathbb{E}}[v]$ arising from pseudodistributions $\tilde{\mathbb{E}}$ satisfying Axioms 1 is some convex set $\mathcal{J} \subset \mathbb{R}^n$.

Lemma 4.6.10. *Let $\mathcal{J}$ be the convex set of all vectors of the form $\tilde{\mathbb{E}}[v]$ for some degree-t pseudodistribution $\tilde{\mathbb{E}}$ over the variables $v, \mathbf{W}$ satisfying Axioms 1.*

Additionally, let $\mathcal{J}_1, \mathcal{J}_2 \subset \mathbb{R}^n$ consist of all vectors u for which $\sum_i \mu^{(i)}|u_i| \le \ell \log n + 1$ and for which $\|u\|_2 \le \sqrt{n}$ respectively. Then

$$\mathcal{J} \subset H^{-1}(\mathcal{J}_1 \cap \mathcal{J}_2)$$

Proof. Take any $u \in \mathcal{J}$. We first show that $u \in H^{-1} \cdot \mathcal{J}_1$. By linearity of $\tilde{\mathbb{E}}$, we may write u as

$$u = H^{-1} \cdot \tilde{\mathbb{E}}[Hv].$$

For any $i \in [n]$, the second of Axioms 1 immediately implies that

$$-\mathbf{W}_i \le \tilde{\mathbb{E}}\left[(Hv)_i\right] \le \mathbf{W}_i.$$

We emphasize that this is the only place where we use the second of Axioms 1, and only in a linear fashion, hence Remark 4.6.7.

So $\sum_i \mu^{(i)}|(Hu)_i| \le \tilde{\mathbb{E}}[\sum_i \mu^{(i)}\mathbf{W}_i] \le \ell \log n + 1$, where the last inequality follows by the third of Axioms 1.

Finally, to show that $u \in H^{-1} \cdot \mathcal{J}_2$, note first that by orthonormality of H, it is enough to show that $u \in \mathcal{J}_2$. But this follows immediately from the fact that $\tilde{\mathbb{E}}$ satisfies the first of Axioms 1, which by (1.20) implies that $-1 \le \tilde{\mathbb{E}}[v_i] \le 1$ for all $i \in [n]$, from which we conclude that $\|u\|_2^2 = n$ and thus $u \in \mathcal{J}_2$. □

Lemma 4.6.11. *For every $\eta \le (\ell \log n + 1)^{-1}$, there exists a set $\mathcal{N} \subset \mathbb{P}_{n-1}(\mathbb{R})$ of size $O(n^{3/2}/\eta)^s$ such that for every $u \in H^{-1}(\mathcal{J}_1 \cap \mathcal{J}_2)$, there exists some $\tilde{u} = \sum_\nu \alpha_\nu \cdot u^*_\nu$ for $u^*_\nu \in \mathcal{N}$ such that 1) $\|u - \tilde{u}\|_2 \le \eta$, 2) $\sum_\nu \alpha_\nu \le 1$, and 3) $\|u^*_\nu\|_\infty \le 2(\ell \log n + 1)$ for all ν.*

Proof. Let $s = \ell \log n + 1$, and let $m = \log n$. Let $\mathcal{N}'$ be an $\frac{\eta}{(m+1)\sqrt{n}}$-net in L_2 norm for all s^2-sparse vectors in $\mathbb{S}^{n-1}$. Because $\mathbb{S}^{s^2-1}$ has an $\frac{\eta}{(m+1)\sqrt{n}}$-net in L_2 norm of size $(3(m+1)\sqrt{n}/\eta)^{s^2}$, by a union bound we have that $|\mathcal{N}'| \le \binom{n}{s^2} \cdot (3(m+1)\sqrt{n}/\eta)^{s^2} = O(n^{3/2}\log n/\eta)^{s^2}$.

Take any $u \in H^{-1}(\mathcal{J}_1 \cap \mathcal{J}_2)$ and consider $w \triangleq Hu \in \mathcal{J}_1 \cap \mathcal{J}_2$. We may write w as $\sum_{i \in \mathcal{T}} w[i]$, where

$$w[i] = \sum_{\nu \in T_i} w_\nu \cdot e_\nu$$

for e_ν the ν-th standard basis vector in $\mathbb{R}^n$.

As the nonzero entries of $w[i]$ are just a subset of those of w, we clearly have $\|w[i]\|_2 \le \sqrt{n}$ for all $i \in \mathcal{T}$. Moreover, because $w \in \mathcal{J}_1$, we have that

$$\sum_i 2^{-(m-i)/2} |w[i]| \le s, \tag{4.30}$$

so in particular

$$\|w[i]\|_1 \le 2^{(m-i)/2} \cdot s = 2^{-i/2} \cdot s\sqrt{n}.$$

We can thus apply Lemma 4.6.8 to conclude that for each $i \in [m]$, $w[i] = \sum_j w^{i,j}$ for some vectors $\{w^{i,j}\}_j$ of sparsity at most $\lceil 2^{-i} \cdot s^2 \rceil \le s^2$ and for which

$$\sum_j \|w^{i,j}\|_2 \le \sqrt{n}.$$

For each $w^{i,j}$, there is some $(w')^{i,j} \in \mathcal{N}'$ such that if we define $\tilde{w}^{i,j} \triangleq \|w^{i,j}\|_2 \cdot (w')^{i,j}$, then we have

$$\|w^{i,j} - \tilde{w}^{i,j}\|_2 \le \frac{\eta}{(m+1)\sqrt{n}} \cdot \|w^{i,j}\|_2. \tag{4.31}$$

Defining $\tilde{w}[i] \triangleq \sum_j \tilde{w}^{i,j}$, we get that

$$\|w[i] - \tilde{w}[i]\|_2 \le \frac{\eta}{(m+1)\sqrt{n}} \sum_j \|w^{i,j}\|_2 \le \frac{\eta}{m+1}.$$

So if we define $\tilde{w} \triangleq \sum_{i\in\mathcal{T}} \tilde{w}[i] = \sum_{i\in\mathcal{T}} \sum_j \tilde{w}^{i,j}$, we have that $\|w - \tilde{w}\|_2 \le \eta$.

Now let $\mathcal{N} \triangleq \mathbb{P}(H^{-1}\mathcal{N}')$. As $u = H^{-1}w$ and H^{-1} is an isometry, if we define $\tilde{u}^{i,j} \triangleq H^{-1}\tilde{w}^{i,j}$ and $\tilde{u} \triangleq \sum_{i\in\mathcal{T}} \sum_j \tilde{u}^{i,j}$, then we likewise get that $\|u - \tilde{u}\|_2 \le \eta$, and clearly $\tilde{u}^{i,j} \in \mathcal{N}$, concluding the proof of part 1) of the lemma.

For each $\tilde{u}^{i,j}$, define

$$u_*^{i,j} \triangleq \tilde{u}^{i,j}/\alpha_{i,j} \quad \text{for} \quad \alpha_{i,j} \triangleq s^{-1} \cdot 2^{-(m-i)/2} \|w^{i,j}\|_\infty \tag{4.32}$$

so that

$$\tilde{u} = \sum_{i,j} \alpha_{i,j} u_*^{i,j}.$$

Note that

$$\begin{aligned}\sum_{i,j} \alpha_{i,j} &\leq \frac{1}{s} \sum_i 2^{-(m-i)/2} \sum_j \|w^{i,j}\|_\infty \\ &\leq \frac{1}{s} \sum_i 2^{-(m-i)/2} \|w[i]\|_1 \\ &\leq \frac{1}{s} \cdot s = 1,\end{aligned}$$

where the second inequality follows by the fact that for fixed i, the supports of the vectors $w^{i,j}$ are disjoint for different j so that $\sum_j \|w^{i,j}\|_\infty \leq \|w[i]\|_1$, and the third inequality follows from (4.30). This concludes the proof of part 2) of the lemma.

Finally, we need to bound $\|u_*^{i,j}\|_\infty$. Note first that for any vector z supported only on indices $\nu \in T_i$,

$$\|H^{-1} z\|_\infty \leq 2^{-(m-i)/2} \cdot \|z\|_\infty \tag{4.33}$$

because the Haar wavelets $\{\psi_{i,j}\}_j$ have disjoint supports and L_∞ norm $2^{-(m-i)/2}$. It follows that

$$\begin{aligned}\|\tilde{u}^{i,j}\|_\infty &\leq \|H^{-1} w^{i,j}\|_\infty + \|H^{-1}(w^{i,j} - \tilde{w}^{i,j})\|_\infty \\ &\leq 2^{-(m-i)/2} \cdot \|w^{i,j}\|_\infty + 2^{-(m-i)/2} \|w^{i,j} - \tilde{w}^{i,j}\|_\infty \\ &\leq 2^{-(m-i)/2} \cdot \|w^{i,j}\|_\infty + 2^{-(m-i)/2} \|w^{i,j} - \tilde{w}^{i,j}\|_2 \\ &\leq 2^{-(m-i)/2} \cdot \|w^{i,j}\|_\infty + 2^{-(m-i)/2} \cdot \frac{\eta}{(m+1)\sqrt{n}} \|w^{i,j}\|_2 \\ &\leq 2^{-(m-i)/2} \cdot \|w^{i,j}\|_\infty + 2^{-(m-i)/2} \cdot \frac{\eta}{(m+1)\sqrt{n}} \|w^{i,j}\|_\infty \cdot s \\ &= 2^{-(m-i)/2} \cdot \|w^{i,j}\|_\infty \left(1 + \frac{\eta \cdot s}{(m+1)\sqrt{n}}\right) \\ &\leq 2 \cdot 2^{-(m-i)/2} \cdot \|w^{i,j}\|_\infty,\end{aligned}$$

where the first inequality is triangle inequality, the second inequality follows by (4.33), the third inequality follows from monotonicity of L_p norms, the fourth inequality follows from (4.31), the fifth inequality follows from the fact that $w^{i,j}$ is s^2-sparse, and the final inequality follows from the hypothesis that $\eta \leq 1/s$. Recalling (4.32), we conclude that $\|u_*^{i,j}\|_\infty \leq 2s$ as claimed. □

Next we show that we can control $\frac{1}{m}\sum_{j \in S_G} \langle X_j - \mu_j, u \rangle$ for all directions u in the net $\mathcal{N}$.

Lemma 4.6.12. *Let $\xi > 0$ and let $\mathcal{N} \in \mathbb{P}_{n-1}(\mathbb{R})$ be any collection of M directions. Then*

$$\Pr\left[\frac{1}{m}\sum_{j \in S_G} \langle X_j - \mu_j, u \rangle > \xi \cdot \|u\|_\infty \ \forall u \in \mathcal{N}\right] < 2M \cdot e^{-2m\xi^2},$$

where the probability is over the samples X_j for $j \in S_G$.

Proof. Without loss of generality, assume that $\|u\|_\infty = 1$. For any $j \in S_G$, note that $\|X_j - \mu_j\|_1 \leq 2$, so $\langle X_j - \mu_j, u \rangle$ is a $[-2, 2]$-valued random variable, call it A_j. By Hoeffding's inequality,

$$\Pr\left[\left|\frac{1}{m}\sum_{j \in S_G} A_j - \frac{1}{m}\sum_{j \in S_G} \mathbb{E}[A_j]\right| \geq \xi\right] \leq 2e^{-2m\xi^2},$$

so we are done by a union bound over the M directions in $\mathcal{N}$. □

We may now proceed with the proof of (4.29). For $u \in \mathcal{J}$, by Lemmas 4.6.10 and 4.6.11, there is some $\tilde{u} = \sum_\nu \alpha_\nu u_\nu^*$ such that $u_\nu^* \in \mathcal{N}$ and $\|u - \tilde{u}\|_2 \leq \eta$. We may write

$$\begin{aligned}
\frac{1}{m}\sum_{j \in S_G} \langle X_j - \mu_j, u \rangle &\leq \frac{1}{m}\sum_{j \in S_G} \langle X_j - \mu_j, \tilde{u} \rangle + \|\frac{1}{m}\sum_{j \in S_G} X_j\|_2 \cdot \|u - \tilde{u}\|_2 \\
&\leq \frac{1}{m}\sum_{j \in S_G} \langle X_j - \mu_j, \tilde{u} \rangle + \eta \\
&= \sum_\nu \alpha_\nu \left(\frac{1}{m}\sum_{j \in S_G} \langle X_j - \mu_j, u_\nu^* \rangle\right) + \eta \\
&\leq \sum_\nu \alpha_\nu \cdot \xi \cdot \|u_\nu^*\|_\infty + \eta \\
&\leq 2\xi(\ell \log n + 1)(\log n + 1) + \eta,
\end{aligned}$$

where the second inequality follows from the fact that $\frac{1}{m}\sum_{j \in S_G} X_j$ is a vector in Δ^n and thus has L_2 norm at most 1, and the penultimate step holds with probability $2|\mathcal{N}|e^{-8m\xi^2}$.

So if $\eta = \omega/2$ and $\xi = \frac{\omega}{4(\ell \log n+1)(\log n+1)}$, then as long as

$$m = \Omega(\xi^{-2} \log |\mathcal{N}|) = \Omega\left(\frac{\log(1/\omega)}{\omega} \cdot \ell^4 \log^7 n\right),$$

then with probability at least $1-\text{poly}(n)$, there exists an SoS proof of (4.29) using Axioms 1.

□

Lemma 4.6.13. *Under the canonical assignment, with high probability there is some choice of* $\{(P_S^{T_1,T_2})_{\gamma,\rho}\}$ *and* $\{(Q_S^{T_1,T_2})_{\gamma,\rho}\}$ *for which Constraints 5 and 7 are satisfied.*

The proof of Lemma 4.6.13 is conceptually very similar to that of Lemma 4.6.9, so we defer it to Appendix 4.7.

4.7 Appendix: Proof of Lemma 4.6.13

In the arguments that follow, it will be useful to define the notion of projectivization. Given a set $S \subset \mathbb{R}^m$, let $\mathbb{P}S$ denote its projectivization, namely the quotient of S by the equivalence relation $u \sim v$ if $u = \lambda v$ for some $\lambda \in \mathbb{R}$. We will denote the projectivization of $\mathbb{R}^m$ by $\mathbb{P}_{n-1}(\mathbb{R})$. Occasionally we will abuse notation and implicitly associate $S \subset \mathbb{P}_{n-1}(\mathbb{R})$ with its fiber under the quotient map $\mathbb{R}^n \to \mathbb{P}_{n-1}(\mathbb{R})$.

Proof of Lemma 4.6.13. As in the proof of Lemma 4.4.2, because of Lemma 4.3.3 it is enough to show an SoS proof using Axioms 2 that

$$\frac{1}{m}\sum_{i\in S_G}\langle X_i - \mu_i, v\rangle^t - \frac{1}{m}\sum_{i\in S_G}\mathbb{E}_{X\sim\mathcal{D}_i}\langle X - \mu_i, v\rangle^t \le (8t/k)^{t/2}.$$

Equivalently, we must show that for any degree-t pseudodistribution $\tilde{\mathbb{E}}$ over the variables v and $\mathbf{U}$ which satisfies Axioms 2, we have that

$$\left\langle \mathbf{Z}, \tilde{\mathbb{E}}\left[v^{\otimes t/2}(v^{\otimes t/2})^\top\right]\right\rangle \le (8t/k)^{t/2}, \tag{4.34}$$

where $\mathbf{Z} \triangleq \mathbf{Z}[S_G]$. The set of matrices $\tilde{\mathbb{E}}[v^{\otimes t/2}(v^{\otimes t/2})^\top]$ arising from pseudodistributions $\tilde{\mathbb{E}}$

satisfying Axioms 2 is some convex set $\mathcal{K}$ in $\mathbb{R}^{n^{t/2}\times n^{t/2}}$.

Lemma 4.7.1. *Let $\mathcal{K} \subset \mathbb{R}^{n^{t/2}\times n^{t/2}}$ be the convex set of all matrices of the form $\tilde{\mathbb{E}}[v^{\otimes t/2}(v^{\otimes t/2})^\top]$ for some degree-t pseudodistribution $\tilde{\mathbb{E}}$ over the variables $v, \mathbf{U}$ satisfying Axioms 2.*

Additionally, let $\mathcal{K}_1, \mathcal{K}_2 \subset \mathbb{R}^{n^{t/2}\times n^{t/2}}$ consist of all matrices $\mathbf{M}$ for which $\sum_{\alpha,\beta}\mu^{(\alpha)}\mu^{(\beta)}|\mathbf{M}_{\alpha,\beta}| \leq (\ell\log n+1)^t$ and for which $\|\mathbf{M}\|_F \leq n^{t/2}$ respectively. Then

$$\mathcal{K} \subset \left[(H^{-1})^{\otimes t/2}\right](\mathcal{K}_1\cap\mathcal{K}_2)\left[(H^{-1})^{\otimes t/2}\right]^\top,$$

Proof. Take any $\mathbf{M}\in\mathcal{K}$. We first show that $\mathbf{M}\in\left[(H^{-1})^{\otimes t/2}\right]\mathcal{K}_1\left[(H^{-1})^{\otimes t/2}\right]^\top$. By linearity of $\tilde{\mathbb{E}}$, we may write $\mathbf{M}$ as

$$\mathbf{M} = (H^{-1})^{\otimes t/2}\cdot\tilde{\mathbb{E}}\left[\left[H^{\otimes t/2}v^{\otimes t/2}\right]\cdot\left[H^{\otimes t/2}v^{\otimes t/2}\right]^\top\right]\cdot((H^{-1})^{\otimes t/2})^\top.$$

For any monomials α,β each of degree $t/2$, the second of Axioms 2 immediately implies that

$$-\mathbf{U}_\alpha\mathbf{U}_\beta \leq \tilde{\mathbb{E}}\left[\left[H^{\otimes t/2}v^{\otimes t/2}\right]_\alpha\cdot\left[H^{\otimes t/2}v^{\otimes t/2}\right]_\beta\right] \leq \mathbf{U}_\alpha\mathbf{U}_\beta.$$

We emphasize that this is the only place where we use the second of Axioms 2, and only in a degree-2 fashion, hence Remark 4.6.7. So $\sum_{\alpha,\beta}\mu^{(\alpha)}\mu^{(\beta)}|\mathbf{M}_{\alpha,\beta}| \leq \tilde{\mathbb{E}}\left[\sum_{\alpha,\beta}\mu^{(\alpha)}\mu^{(\beta)}\mathbf{U}_\alpha\mathbf{U}_\beta\right] \leq (\ell logn+1)^t$, where the last inequality follows by axiom 3.

Finally, to show that $\mathbf{M}\in\left[(H^{-1})^{\otimes t/2}\right]\mathcal{K}_2\left[(H^{-1})^{\otimes t/2}\right]^\top$, note first that by orthonormality of H, it is enough to show that $\mathbf{M}\subset\mathcal{K}_2$. But this follows immediately from the fact that $\tilde{\mathbb{E}}$ satisfies the first of Axioms 2. Indeed, from Fact 1.3.44 and the fact that $\tilde{\mathbb{E}}$ is degree-$O(t)$ we get that $-1\leq\tilde{\mathbb{E}}[v_\alpha v_\beta]\leq 1$, so

$$\sum_{|\alpha|,|\beta|=t/2}\mathbf{M}_{\alpha,\beta}^2 = \sum_{\alpha,\beta}\tilde{\mathbb{E}}[v_\alpha v_\beta]^2 \leq n^t$$

as claimed. □

Lemma 4.7.2. *For every $\eta\leq(\ell\log n+1)^{-1}$, there exists a set $\mathcal{N}\subset\mathbb{P}(\mathbb{R}^{n^{t/2}\times n^{t/2}})$ of size $O(n^{3t/2}\log^t n/\eta)^{(\ell\log n+1)^{2t}}$ such that for every $\mathbf{M}\in\left[(H^{-1})^{\otimes t/2}\right](\mathcal{K}_1\cap\mathcal{K}_2)\left[(H^{-1})^{\otimes t/2}\right]^\top$, there* *exists some* $\tilde{\mathbf{M}} = \sum_\nu \alpha_\nu \cdot \mathbf{M}^*_\nu$ *for* $\mathbf{M}^*_\nu \in \mathcal{N}$ *such that 1)* $\|\mathbf{M} - \tilde{\mathbf{M}}\|_F \le \eta$, *2)* $\sum_\nu \alpha_\nu \le 1$, *and 3)* $\|\mathbf{M}^*_\nu\|_{\max} \le 2(\ell \log n + 1)^t$.

Proof. Let $s = \ell \log n + 1$, and let $m = \log n$. Let $\mathcal{N}'$ be an $\frac{\eta}{(m+1)^t n^{t/2}}$-net in Frobenius norm for all s^{2t}-sparse $n^{t/2} \times n^{t/2}$ matrices of unit Frobenius norm. Because $\mathbb{S}^{s^{2t}-1}$ has an $\frac{\eta}{m\sqrt{n}}$-net in L_2 norm of size $(3(m+1)^t n^{t/2}/\eta)^{s^{2t}}$, by a union bound we have that

$$|\mathcal{N}'| \le \binom{n^t}{s^{2t}} \cdot (3(m+1)^t n^{t/2}/\eta)^{s^{2t}} = O(n^{3t/2} \log^t n/\eta)^{s^{2t}}$$

Take any $\mathbf{M} \in \left[(H^{-1})^{\otimes t/2}\right](\mathcal{K}_1 \cap \mathcal{K}_2)\left[(H^{-1})^{\otimes t/2}\right]^\top$ and consider $\mathbf{L} \triangleq H^{\otimes t/2}\mathbf{M}\left[H^{\otimes t/2}\right]^\top$. Define $\mathcal{T} \triangleq \{0_{\text{father}}, 0_{\text{mother}}, 1, ..., m-1\}$. We may write $\mathbf{L}$ as $\sum_{\sigma,\tau} \mathbf{L}[\sigma,\tau]$, where σ, τ are monomials of degree $t/2$ in the indices $\mathcal{T}$, and where $\mathbf{L}[\sigma,\tau]$ the submatrix of $\mathbf{L}$ consisting of all entries from the rows α (resp. columns β) for which $\alpha_i \in T_{\sigma_i}$ (resp. $\beta_i \in T_{\tau_i}$) for all $1 \le i \le t/2$.

As the nonzero entries of $\mathbf{L}[\sigma,\tau]$ are just a subset of those of $\mathbf{L}$, we clearly have $\|\mathbf{L}[\sigma,\tau]\|_F \le n^{t/2}$ for all σ, τ. Moreover, because $\mathbf{L} \in \mathcal{K}_1$, we have that

$$\sum_{\sigma,\tau} \prod_{i=1}^{t/2} 2^{-(m-\sigma_i)/2} \cdot \prod_{j=1}^{t/2} 2^{-(m-\tau_j)/2} \cdot \|\mathbf{L}[\sigma,\tau]\|_{1,1} \le s^t$$

so in particular

$$\|\mathbf{L}[\sigma,\tau]\|_{1,1} \le \prod_{i=1}^{t/2} 2^{(m-\sigma_i)/2} \cdot \prod_{j=1}^{t/2} 2^{(m-\tau_j)/2} \cdot s^t = 2^{-(\sum_i \sigma_i + \sum_j \tau_j)/2} \cdot s^t \cdot n^{t/2}. \tag{4.35}$$

We can thus apply Lemma 4.6.8 to conclude that for each σ, τ, $\mathbf{L}[\sigma,\tau] = \sum_j \mathbf{L}^{\sigma,\tau;j}$ for some matrices $\{\mathbf{L}^{\sigma,\tau;j}\}_j$ of sparsity at most $\lceil 2 \rceil^{-\sum_i \sigma_i - \sum_j \tau_j} \cdot s^{2t} \le s^{2t}$ and for which

$$\sum_j \|\mathbf{L}^{\sigma,\tau;j}\|_F \le n^{t/2}.$$

For each $\mathbf{L}^{\sigma,\tau;j}$, there is some $(\mathbf{L}')^{\sigma,\tau;j} \in \mathcal{N}'$ such that if we define $\tilde{\mathbf{L}}^{\sigma,\tau;j} \triangleq \|\mathbf{L}^{\sigma,\tau;j}\|_F \cdot (\mathbf{L}')^{\sigma,\tau;j}$, then we have

$$\|\mathbf{L}^{\sigma,\tau;j} - \tilde{\mathbf{L}}^{\sigma,\tau;j}\|_F \le \frac{\eta}{(m+1)^t n^{t/2}} \cdot \|\mathbf{L}^{\sigma,\tau;j}\|_F. \tag{4.36}$$

Defining $\tilde{\mathbf{L}}[\sigma,\tau] \triangleq \sum_j \tilde{\mathbf{L}}^{\sigma,\tau;j}$, we get that

$$\|\mathbf{L}[\sigma,\tau] - \tilde{\mathbf{L}}[\sigma,\tau]\|_F \le \frac{\eta}{(m+1)^t n^{t/2}} \sum_j \|\mathbf{L}^{\sigma,\tau;j}\|_F \le \frac{\eta}{(m+1)^t}.$$

So if we define $\tilde{\mathbf{L}} = \sum_{\sigma,\tau} \tilde{\mathbf{L}}[\sigma,\tau] = \sum_{\sigma,\tau}\sum_j \tilde{\mathbf{L}}^{\sigma,\tau;j}$, we have that $\|\mathbf{L} - \tilde{\mathbf{L}}\|_F \le \eta$.

Now let $\mathcal{N} \triangleq (H^{-1})^{\otimes t/2}\mathcal{N}' \left[(H^{-1})^{\otimes t/2}\right]^\top$. As $\mathbf{M} = (H^{-1})^{\otimes t/2}\mathbf{L}\left[(H^{-1})^{\otimes t/2}\right]$ and $(H^{-1})^{\otimes t/2}$ is an isometry, if we define $\tilde{\mathbf{M}}^{\sigma,\tau;j} \triangleq (H^{-1})^{\otimes t/2}\tilde{\mathbf{L}}^{\sigma,\tau;j}\left[(H^{-1})^{\otimes t/2}\right]$ and $\tilde{\mathbf{M}} \triangleq \sum_{\sigma,\tau}\sum_j \tilde{\mathbf{M}}^{\sigma,\tau;j}$, then we likewise get that $\|\mathbf{M} - \tilde{\mathbf{M}}\|_F \le \eta$, and clearly $\tilde{\mathbf{M}}^{\sigma,\tau;j} \in \mathcal{N}$, concluding the proof of part 1) of the lemma.

For each $\tilde{\mathbf{M}}^{\sigma,\tau;j}$, define

$$\mathbf{M}_*^{\sigma,\tau;j} \triangleq \tilde{\mathbf{M}}^{\sigma,\tau;j}/\alpha_{\sigma,\tau;j} \quad \text{for} \quad \alpha_{\sigma,\tau;j} \triangleq s^{-t} \cdot \prod_{i=1}^{t/2} 2^{-(m-\sigma_i)/2} \cdot \prod_{j=1}^{t/2} 2^{-(m-\tau_j)/2} \|\mathbf{L}^{\sigma,\tau;j}\|_{\max} \tag{4.37}$$

so that

$$\mathbf{M} = \sum_{\sigma,\tau,j} \alpha_{\sigma,\tau;j}\mathbf{M}_*^{\sigma,\tau;j}.$$

Note that

$$\begin{aligned}\sum_{\sigma,\tau,j}\alpha_{\sigma,\tau;j} &\le \frac{1}{s^t}\sum_{\sigma,\tau}\prod_{i=1}^{t/2} 2^{-(m-\sigma_i)/2} \cdot \prod_{j=1}^{t/2} 2^{-(m-\tau_j)/2} \sum_j \|\mathbf{L}^{\sigma,\tau;j}\|_{\max} \\ &\le \frac{1}{s^t}\sum_{\sigma,\tau}\prod_{i=1}^{t/2} 2^{-(m-\sigma_i)/2} \cdot \prod_{j=1}^{t/2} 2^{-(m-\tau_j)/2} \|\mathbf{L}[\sigma,\tau]\|_{1,1} \\ &\le \frac{1}{s^t} \cdot s^t = 1,\end{aligned}$$

where the second inequality follows by the fact that for fixed σ,τ, the supports of the matrices $\mathbf{L}^{\sigma,\tau;j}$ are disjoint for different j so that $\sum_j \|\mathbf{L}^{\sigma,\tau;j}\|_{\max} \le \|\mathbf{L}[\sigma,\tau]\|_{1,1}$, and the third inequality follows from (4.35). This concludes the proof of part 2) of the lemma.

Finally, we need to bound $\|\mathbf{M}_*^{\sigma,\tau;j}\|_{\max}$. Note first that for any matrix $\mathbf{J}$ supported only on the support of some $\mathbf{L}[\sigma, \tau]$,

$$\|(H^{-1})^{\otimes t/2}\mathbf{J}\left[(H^{-1})^{\otimes t/2}\right]\|_{\max} \leq \prod_{i=1}^{t/2} 2^{-(m-\sigma_i)/2} \cdot \prod_{j=1}^{t/2} 2^{-(m-\tau_j)/2} \cdot \|\mathbf{J}\|_{\max} \tag{4.38}$$

because the tensored Haar wavelets $\{\psi_{\sigma_1,j_1} \otimes \cdots \otimes \psi_{\sigma_{t/2},j_{t/2}}\}_{j_1,\ldots,j_{t/2}}$ (resp. $\{\psi_{\tau_1,j_1} \otimes \cdots \otimes \psi_{\tau_{t/2},j_{t/2}}\}_{j_1,\ldots,j_{t/2}}$) have disjoint supports and max-norm $\prod_{i=1}^{t/2} 2^{-(m-\sigma_i)/2}$ (resp. $\prod_{j=1}^{t/2} 2^{-(m-\tau_j)/2}$). It follows that

$$\begin{aligned}
\|\tilde{\mathbf{M}}^{\sigma,\tau;j}\|_{\max} &\leq \|(H^{-1})^{\otimes t/2}\mathbf{L}^{\sigma,\tau;j}\left[(H^{-1})^{\otimes t/2}\right]^\top\|_{\max} + \|(H^{-1})^{\otimes t/2}\left(\mathbf{L}^{\sigma,\tau;j} - \tilde{\mathbf{L}}^{\sigma,\tau;j}\right)\left[(H^{-1})^{\otimes t/2}\right]^\top\|_{\max} \\
&\leq \prod_{i=1}^{t/2} 2^{-(m-\sigma_i)/2} \prod_{j=1}^{t/2} 2^{-(m-\tau_j)/2} \cdot \left(\|\mathbf{L}^{\sigma,\tau;j}\|_{\max} + \|\mathbf{L}^{\sigma,\tau;j} - \tilde{\mathbf{L}}^{\sigma,\tau;j}\|_{\max}\right) \\
&\leq \prod_{i=1}^{t/2} 2^{-(m-\sigma_i)/2} \prod_{j=1}^{t/2} 2^{-(m-\tau_j)/2} \cdot \left(\|\mathbf{L}^{\sigma,\tau;j}\|_{\max} + \|\mathbf{L}^{\sigma,\tau;j} - \tilde{\mathbf{L}}^{\sigma,\tau;j}\|_{F}\right) \\
&\leq \prod_{i=1}^{t/2} 2^{-(m-\sigma_i)/2} \prod_{j=1}^{t/2} 2^{-(m-\tau_j)/2} \cdot \left(\|\mathbf{L}^{\sigma,\tau;j}\|_{\max} + \frac{\eta}{(m+1)^t n^{t/2}}\|\mathbf{L}^{\sigma,\tau;j}\|_{F}\right) \\
&\leq \prod_{i=1}^{t/2} 2^{-(m-\sigma_i)/2} \prod_{j=1}^{t/2} 2^{-(m-\tau_j)/2} \cdot \left(\|\mathbf{L}^{\sigma,\tau;j}\|_{\max} + \frac{\eta}{(m+1)^t n^{t/2}}\|\mathbf{L}^{\sigma,\tau;j}\|_{\max} \cdot s^t\right) \\
&= \prod_{i=1}^{t/2} 2^{-(m-\sigma_i)/2} \prod_{j=1}^{t/2} 2^{-(m-\tau_j)/2} \cdot \|\mathbf{L}^{\sigma,\tau;j}\|_{\max} \cdot \left(1 + \frac{\eta \cdot s^t}{(m+1)^t n^{t/2}}\right) \\
&\leq 2 \cdot \prod_{i=1}^{t/2} 2^{-(m-\sigma_i)/2} \prod_{j=1}^{t/2} 2^{-(m-\tau_j)/2} \cdot \|\mathbf{L}^{\sigma,\tau;j}\|_{\max},
\end{aligned}$$

where the first inequality is triangle inequality, the second inequality follows by (4.38), the third inequality follows from monotonicity of L_p norms, the fourth inequality follows from (4.36), the fifth inequality follows from the fact that $\mathbf{L}^{\sigma,\tau;j}$ is s^{2t} sparse, and the final inequality follows from the hypothesis that $\eta \leq s^{-t}$. Recalling (4.37), we conclude that $\|\mathbf{M}_*^{\sigma,\tau;j}\|_{\max} \leq 2s^t$ as claimed. □

Next we show that we can control $\langle \mathbf{Z}, \mathbf{M} \rangle$ for all directions in the net $\mathcal{N}$.

Lemma 4.7.3. *Let $\xi > 0$ and let $\mathcal{N} \in \mathbb{P}\left(\mathbb{R}^{n^{t/2} \times n^{t/2}}\right)$ be any collection of M directions. Then*

$$\Pr[\langle \mathbf{Z}, \mathbf{M} \rangle > \xi \cdot \|\mathbf{M}\|_{\max} \ \forall \ \mathbf{M} \in \mathcal{N}] < 2M \cdot e^{-8m\xi^2},$$

where the probability is over the samples X_j for $j \in S_G$.

Proof. Without loss of generality, assume that $\|\mathbf{M}\|_{\max} = 1$. For any $j \in S_G$, note that the sum of the absolute values of the entries of the matrix $\left[(X - \mu_i)^{\otimes t/2}\right]\left[(X - \mu_i)^{\otimes t/2}\right]^\top$ is $\left(\sum_\alpha |(X - \mu_i)_\alpha|\right)^2 \le \left(\sum_\alpha X_\alpha + \sum_\alpha (\mu_i)_\alpha\right)^2 \le 4$. So for any $j \in S_G$ and $\mathbf{M} \in \mathcal{N}$,

$$\left\langle \left[(X_i - \mu_i)^{\otimes t/2}\right]\left[(X_i - \mu_i)^{\otimes t/2}\right]^\top, \mathbf{M} \right\rangle$$

is a $[-4, 4]$-valued random variable, call it A_i. By Hoeffding's inequality,

$$\Pr\left[\left|\frac{1}{m}\sum_{i \in S_G} A_i - \frac{1}{m}\sum_{i \in S_G} \mathbb{E}[A_i]\right| \ge \xi\right] \le 2e^{-8m\xi^2},$$

so we are done by a union bound over the M directions in $\mathcal{N}$ □

We may now proceed with the proof of (4.34). For $\mathbf{M} \in \mathcal{K}$, by Lemmas 4.7.1 and 4.7.2, there is some $\tilde{\mathbf{M}} = \sum_\nu \alpha_\nu \mathbf{M}^*_\nu$ such that $\mathbf{M}^*_\nu \in \mathcal{N}$ and $\|\mathbf{M} - \tilde{\mathbf{M}}\| \le \eta$. We may write

$$\begin{aligned}
\langle \mathbf{Z}, \mathbf{M} \rangle &\le \langle \mathbf{Z}, \tilde{\mathbf{M}} \rangle + \|\mathbf{Z}\|_F \|\mathbf{M} - \tilde{\mathbf{M}}\|_F \\
&\le \langle \mathbf{Z}, \tilde{\mathbf{M}} \rangle + \eta \cdot \|\mathbf{Z}\|_F \\
&= \sum_\nu \alpha_\nu \langle \mathbf{Z}, \mathbf{M}^*_\nu \rangle + \eta \cdot \|\mathbf{Z}\|_F \\
&\le \sum_\nu \alpha_\nu \cdot \xi \cdot \|\mathbf{M}^*_\nu\|_{\max} + \eta \cdot \|\mathbf{Z}\|_F \\
&\le 2\xi(\ell \log n + 1)^t + \eta \cdot \|\mathbf{Z}\|_F.
\end{aligned}$$

where the penultimate step holds with probability $2|\mathcal{N}|e^{-8m\xi^2}$. But observe that because $\|\mu_i\|_\infty \le 1$ for all $i \in [N]$, we have the simple bound that for any $X \in \Delta^n$ and any $i \in [n]$,

$$\|\left[(X - \mu_i)^{\otimes t/2}\right]\left[(X - \mu_i)^{\otimes t/2}\right]^\top\|_F = \|X - \mu_i\|_2^2 \le 2,$$

from which we conclude by triangle inequality that $\|\mathbf{Z}\|_F \leq 4$.

We conclude that $\langle \mathbf{Z}, \mathbf{M} \rangle \leq 2\xi(\ell \log n + 1)^t + 4\eta$, so if $\eta = \frac{1}{8}(8t/k)^{t/2}$ and $\xi = \frac{(8t/k)^{t/2}}{4(\ell \log n+1)^t}$, then as long as

$$m = \Omega(\xi^{-2} \log |\mathcal{N}|) = \Omega\left(\ell^4 \log^4 n\right)^t \cdot \frac{k^t}{t^{t-1}} \cdot \log(nk/t),$$

then with probability at least $1-\mathrm{poly}(n)$, there exists an SoS proof of (4.34) using Axioms 2.

□

Chapter 5

Learning From Untrusted Batches With Alternating Minimization

5.1 Introduction

We now show how to improve upon the guarantees of the preceding chapter by replacing sum-of-squares programming with alternating minimization. As we mentioned in Section 4.1.1, in the case of general, unstructured distributions μ, in a work concurrent with and independent of ours from the previous chapter, Jain and Orlitsky [JO19] gave a polynomial time algorithm based on a much simpler semidefinite program that estimates μ to within the same total variation distance as our Theorem 4.4.1. Their approach was based on an elegant way to combine approximation algorithms for the cut-norm [AN04] with the filtering approach for robust estimation [DKK+19a, SCV18, DKK+17, DKK+19b, DHL19].

An appealing aspect of our relaxation from the previous chapter however was that it was possible to incorporate shape-constraints into the relaxation, through the Haar wavelet basis, which allowed us to improve the sample complexity to quasipolynomial in d and s, respectively the degree and number of parts in the piecewise polynomial approximation, and quasipolylogarithmic in n. Unfortunately, while [JO19] achieves better runtime and sample complexity in the unstructured setting, their techniques do not obviously extend to obtain a similar sample complexity under structural assumptions.

This raises a natural question: can we build on [JO19] and the tools from the previous chapter, to incorporate shape constraints into a simple semidefinite programming approach, that can achieve nearly-optimal robustness, in polynomial runtime, and with sample complexity which is sublinear in n? In this chapter, we answer this question in the affirmative:

Theorem 5.1.1 (Informal, see Theorem 5.4.1)**.** *Let μ be a distribution over $[n]$ that is approximated by an s-part piecewise polynomial function with degree at most d. Then there is a polynomial-time algorithm which estimates μ to within*

$$O\left(\omega + \frac{\varepsilon}{\sqrt{k}}\sqrt{\log 1/\varepsilon}\right)$$

in total variation distance after drawing N ε-corrupted batches, each of size k, where

$$N = \widetilde{O}\left((s^2d^2/\varepsilon^2)\cdot\log^3(n)\right)$$

is the number of batches needed.

5.1.1 High-Level Argument

In the discussion in this section, we will specialize to the case of $\omega = 0$ for the sake of clarity.

Learning via Filtering Recall from Section 4.2 that we can easily view learning from untrusted batches as robust mean estimation of multinomial distributions in L_1 distance: given a batch of samples $Y_i = (Y_i^1, ..., Y_i^k)$ from a distribution μ over $[n]$, the frequency vector $\{\frac{1}{k}\sum_{j=1}^k \mathbb{1}[Y_i^j = a]\}_{a\in[n]}$ is distributed according to the normalized multinomial distribution $\mathrm{Mul}_k(\mu)$ given by k draws from μ. Note that μ is precisely the mean of $\mathrm{Mul}_k(\mu)$, so the problem of estimating μ from an ε-corrupted set of N frequency vectors is equivalent to that of robustly estimating the mean of a multinomial distribution.

As such, it is natural to try to adapt the existing algorithms for robust mean estimation of other distributions; the fastest of these are based on a simple filtering approach which works as follows. We maintain weights for each point, initialized to uniform. At every step, we measure the maximum "skew" of the weighted dataset in any direction, and if this skew is still too high, update the weights by

1. Finding the direction v in which the corruptions "skew" the dataset the most.
2. Giving a "score" to each point based on how badly it skews the dataset in the direction v
3. Downweighting or removing points with high scores.

Otherwise, if the skew is low, output the empirical mean of the weighted dataset.

To prove correctness of this procedure, one must show three things for the particular skewness measure and score function chosen:

- **Regularity**: For any sufficiently large collection of ε-corrupted samples, a particular deterministic regularity condition holds (Definition 5.4.3 and Lemma 5.4.6)
- **Soundness**: Under the regularity condition, if the skew of the weighted dataset is small, then the empirical mean of the weighted dataset is sufficiently close to the true mean (Lemma 5.4.7).
- **Progress**: Under the regularity condition, if the skew of the weighted dataset is large, then one iteration of the above update scheme will remove more weight from the bad samples than from the good samples (Lemma 5.4.10).

For isotropic Gaussians, skewness is just given by the maximum variance of the weighted dataset in any direction, i.e. $\max_{v \in \mathbb{S}^{n-1}} \langle vv^\top, \tilde{\Sigma} \rangle$ where $\tilde{\Sigma}$ is the empirical covariance of the weighted dataset. Given maximizing v, the "score" of a point X is then simply its contribution to the skewness.

To learn in L_1 distance, the right set of test vectors v to use is the Hamming cube $\{0,1\}^n$, so a natural attempt at adapting the above skewness measure to robust mean estimation of multinomials is to consider the quantity $\max_{v \in \{0,1\}^n} \langle vv^\top, \tilde{\Sigma} \rangle$. But one of the key challenges in passing from isotropic Gaussians to multinomial distributions is that this quantity above is not very informative because we do not have a good handle on the covariance of $\mathrm{Mul}_k(\mu)$. In particular, it could be that for a direction v, $\langle vv^\top, \tilde{\Sigma} \rangle$ is high simply because the good points have high variance to begin with.

The Jain-Orlitsky Correction Term The clever workaround of [JO19] was to observe that we know exactly what the projection of a multinomial distribution $\mathrm{Mul}_k(\mu)$ in any $\{0,1\}^n$ direction v is, namely $\mathrm{Bin}(k, \langle v, \mu\rangle)$. And so to discern whether the corrupted points skew our estimate in a given direction v, one should measure not the variance in the direction v, but rather the following *corrected quantity*: the variance in the direction v, *minus* what the variance would be if the distribution of the projections in the v direction were actually given by $\mathrm{Bin}(k, \langle v, \tilde{\mu}\rangle)$, where $\tilde{\mu}$ is the empirical mean of the weighted dataset. This new skewness measure can be written as

$$\max_{v\in\{0,1\}^n} \left\{\langle vv^\top, \tilde{\Sigma}\rangle - \frac{1}{k}(\langle v, \tilde{\mu}\rangle - \langle v, \tilde{\mu}\rangle^2)\right\}. \tag{5.1}$$

Finding the direction $v \in \{0,1\}^n$ which maximizes this corrected quantity is some Boolean quadratic programming problem which can be solved approximately by solving the natural SDP relaxation and rounding to a Boolean vector v using the machinery of [AN04]. Using this approach, [JO19] obtained a polynomial-time algorithm for learning *general* discrete distributions from untrusted batches.

Learning Structured Distributions Recall from the previous chapter that it suffices to be able to learn with respect to the $\mathcal{A}_K$ norm, and the key difficulty we had to address was that unlike the Hamming cube or $\mathbb{S}^{n-1}$, it is unclear how to optimize over the set of test vectors dual to the $\mathcal{A}_K$ norm. Our main observation was that vectors with few sign changes admit sparse representations in the *Haar wavelet basis*, so instead of working with $\mathcal{V}^n_{2K}$, one can simply work with a convex relaxation of this Haar-sparsity constraint. As such, if we let $\mathcal{K} \subseteq \mathbb{R}^{n\times n}$ denote the relaxation of the set of $\{vv^\top | v \in \mathcal{V}^n_{2K}\}$ to all matrices Σ whose Haar transforms are "analytically sparse" in some appropriate, convex sense (see Section 5.3 for a formal definition), then as this set of test matrices contains the set of test matrices $vv^\top$ for $v \in \mathcal{V}^n_{2K}$, it is enough to learn μ in the norm associated to $\mathcal{K}$, which is strictly stronger than the $\mathcal{A}_K$ norm.[1]

Our goal then is to produce $\hat{\mu}$ for which $\|\hat{\mu} - \mu\|_{\mathcal{K}} \triangleq \sup_{\Sigma\in\mathcal{K}}\langle \Sigma, (\hat{\mu} - \mu)^{\otimes 2}\rangle^{1/2}$ is small.

[1] Note that in the previous chapter, because moment bounds beyond degree 2 were used, we also needed to use higher-order tensor analogues of $\mathcal{K}$, but in this chapter it will suffice to work with degree 2.

And even though $\|\cdot\|_{\mathcal{K}}$ is a stronger norm, it turns out that the metric entropy of $\mathcal{K}$ is still small enough that one can get good sample complexity guarantees. Indeed, showing that this is the case (see Lemma 5.6.1) was where the bulk of the technical machinery of the previous chapter went, and as we elaborate on in Appendix 5.7, the analysis there left some room for tightening. In this chapter, we give a refined analysis of $\mathcal{K}$ which allows us to get nearly tight sample complexity bounds.

Putting Everything Together Almost all of the pieces are in place to instantiate the filtering framework: in lieu of the quantity in (5.1), which can be phrased as the maximization of some quadratic $\langle vv^\top, M(w)\rangle$ over $\{\pm 1\}^n$, where $M(w) \in \mathbb{R}^{n\times n}$ depends on the dataset and the weights w on its points,[2] we can define our skewness measure as $\max_{\Sigma\in\mathcal{K}}\langle \Sigma, M(w)\rangle = \|M(w)\|_{\mathcal{K}}$, and we can define the score for each point in the dataset to be its contribution to the skewness measure (see Section 5.4.2).

At this point the reader may be wondering why we never round Σ to an actual vector $v \in \mathcal{V}^n_{2K}$ before computing skewness and scores. As our subsequent analysis will show, it turns out that rounding is unnecessary, both in our setting and even in the unstructured distribution setting considered in [JO19]. Indeed, if one examines the three proof ingredients of regularity, soundness, and progress that we enumerated above, it becomes evident that the filtering framework for robust mean estimation does not actually require finding a concrete direction in $\mathbb{R}^n$ in which to filter, merely a skewness measure and score functions which are amenable to showing the above three statements. That said, as we will see, it becomes more technically challenging to prove these ingredients when Σ is not rounded to an actual direction (see e.g. the discussion after Lemmas 5.6.2 and 5.6.3 in Appendix 5.6), though nevertheless possible. We hope that this observation will prove useful in future applications of filtering

[2] Note that we have switched to $\{\pm 1\}^n$ in place of $\{0, 1\}^n$. We do not belabor this point here, as the difference turns out to be immaterial, and the former is more convenient for understanding how we handle $\mathcal{V}^n_{2K}$, which is a subset of $\{\pm 1\}^n$.

5.1.2 Concurrent and Subsequent Work

Concurrently and independently of the work in this chapter, a newer work of Jain and Orlitsky [JO20] obtains very similar results, though our quantitative guarantees are incomparable: the number of batches N they need scales linearly in $s \cdot d$ and independently of n, but also scales with $\sqrt{k}$ and $1/\varepsilon^3$. Finally, we remark that in a very recent follow-up to these two works, Jain and Orlitsky [JO21] managed to answer the remaining open question of achieving the tight sample complexity scaling as $s \cdot d/\varepsilon^2$.

Roadmap In Section 5.2, we overview notation and give miscellaneous technical tools. In Section 5.3, we define the semidefinite program that we use to compute skewness. In Section 5.4, we give our algorithm LearnWithFilter and prove our main result, Theorem 5.1.1. In Section 5.5, we describe our empirical evaluations of LearnWithFilter on synthetic data. In Appendices 5.6, 5.7, and 5.8, we complete the proofs of some deferred technical statements relating to deterministic regularity conditions and metric entropy bounds.

5.2 Technical Preliminaries

5.2.1 Weights, Means, and Covariances

Given samples $X_1, \cdots, X_N \sim \mathrm{Mul}_k(\mu)$ and $U \subseteq [N]$, define $w(U) : [N] \to [0, 1/N]$ to be the set of weights which assigns $1/N$ to all points in U and 0 to all other points. Also define its normalization $\hat{w}(U) \triangleq w(U)/\|w\|_1$. Let $\mathcal{W}_\varepsilon$ denote the set of weights $w : [N] \to [0, 1/N]$ which are convex combinations of such weights for $|U| \geq (1-\varepsilon)N$.

Given w, define $\mu(w) \triangleq \sum_{i=1}^N \frac{w_i}{\|w\|_1} X_i$, and define $\mu(U) \triangleq \mu(w(U))$, that is, the empirical mean of the samples indexed by U.

Given samples $X_1, \cdots, X_N \sim \mathrm{Mul}_k(\mu)$, weights w, and $\nu_1, ..., \nu_N \in \Delta^n$, define the matrices

$$A(w, \{\nu_i\}) = \sum_{i=1}^N w_i (X_i - \nu_i)^{\otimes 2} \qquad \text{and} \qquad B(\{\nu_i\}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{X \sim \mathrm{Mul}_k(\nu_i)}[(X - \nu_i)^{\otimes 2}].$$

When $\nu_1 = \cdots = \nu_N = \nu$, denote these matrices by $A(w, \nu)$ and $B(\nu)$ and note that

$$B(\nu) = \frac{1}{k}\left(\operatorname{diag}(\nu) - \nu^{\otimes 2}\right). \tag{5.2}$$

Also define $M(w, \{\nu_i\})) \triangleq A(w, \{\nu_i\}) - B(\{\nu_i\})$ and $M(w, \nu) \triangleq A(w, \nu) - B(\nu)$. We will also denote $M(w, \mu(w))$ by $M(w)$ and $M(\hat{w}(U))$ by M_U.

To get intuition for these definitions, note that any bitstring $v \in \{0, 1\}^n$ corresponding to $S \subseteq [n]$ induces the normalized binomial distribution $Y \triangleq \mathrm{Bin}(n, \langle \mu, v\rangle) \in [0, 1]$, and any sample $X_i \sim \mathrm{Mul}_k(\mu)$ induces a corresponding sample $\langle X_i, v\rangle$ from Y. Then $\langle vv^\top, M_U\rangle$ is the difference between the empirical variance of Y and the variance of the binomial distribution $\mathrm{Bin}(n, \langle \mu(U), v\rangle)$.

5.2.2 Some Elementary Facts

In this section we collect miscellaneous elementary facts that will be useful in subsequent sections.

Fact 5.2.1. *For $X_1, \cdots, X_m \in \mathbb{R}^n$, weights $w : [m] \to \mathbb{R}_{\geq 0}$, $v \in \mathbb{R}^n$, $\mu \in \mathbb{R}^n$, and $\Sigma \in \mathbb{R}^{n\times n}$ symmetric,*

$$\sum w_i \left\langle (X_i - \mu)^{\otimes 2}, \Sigma\right\rangle = \sum w_i \left\langle (X_i - \mu(w))^{\otimes 2}, \Sigma\right\rangle + \|w\|_1 \cdot \left\langle (\mu(w) - \mu)^{\otimes 2}, \Sigma\right\rangle. \tag{5.3}$$

In particular, by taking $\Sigma = vv^\top$ for any $v \in \mathbb{R}^n$,

$$\sum w_i \langle X_i - \mu, v\rangle^2 = \sum w_i \langle X_i - \mu(w), v\rangle^2 + \|w\|_1 \cdot \langle \mu(w) - \mu, v\rangle^2.$$

That is, the function $\nu \mapsto \sum_i w_i \langle X_i - \nu, v\rangle^2$ is minimized over $\nu \in \mathbb{R}^n$ by $\nu = \mu(w)$.

Proof. Without loss of generality we may assume $\|w\|_1 = 1$. Using the fact that $\langle u^{\otimes 2}, \Sigma\rangle - \langle v^{\otimes 2}, \Sigma\rangle = (u - v)^\top \Sigma(u + v)$ for symmetric Σ, we see that

$$\left\langle (X_i - \mu^{\otimes 2} - (X_i - \mu(w))^{\otimes 2}, \Sigma\right\rangle = (\mu(w) - \mu)^\top \Sigma(2X_i - \mu - \mu(w)).$$

Because $\sum w_i X_i = \mu(w)$, we see that

$$\sum w_i(\mu(w) - \mu)^\top \Sigma(2X_i - \mu - \mu(w)) = \left\langle (\mu(w) - \mu)^{\otimes 2}, \Sigma \right\rangle ,$$

from which (5.3) follows. The remaining parts of the claim follow trivially. □

Fact 5.2.2. *For any $0 < \varepsilon < 1$, let weights $w : [N] \to [0, 1/N]$ satisfy $\sum_{i \in [N]} w_i \geq 1 - O(\varepsilon)$. If w' is the set of weights defined by $w'_i = w_i$ for $i \in S_G$ and $w'_i = 0$ otherwise, and if $|S_G| \geq (1 - \varepsilon)N$, then we have that $\|\mu(w) - \mu(w')\|_1 \leq O(\varepsilon)$.*

Proof. We may write

$$\begin{aligned}\|\mu(w) - \mu(w')\|_1 &\leq \|\frac{1}{\|w\|_1} \sum_{i \in S_B} w_i X_i\|_1 + \left(\frac{1}{\|w\|_1} - \frac{1}{\|w'\|_1}\right) \|\sum_{i \in S_G} w_i X_i\|_1 \\ &\leq O(\varepsilon) + \left(\frac{1}{\|w\|_1} - \frac{1}{\|w'\|_1}\right) \|\sum_{i \in S_G} w_i X_i\|_1 \leq O(\varepsilon),\end{aligned}$$

where the first step follows by definition of $\mu(\cdot)$ and by triangle inequality, the second step follows by the fact that $|S_B| \leq \varepsilon N$, and the third step follows by the fact that $|\|w\|_1 - \|w'\|_1| = \left|\sum_{i \in S_B} w_i\right| \leq \varepsilon$, while $\|\sum_{i \in S_G} w_i X_i\|_1 \leq 1$ as the samples X_i lie in Δ^n. □

It will be useful to have a basic bound on the Frobenius norm of $M(w, \nu)$.

Lemma 5.2.3. *For any $\nu \in \Delta^n$ and any weights w for which $\sum w_i = 1$, we have that $\|M(w, \nu)\|_F \leq 3$.*

Proof. For any sample $X \in \Delta^n$, we have that

$$\|(X - \nu)(X - \nu)^\top\|_F \leq \|X - \nu\|_2^2 \leq 2$$

and

$$\|B(\nu)\|_F \leq \frac{1}{k}\|\nu\|_2 + \frac{1}{k}\|\nu\|_2^2 \leq 2/k,$$

from which the lemma follows by triangle inequality and the assumption that $\sum w_i = 1$. □

5.2.3 Haar Wavelets Revisited

In the analysis in this chapter, it will be useful to introduce the following notation for the Haar wavelet basis, previously introduced in Section 4.6.3.

- For $\nu \in [n]$, if the ν-th element of the Haar wavelet basis for $\mathbb{R}^n$ is some $\psi_{i,j}$, then define the weight $\mathbf{h}^{(\nu)} \triangleq 2^{-(m-i)/2}$.

- For any index $i \in \{0_{\text{father}}, 0_{\text{mother}}, 1, \cdots, m-1\}$, let $T_i \subset [n]$ denote the set of indices ν for which the ν-th Haar wavelet is of the form $\psi_{i,j}$ for some j.

- Given any $p \geq 1$, define the *Haar-weighted L^p norm* $\|\cdot\|_{p;\mathbf{h}}$ on $\mathbb{R}^n$ by $\|w\|_{p;\mathbf{h}} \triangleq \|w'\|_p$, where for every $a \in [n]$, $w'_a \triangleq \mathbf{h}^{(a)} w_a$. Likewise, given any norm $\|\cdot\|_*$ on $\mathbb{R}^{n\times n}$, define the Haar-weighted $*$-norm $\|\cdot\|_{*;\mathbf{h}}$ on $\mathbb{R}^{n\times n}$ by $\|\mathbf{M}\|_{*;\mathbf{h}} \triangleq \|\mathbf{M}'\|_*$, where for every $a, b \in [n]$, $\mathbf{M}'_{a,b} \triangleq \mathbf{h}^{(a)}\mathbf{h}^{(b)}\mathbf{M}_{a,b}$.

In this notation, we obtain the following version of Lemma 4.6.6:

Lemma 5.2.4. *Let $v \in \{\pm 1\}^n$ have at most ℓ sign changes. Then Hv has at most $\ell \log n + 1$ nonzero entries, and furthermore $\|Hv\|_{\infty;\mathbf{h}} \leq 1$. In particular, $\|Hv\|^2_{2;\mathbf{h}}, \|Hv\|_{1;\mathbf{h}} \leq \ell \log n + 1$.*

Proof. We first show that Hv has at most $\ell \log n + 1$ nonzero entries. For any $\psi_{i,j}$ with nonzero entries at indices $[a, b] \subset [n]$ and such that $i \neq 0_{\text{father}}$, if v has no sign change in the interval $[a, b]$, then $\langle \psi_{i,j}, v\rangle = 0$. For every index $\nu \in [n]$ at which v has a sign change, there are at most $m = \log n$ choices of i, j for which $\psi_{i,j}$ has a nonzero entry at index ν, from which the claim follows by a union bound over all ℓ choices of ν, together with the fact that $\langle \psi_{0_{\text{father}},0}, v\rangle$ may be nonzero.

Now for each (i, j) for which $\langle \psi_{i,j}, v\rangle \neq 0$, note that

$$2^{-(m-i)/2} \cdot |\langle \psi_{i,j}, v\rangle| \leq 2^{-(m-i)/2} \cdot \left(2^{-(m-i)/2} \cdot 2^{m-i}\right) = 1,$$

as claimed. The bounds on $\|Hv\|_{1;\mathbf{h}}, \|Hv\|^2_{2;\mathbf{h}}$ follow immediately. □

5.3 SDP for Finding the Direction of Largest Variance

Recall that in [JO19], the authors consider the binary optimization problem $\max_{v\in\{0,1\}^n}|v^\top M_U v|$. We would like to approximate the optimization problem $\max_{v\in\mathcal{V}_\ell^n}|v^\top M_U v|$. Motivated by the sum-of-squares relaxation from the previous chapter and Lemma 5.2.4, we consider the following convex relaxation:

Definition 5.3.1. *Let ℓ be given by (4.21). Let $\mathcal{K}$ denote the (convex) set of all matrices $\Sigma\in\mathbb{R}^{n\times n}$ for which*

1. $\|\Sigma\|_{\max}\le 1$.
2. $\|H\Sigma H^\top\|_{1,1;\mathbf{h}}\le \ell\log n+1$.
3. $\|H\Sigma H^\top\|^2_{F;\mathbf{h}}\le \ell\log n+1$.
4. $\|H\Sigma H^\top\|_{\max;\mathbf{h}}\le 1$.
5. $\Sigma\succeq 0$.

Let $\|\cdot\|_{\mathcal{K}}$ denote the associated norm given by $\|\mathbf{M}\|_{\mathcal{K}}\triangleq\sup_{\Sigma\in\mathcal{K}}|\langle\mathbf{M},\Sigma\rangle|$. By abuse of notation, for vectors $v\in\mathbb{R}^n$ we will also use $\|v\|_{\mathcal{K}}$ to denote $\|vv^\top\|^{1/2}_{\mathcal{K}}$.

Because $\mathcal{K}$ has an efficient separation oracle, one can compute $\|\cdot\|_{\mathcal{K}}$ in polynomial time.

Remark 5.3.2. *Note that, besides not being a sum-of-squares program like the one considered in the previous chapter, this relaxation is also slightly different because of Constraints 3 and 4. As we will see in Section 5.7, these additional constraints will be crucial for getting refined sample complexity bounds.*

Note that Lemma 5.2.4 immediately implies that $\mathcal{K}$ is a relaxation of $\mathcal{V}_\ell^n$:

Corollary 5.3.3 (Corollary of Lemma 5.2.4)**.** *$vv^\top\in\mathcal{K}$ for any $v\in\mathcal{V}_\ell^n$.*

Note also that Constraint 1 in Definition 5.3.1 ensures that $\|\cdot\|_{\mathcal{K}}$ is weaker than $\|\cdot\|_1$ and more generally that:

Fact 5.3.4. *For any $a,b\in\mathbb{R}^n$ and $\Sigma\in\mathcal{K}$, $a^\top\cdot\Sigma\cdot b\le\|a\|_1\cdot\|b\|_1$. In particular, for any $v\in\mathbb{R}^n$, $\|v\|_{\mathcal{K}}\le\|v\|_1$.*

As a consequence, we conclude the following useful fact about stability of the $B(\cdot)$ matrix.

Corollary 5.3.5. *For any $\mu, \mu' \in \Delta^n$, $\|B(\mu) - B(\mu')\|_{\mathcal{K}} \leq \frac{3}{k}\|\mu - \mu'\|_1$.*

Proof. Take any $\Sigma \in \mathcal{K}$. By symmetry, it is enough to show that $\langle B(\mu) - B(\mu'), \Sigma\rangle \leq \frac{3}{k}\|\mu - \mu'\|_1$. By Constraint 1, we have that $\langle \mu - \mu', \mathrm{diag}(\Sigma)\rangle \leq \|\mu - \mu'\|_1$. On the other hand, note that

$$\mu'^{\top}\Sigma\mu' - \mu^{\top}\Sigma\mu = (\mu' - \mu)^{\top}\Sigma(\mu' + \mu) \leq \|\mu' - \mu\|_1 \cdot \|\mu' + \mu\|_1 \leq 2\|\mu' - \mu\|_1,$$

where the second step follows from Fact 5.3.4. The corollary now follows. □

Note that if the solution to the convex program $\mathrm{argmax}_{\Sigma\in\mathcal{K}}\langle M_U, \Sigma\rangle$ were actually integral, that is, some rank-1 matrix $vv^{\top}$ for $v \in \mathcal{V}_\ell^n$, it would correspond to the direction v in which the samples in U have the largest discrepancy between the empirical variance and the variance predicted by the empirical mean. Then v would correspond to a subset of the domain $[s]$ on which one could filter out bad points as in [JO19]. In the sequel, we will show that this kind of analysis applies *even if the solution to $\mathrm{argmax}_{\Sigma\in\mathcal{K}}\langle M_U, \Sigma\rangle$ is not integral.*

5.4 Filtering Algorithm and Analysis

In this section we prove our main theorem, stated formally below:

Theorem 5.4.1. *Let μ be an (η, s)-piecewise degree-d distribution over $[n]$. Then for any $0 < \varepsilon < 1/2$ smaller than some absolute constant, and any $0 < \delta < 1$, there is a $\mathrm{poly}(n, k, 1/\varepsilon, 1/\delta)$-time algorithm* LearnWithFilter *which, given*

$$N = \widetilde{O}\left(\log(1/\delta)(s^2 d^2/\varepsilon^2)\log^3(n)\right),$$

ε-corrupted, ω-diverse batches of size k from μ, outputs an estimate $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_1 \leq O\left(\eta + \omega + \frac{\varepsilon\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right)$ with probability at least $1 - \delta$ over the samples.

In Section 5.4.1, we first describe and prove guarantees for a basic but important subroutine, 1DFilter, of our algorithm. In Section 5.4.2, we describe our learning algorithm, LearnWithFilter, in full. In Section 5.4.3 we define the deterministic conditions that the dataset must satisfy for LearnWithFilter to succeed, deferring the proof that these deterministic conditions hold with high probability (Lemma 5.4.6) to Appendix 5.6. In Section 5.4.4 we prove a key geometric lemma (Lemma 5.4.7). Finally, in Section 5.4.5, we complete the proof of correctness of LearnWithFilter.

5.4.1 Univariate Filter

In this section, we define and analyze a simple deterministic subroutine 1DFilter which takes as input a set of weights w and a set of scores on the batches $X_1, \cdots, X_N$, and outputs a new set of weights w' such that, if the weighted average of the scores among the bad batches exceeds that of the scores among the good batches, then w' places even less weight relatively on the bad batches than does w. This subroutine is given in Algorithm 13 below.

Algorithm 13: 1DFilter(τ, w)

```
  Input: Scores $\tau : [N] \to \mathbb{R}_{\geq 0}$, weights $w : [N] \to \mathbb{R}_{\geq 0}$
  Output: New weights $w'$ with even less mass on bad points than good points (see
          Lemma 5.4.2)
1 $\tau_{\max} \leftarrow \max_{i: w_i > 0} \tau_i$
2 $w'_i \leftarrow \left(1 - \frac{\tau_i}{\tau_{\max}}\right) w_i$ for all $i \in [N]$
3 return $w'$.
```

Lemma 5.4.2. *Let $\tau : [N] \to \mathbb{R}_{\geq 0}$ be a set of scores, and let $w : [N] \to \mathbb{R}_{\geq 0}$ be a weight. Given a partition $[N] = S_G \sqcup S_B$ for which*

$$\sum_{i \in S_G} w_i \tau_i < \sum_{i \in S_B} w_i \tau_i,$$

then the output w' of 1DFilter(τ, w) *satisfies (a) $w'_i \leq w_i$ for all $i \in [N]$, (b) the support of w' is a strict subset of the support of w, and (c) $\sum_{i \in S_G} w_i - w'_i < \sum_{i \in S_B} w_i - w'_i$.*

Proof. (*a*) and (*b*) are immediate. For (*c*), note that

$$\sum_{i \in S_G} w_i - w'_i = \frac{1}{\tau_{\max}} \sum_{i \in S_G} \tau_i w_i < \frac{1}{\tau_{\max}} \sum_{i \in S_B} \tau_i w_i = \sum_{i \in S_B} w_i - w'_i,$$

from which the lemma follows. □

We note that this kind of downweighting scheme and its analysis are not new, see e.g. Lemma 4.5 from [CSV17] or Lemma 17 from [SCV18].

5.4.2 Algorithm Specification

We can now describe our algorithm LearnWithFilter. At a high level, we maintain weights $w : [N] \to \mathbb{R}_{\geq 0}$ for each of the batches. In every iteration, we compute $\Sigma \in \mathcal{K}$ maximizing $|\langle M(w), \Sigma\rangle|$. If $|\langle M(w), \Sigma\rangle| \leq O\left(\frac{\varepsilon}{k}\log 1/\varepsilon\right)$, then output $\mu(w)$. Otherwise, update the weights as follows: for every batch X_i, compute the score τ_i given by

$$\tau_i \triangleq \left\langle (X_i - \mu(w))^{\otimes 2}, \Sigma \right\rangle, \tag{5.4}$$

and set the weights to be the output of 1DFilter(τ, w). The pseudocode for LearnWithFilter is given in Algorithm 14 below.

Algorithm 14: LearnWithFilter($\{X_i\}_{i\in[N]}, \varepsilon$)

```
Input: Frequency vectors $X_1, \cdots, X_N$ coming from an $\varepsilon$-corrupted, $\omega$-diverse set of
       batches from $\mu$, where $\mu$ is $(\eta, s)$-piecewise, degree $d$
Output: $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_1 \leq O\left(\eta + \omega + \frac{\varepsilon\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right)$, provided uncorrupted
        samples $\varepsilon$-good
$w \leftarrow w([N])$
while $\|M(w)\|_{\mathcal{K}} \geq \Omega(\omega + \frac{\varepsilon}{k}\log 1/\varepsilon)$ do
    $\Sigma \leftarrow \mathrm{argmax}_{\Sigma' \in \mathcal{K}} |\langle M(w), \Sigma\rangle|$
    Compute scores $\tau : [N] \to \mathbb{R}_{\geq 0}$ according to (5.4).
    $w \leftarrow$1DFilter$(\tau, w)$
Using the algorithm of [ADLS17] (see Lemma 4.5.4), form the $s$-piecewise, degree-$d$
  distribution $\hat{w}$ minimizing $\|\mu(w) - \hat{\mu}\|_{s(d+1)}$ (up to additive error $\eta$).
return $\hat{w}$.
```

5.4.3 Deterministic Condition

Definition 5.4.3 (ε-goodness)**.** *Take a set of points $U \subset [N]$, and let $\{\mu_i\}_{i\in U}$ be a collection of distributions over $[n]$. For any $W \subseteq U$, define $\overline{\mu}_W \triangleq \frac{1}{|W|}\sum_{i\in W}\mu_i$. Denote $\overline{\mu} \triangleq \overline{\mu}_U$.*

We say U is ε-good *if it satisfies that for all $W \subset U$ for which $|W| = \varepsilon|U|$,* *(I) (Concentration of mean)*

$$\|\mu(U) - \overline{\mu}\|_\mathcal{K} \leq O\left(\frac{\varepsilon\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right) \quad \textit{and} \quad \|\mu(W) - \overline{\mu}_W\|_\mathcal{K} \leq O\left(\frac{\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right)$$

(II) (Concentration of covariance)

$$\|M(\hat{w}(U), \{\mu_i\}_{i\in U})\|_\mathcal{K} \leq O\left(\frac{\varepsilon \log 1/\varepsilon}{k}\right) \quad \textit{and} \quad \|A(\hat{w}(W), \{\mu_i\}_{i\in W}\|_\mathcal{K} \leq O\left(\frac{\log 1/\varepsilon}{k}\right)$$

(III) (Concentration of variance proxy)

$$\|B(\hat{\mu}(U)) - B(\{\mu_i\}_{i\in U})\|_\mathcal{K} \leq O(\omega^2/k + \varepsilon/k)$$

(IV) (Heterogeneity has negligible effect, see Lemma 5.4.4)

$$\sup_{\Sigma\in\mathcal{K}}\left\{\frac{1}{|U|}\sum_{i\in U}(\mu_i - \overline{\mu})^\top \cdot \Sigma \cdot (X_i - \mu_i)\right\} \leq O\left(\omega \cdot \frac{\varepsilon\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right).$$

$$\sup_{\Sigma\in\mathcal{K}}\left\{\frac{1}{|W|}\sum_{i\in W}(\mu_i - \overline{\mu})^\top \cdot \Sigma \cdot (X_i - \mu_i)\right\} \leq O\left(\omega \cdot \frac{\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right).$$

We first remark that we only need extremely mild concentration in Condition (III), but it turns out this suffices in the one place where we use it (see Lemma 5.4.9).

Additionally, note that we can completely ignore Condition (IV) when $\omega = 0$. The following makes clear why it is useful when $\omega > 0$.

Lemma 5.4.4. *For ε-good U, all $W \subset U$ of size $\varepsilon|U|$, and all $\Sigma \in \mathcal{K}$,*

$$\|A(\hat{\mu}(U), \overline{\mu}) - A(\hat{\mu}(U), \{\mu_i\})\|_\mathcal{K} \leq O\left(\omega + \frac{\varepsilon\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right)^2$$

$$\|A(\hat{\mu}(W),\overline{\mu}) - A(\hat{\mu}(W),\{\mu_i\})\|_{\mathcal{K}} \le O\left(\omega + \frac{\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right)^2.$$

Proof. For $S = U$ or $S = W$ and any $\Sigma \in \mathcal{K}$,

$$\begin{aligned}
&\langle \Sigma, A(\hat{\mu}(S),\overline{\mu}) - A(\hat{\mu}(S),\{\mu_i\})\rangle \\
&\quad= \frac{1}{|S|}\sum_{i\in S}\langle (X_i - \overline{\mu})^{\otimes 2} - (X_i-\mu_i)^{\otimes 2}, \Sigma\rangle \\
&\quad= \frac{1}{|S|}\sum_{i\in S}(\mu_i - \overline{\mu})^\top \cdot \Sigma \cdot (2X_i - \mu_i - \overline{\mu}) \\
&\quad= \frac{2}{|S|}\sum_{i\in S}(\mu_i - \overline{\mu})^\top \cdot \Sigma \cdot (X_i - \mu_i) + \frac{1}{|S|}\sum_{i\in S}\langle (\mu_i - \overline{\mu})^{\otimes 2}, \Sigma\rangle. \qquad (5.5)
\end{aligned}$$

The first (resp. second) part of the lemma follows by taking $S = U$ (resp. $S = W$) and invoking the first (resp. second) part of Condition (IV) of ε-goodness to upper bound the first term in (5.5), and Fact 5.3.4 and the fact that $\|\mu_i - \overline{\mu}\|_1 \le \omega$ for all i to upper bound the second term in (5.5). □

Corollary 5.4.5. *If U is ε-good and $\overline{\mu} \triangleq \frac{1}{|U|}\sum_{i\in U}\mu_i$, then*

$$\|A(\hat{w}(U),\overline{\mu}) - B(\{\mu_i\})\|_{\mathcal{K}} \le O\left(\omega + \frac{\varepsilon\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right)^2.$$

Proof. This follows immediately from Lemma 5.4.4 and the first part of Condition (II) of ε-goodness. □

In Appendix 5.6, we will show that for N sufficiently large, the set S_G of uncorrupted batches will satisfy the above deterministic condition.

Lemma 5.4.6 (Regularity of good samples)**.** *If U is a set of $\widetilde{\Omega}\left(\log(1/\delta)(\ell^2/\varepsilon^2)\cdot\log^3(n)\right)$ independent samples from $Mul_k(\mu_1), ..., Mul_k(\mu_{|U|})$, then U is ε-good with probability at least $1-\delta$.*

5.4.4 Key Geometric Lemma

The key property of ε-good sets is the following geometric lemma bounding the accuracy of an estimate $\mu(w)$ given by weights w in terms of $\|M(w)\|_{\mathcal{K}}$.

Lemma 5.4.7 (Spectral signatures)**.** *If S_G is ε-good and $|S_G| \geq (1-\varepsilon)N$, then for any $w \in \mathcal{W}_\varepsilon$,*

$$\|\mu(w) - \overline{\mu}\|_{\mathcal{K}} \leq O\left(\frac{\varepsilon}{\sqrt{k}}\sqrt{\log 1/\varepsilon} + \varepsilon \cdot \omega + \sqrt{\varepsilon\left(\|M(w)\|_{\mathcal{K}} + \omega^2 + \frac{\varepsilon}{k}\log 1/\varepsilon\right)}\right).$$

It turns out the proof ingredients for Lemma 5.4.7 will also be useful in our analysis of LearnWithFilter later, so we will now prove this lemma in full.

Proof. Take any $\Sigma \in \mathcal{K}$. Recalling that Σ is psd by Constraint 5 in Definition 5.3.1, we will sometimes write it as $\Sigma = \mathbb{E}_v[vv^\top]$, where the distribution over v is defined according to the eigendecomposition of Σ. We wish to bound $\mathbb{E}_v\left[\langle \mu(w) - \mu, v\rangle^2\right]$. By splitting $w_i \triangleq 1/N - \delta_i$ for $i \in S_G$, we have that

$$\begin{aligned}
\langle \mu(w) - \overline{\mu}, v\rangle &= \sum_{i=1}^N w_i \langle X_i - \overline{\mu}, v\rangle \\
&= \left\langle \frac{|S_G|}{N}(\mu(S_G) - \overline{\mu}), v\right\rangle - \sum_{i\in S_G}\delta_i\langle X_i - \overline{\mu}, v\rangle + \sum_{i\in S_B} w_i\langle X_i - \overline{\mu}, v\rangle, \\
&= \left\langle \frac{|S_G|}{N}(\mu(S_G) - \overline{\mu}), v\right\rangle - \sum_{i\in S_G}\delta_i\langle X_i - \overline{\mu}, v\rangle + \sum_{i\in S_B} w_i\langle X_i - \mu(w), v\rangle + \langle \mu(w) - \overline{\mu}, v\rangle \sum_{i\in S_B} w_i.
\end{aligned}$$

We may rewrite this as

$$\left(1 - \sum_{i\in S_B} w_i\right)\langle \overline{\mu}(w) - \mu, v\rangle = \left\langle \frac{|S_G|}{N}(\mu(S_G) - \overline{\mu}), v\right\rangle - \sum_{i\in S_G}\delta_i\langle X_i - \overline{\mu}, v\rangle + \sum_{i\in S_B} w_i\langle X_i - \overline{\mu}(w), v\rangle.$$

Note further that

$$\sum_{i\in S_G}\delta_i\langle X_i - \overline{\mu}, v\rangle = \sum_{i\in S_G}\delta_i\langle X_i - \mu_i, v\rangle + \sum_{i\in S_G}\delta_i\langle \mu_i - \overline{\mu}, v\rangle,$$

so in particular,

$$\frac{1}{4}\left(1-\sum_{i\in S_B} w_i\right)^2 \cdot \mathbb{E}_v\left[\langle \mu(w)-\overline{\mu}, v\rangle^2\right] \le \textcircled{1}+\textcircled{2}+\textcircled{3}+\textcircled{4} \tag{5.6}$$

where

$$\textcircled{1} \triangleq \frac{|S_G|^2}{N^2}\mathbb{E}_v\left[\langle \mu(S_G)-\overline{\mu}, v\rangle^2\right] \qquad \textcircled{2} \triangleq \mathbb{E}_v\left[\left(\sum_{i\in S_G}\delta_i\langle X_i-\mu_i, v\rangle\right)^2\right]$$

$$\textcircled{3} \triangleq \mathbb{E}_v\left[\left(\sum_{i\in S_G}\delta_i\langle \mu_i-\mu, v\rangle\right)^2\right] \qquad \textcircled{4} \triangleq \mathbb{E}_v\left[\left(\sum_{i\in S_B}w_i\langle X_i-\mu(w), v\rangle\right)^2\right]$$

For $\textcircled{1}$, note that

$$\textcircled{1} \le \frac{|S_G|^2}{N^2}\|\mu(S_G)-\overline{\mu}\|^2_{\mathcal{K}} \le O\left(\frac{\varepsilon^2\log 1/\varepsilon}{k}\right)$$

by the first part of Condition (I) of ε-goodness of S_G and the fact that $|S_G|/N \ge 1-\varepsilon$.

For $\textcircled{2}$, by Cauchy-Schwarz we have that

$$\begin{aligned}
\textcircled{2} &\le \left(\sum_{i\in S_G}\delta_i\right)\cdot\mathbb{E}_v\left[\sum_{i\in S_G}\delta_i\langle X_i-\mu_i, v\rangle^2\right] \\
&\le \varepsilon\cdot\left\langle\sum_{i\in S_G}\delta_i(X_i-\mu_i)^{\otimes 2}, \mathbb{E}_v[vv^\top]\right\rangle \\
&= \varepsilon\left\langle A(\delta,\{\mu_i\}),\Sigma\right\rangle \\
&\le O\left(\frac{\varepsilon^2}{k}\log 1/\varepsilon\right),
\end{aligned} \tag{5.7}$$

where the last step follows by Lemma 5.4.8 below.

For $\textcircled{3}$, again by Cauchy-Schwarz,

$$\begin{aligned}
\textcircled{3} &\le \left(\sum_{i\in S_G}\delta_i\right)\cdot\mathbb{E}_v\left[\sum_{i\in S_G}\delta_i\langle \mu_i-\mu, v\rangle^2\right] \\
&\le \varepsilon\cdot\sum_{i\in S_G}\delta_i\|\mu_i-\mu\|^2_{\mathcal{K}} \\
&\le \varepsilon^2\cdot\max_{i\in S_G}\|\mu_i-\mu\|^2_1 \\
&\le \varepsilon^2\cdot\omega^2,
\end{aligned}$$

where the penultimate step follows by Fact 5.3.4.

Finally, we will relate ④ to $\|M(w)\|_{\mathcal{K}}$. Let w' be the set of weights given by $w'_i = w_i$ for $i \in S_G$ and $w'_i = 0$ for $i \notin S_G$. By another application of Cauchy-Schwarz,

$$\begin{aligned}
④ &\le \left(\sum_{i\in S_B} w_i\right) \cdot \mathbb{E}_v\left[\sum_{i\in S_B} w_i \langle X_i - \mu(w), v\rangle^2\right] \\
&\le \varepsilon \left(\mathbb{E}_v\left[\sum_{i=1}^N w_i \langle X_i - \mu(w), v\rangle^2\right] - \mathbb{E}_v\left[\sum_{i\in S_G} w_i\langle X_i - \mu(w), v\rangle^2\right]\right) \\
&= \varepsilon \left\langle A(w, \mu(w)) - A(w', \mu(w)), \Sigma\right\rangle && (5.8) \\
&\le \varepsilon \left\langle A(w, \mu(w)) - A(w', \mu(w')), \Sigma\right\rangle && (5.9) \\
&\le \varepsilon \left\langle A(w, \mu(w)) - \frac{1}{\sum w'_i} B(\mu(w')), \Sigma\right\rangle + O\left(\varepsilon \cdot \omega^2 + \frac{\varepsilon^2}{k}\log 1/\varepsilon\right) && (5.10) \\
&= \varepsilon\langle M(w), \Sigma\rangle + \varepsilon\left\langle B(\mu(w)) - \frac{1}{\sum w'_i} B(\mu(w')), \Sigma\right\rangle + O\left(\varepsilon \cdot \omega^2 + \frac{\varepsilon^2}{k}\log 1/\varepsilon\right) \\
&\le \varepsilon\|M(w)\|_{\mathcal{K}} + \varepsilon\|B(\mu(w)) - \frac{1}{\sum w'_i} B(\mu(w'))\|_{\mathcal{K}} + O\left(\varepsilon \cdot \omega^2 + \frac{\varepsilon^2}{k}\log 1/\varepsilon\right) && (5.11)
\end{aligned}$$

where (5.8) follows by the definition of $A(w, \nu)$, (5.9) follows by Fact 5.2.1, (5.10) follows by Lemma 5.4.9 below. Lastly, by triangle inequality, we may upper bound $\|B(\mu(w)) - \frac{1}{\sum w'_i} B(\mu(w'))\|_{\mathcal{K}}$ by

$$\|B(\mu(w)) - B(\mu(w'))\|_{\mathcal{K}} + O(\varepsilon) \cdot \|B(\mu(w'))\|_{\mathcal{K}} \le \frac{3}{k}\|\mu(w) - \mu(w')\|_1 + O(\varepsilon/k) \le O(\varepsilon/k), \quad (5.12)$$

where the first inequality follows by Corollary 5.3.5, and the bound on $\|\mu(w) - \mu(w')\|_1$ in the last step follows from Fact 5.2.2. The lemma then follows from (5.6), (5.7), (5.11), and (5.12). □

Next, we show in Lemma 5.4.8 that small subsets of the good samples cannot contribute too much to the total energy. Lemma 5.4.9, which bounds the norm of $M(w)$ for any set of weights w which is close to the uniform set of weights over S_G, will follow as a consequence.

Lemma 5.4.8. *For any $0 < \varepsilon < 1/2$, if U is ε-good, and $\delta : U \to [0, 1/|U|]$ is a set of weights satisfying $\sum_{i\in U} \delta_i \le \varepsilon$, then we have the following bounds:*

1. $\|A(\delta, \{\mu_i\})\|_{\mathcal{K}} \le O(\frac{\varepsilon}{k}\log 1/\varepsilon)$

2. $\|\sum_{i\in U}\delta_i(X_i-\mu_i)\|_{\mathcal{K}} \le O(\frac{\varepsilon}{\sqrt{k}}\sqrt{\log 1/\varepsilon})$

3. $\|A(\delta,\overline{\mu})\|_{\mathcal{K}} \le O\left(\varepsilon\cdot\omega^2+\frac{\varepsilon\log 1/\varepsilon}{k}\right)$

4. $\|\sum_{i\in U}\delta_i(X_i-\overline{\mu})\|_{\mathcal{K}} \le O(\frac{\varepsilon}{\sqrt{k}}\sqrt{\log 1/\varepsilon}+\varepsilon\cdot\omega)$.

Proof. For the first part, we may assume without loss of generality that $\sum_{i\in U}\delta_i=\varepsilon$. But then we may write δ as $\varepsilon\mathbb{E}_W[\hat{w}(W)]$ for some distribution over subsets $W\subset U$ of size $\varepsilon|U|$. By Jensen's inequality and the second part of Condition (II) of ε-goodness of U, we conclude that

$$A(\delta,\{\mu_i\}) \le \varepsilon\cdot\mathbb{E}_W\left[\|A(\hat{w}(W),\{\mu_i\})\|_{\mathcal{K}}\right] \le O\left(\frac{\varepsilon}{k}\log 1/\varepsilon\right),$$

giving the first part of the lemma.

For the second part, for any $\Sigma\in\mathcal{K}$ of the form $\Sigma=\mathbb{E}[vv^\top]$,

$$\begin{aligned}\left\langle\Sigma,\left(\sum_{i\in U}\delta_i(X_i-\mu_i)\right)^{\otimes 2}\right\rangle &= \mathbb{E}\left[\left(\sum_{i\in U}\delta_i\langle X_i-\mu_i,v\rangle\right)^2\right]\\ &\le \mathbb{E}\left[\left(\sum_{i\in U}\delta_i\right)\cdot\left(\sum_{i\in U}\delta_i\langle X_i-\mu_i,v\rangle^2\right)\right]\\ &\le \varepsilon\|A(\delta,\{\mu_i\})\| \le O\left(\frac{\varepsilon^2}{k}\log 1/\varepsilon\right),\end{aligned}$$

where the second step follows by Cauchy-Schwarz, the fourth step follows by the first part of the lemma. As this holds for all $\Sigma\in\mathcal{K}$, we get the second part of the lemma.

This also implies the fourth part of the lemma because

$$\begin{aligned}\|\sum_{i\in U}\delta_i(X_i-\overline{\mu})\|_{\mathcal{K}} &\le \|\sum_{i\in U}\delta_i(X_i-\mu_i)\|_{\mathcal{K}}+\|\sum_{i\in U}\delta_i(\mu_i-\overline{\mu})\|_{\mathcal{K}}\\ &\le O\left(\frac{\varepsilon}{\sqrt{k}}\sqrt{\log 1/\varepsilon}\right)+\sum_{i\in U}\delta_i\|\mu_i-\overline{\mu}\|_1\\ &\le O\left(\frac{\varepsilon}{\sqrt{k}}\sqrt{\log 1/\varepsilon}+\varepsilon\cdot\omega\right),\end{aligned}$$

where the second step follows by the above together with Fact 5.3.4 and triangle inequality.

Finally, for the third part of the lemma, upon regarding the weights δ as $\varepsilon\mathbb{E}_W[\hat{w}(W)]$ as before and applying Jensen's to the second part of Lemma 5.4.4, we get that

$$\|A(\delta,\overline{\mu}) - A(\delta,\{\mu_i\})\|_{\mathcal{K}} \le \varepsilon \cdot O\left(\omega + \frac{\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right)^2 \le O\left(\varepsilon \cdot \omega^2 + \frac{\varepsilon \log 1/\varepsilon}{k}\right).$$

The third part of the lemma then follows by the first part, together with triangle inequality.

□

Lemma 5.4.9. *If S_G is ε-good, and $w : S_G \to [0,1]$ satisfies $\|w - \hat{w}(S_G)\|_1 \le \varepsilon$ and $\sum_{i\in S_G} w_i = 1$, then $\|M(w)\|_{\mathcal{K}} \le O(\omega^2 + \frac{\varepsilon}{k}\log 1/\varepsilon)$.*

Proof. Define $\delta_i = 1/|S_G| - w_i$ for all $i \in S_G$ and take any $\Sigma \in \mathcal{K}$.

By Fact 5.2.1 and the assumption that $\|w\|_1 = 1$,

$$\langle A(w,\mu(w)),\Sigma\rangle = \langle A(w,\overline{\mu}),\Sigma\rangle - \|\mu(w) - \overline{\mu}\|_{\mathcal{K}}^2. \tag{5.13}$$

For the second term on the right-hand side of (5.13), note that we can write

$$\begin{aligned}
\mu(w) - \overline{\mu} &= \sum_{i\in S_G} w_i(X_i - \overline{\mu}) \\
&= \sum_{i\in S_G} (1/|S_G| - \delta_i)(X_i - \overline{\mu}) \\
&= (\mu(S_G) - \overline{\mu}) - \sum_{i\in S_G} \delta_i(X_i - \overline{\mu}) \\
&= (\mu(S_G) - \overline{\mu}) - \sum_{i\in S_G} \delta_i(X_i - \mu_i) - \sum_{i\in S_G} \delta_i(\mu_i - \overline{\mu}),
\end{aligned}$$

where the first step follows by the fact that $\sum_{i\in S_G} w_i = 1$. So by triangle inequality,

$$\|\mu(w) - \overline{\mu}\|_{\mathcal{K}} \le \|\mu(S_G) - \overline{\mu}\|_{\mathcal{K}} + \|\sum_{i\in S_G} \delta_i(X_i - \overline{\mu})\|_{\mathcal{K}} \le O\left(\frac{\varepsilon}{\sqrt{k}}\sqrt{\log 1/\varepsilon} + \varepsilon \cdot \omega\right) \tag{5.14}$$

where the second step follows by the first part of Condition (I) in the definition of ε-goodness for S_G, together with the second part of Lemma 5.4.8.

Next, we bound the first term on the right-hand side of (5.13). We have

$$\begin{aligned}
|\langle A(w,\overline{\mu}),\Sigma\rangle| &\le |\langle A(\hat{w}(S_G),\overline{\mu}),\Sigma\rangle| + |\langle A(\delta,\overline{\mu}),\Sigma\rangle| \\
&\le |\langle A(\hat{w}(S_G),\overline{\mu}),\Sigma\rangle| + O\left(\frac{\varepsilon}{k}\log 1/\varepsilon + \varepsilon\cdot\omega^2\right) \\
&\le |\langle B(\{\mu_i\}),\Sigma\rangle| + O\left(\omega^2 + \frac{\varepsilon\log 1/\varepsilon}{k}\right) \\
&\le |\langle B(\hat{\mu}(S_G)),\Sigma\rangle| + O\left(\omega^2 + \frac{\varepsilon\log 1/\varepsilon}{k}\right), \qquad (5.15)
\end{aligned}$$

where the second step follows by the third part of Lemma 5.4.8, the third step follows by Corollary 5.4.5, and the fourth step follows by Condition (III) of ε-goodness.

Additionally, by Corollary 5.3.5, we can bound

$$|\langle B(\mu(w)),\Sigma\rangle - \langle B(\hat{\mu}(S_G)),\Sigma\rangle| \le \frac{3}{k}\|\mu(w)-\hat{\mu}(S_G)\|_1 \le \frac{3}{k}\|w-\hat{w}(S_G)\|_1 \le O(\varepsilon/k). \quad (5.16)$$

By (5.15) and (5.16) we conclude that $\langle A(w,\overline{\mu}),\Sigma\rangle \le \langle B(\mu(w)),\Sigma\rangle + O(\frac{\varepsilon}{k}\log 1/\varepsilon)$, so this together with (5.13) and (5.14) yields the desired bound. □

5.4.5 Analyzing the Filter With Spectral Signatures

We now use Lemma 5.4.7 to show that under the deterministic condition that the uncorrupted points are ε-good, LearnWithFilter satisfies the guarantees of Theorem 5.4.1.

The main step is to show that as long as we remain in the main loop of LearnWithFilter, and we have so far thrown out more bad weight than good weight, we are guaranteed to throw out more bad weight than good weight in the next iteration of the main loop:

Lemma 5.4.10. *Let w and w' be the weights at the start and end of a single iteration of the main loop of* LearnWithFilter*. There is an absolute constant $C > 0$ such that if $\|M(w)\|_{\mathcal{K}} > C\cdot\frac{\varepsilon}{k}\log 1/\varepsilon$ and $\sum_{i\in S_G}\frac{1}{N} - w_i < \sum_{i\in S_B}\frac{1}{N} - w_i$, then $\sum_{i\in S_G} w_i - w'_i < \sum_{i\in S_B} w_i - w'_i$.*

Proof. Suppose the scores $\tau_1,\cdots,\tau_N$ in this iteration are sorted in decreasing order, and let T denote the smallest index for which $\sum_{i\in[T]} w_i \ge 2\varepsilon$. As Filter does not modify w_i for $i > T$, we just need to show that $\sum_{i\in S_G\cap[T]} w_i - w'_i < \sum_{i\in S_B\cap[T]} w_i - w'_i$, and by Lemma 5.4.2

it is enough to show that

$$\sum_{i\in S_G\cap[T]} w_i\tau_i < \sum_{i\in S_B\cap[T]} w_i\tau_i. \tag{5.17}$$

First note that because each weight is at most ε, we may assume that $\sum_{i\in[T]} w_i \le 3\varepsilon$. We begin by upper bounding the left-hand side of (5.17).

Lemma 5.4.11. $\sum_{i\in S_G\cap[T]} w_i\tau_i \le O\left(\frac{\varepsilon}{k}\log 1/\varepsilon + \varepsilon\cdot\omega^2 + \varepsilon^2\|M(w)\|_{\mathcal{K}}\right)$.

Proof. Let w'' be the weights given by w''_i for $i \in S_G\cap[T]$ and $w''_i = 0$ otherwise. Then $\sum_{S_G\cap[T]} w_i\tau_i$ is equal to

$$\begin{aligned}
\sum_{i\in[N]} w''_i\tau_i &= \sum_{i\in[N]} w''_i\left\langle (X_i-\mu(w))^{\otimes 2},\Sigma\right\rangle \\
&= \sum_{i\in[N]} w''_i\left\langle (X_i-\mu(w''))^{\otimes 2},\Sigma\right\rangle + \|w''\|_1\cdot\left\langle(\mu(w'')-\mu(w))^{\otimes 2},\Sigma\right\rangle && (5.18)\\
&\le \sum_{i\in[N]} w''_i\left\langle (X_i-\mu(w''))^{\otimes 2},\Sigma\right\rangle + O(\varepsilon)\cdot\|\mu(w'')-\mu(w)\|^2_{\mathcal{K}} && (5.19)\\
&\le \sum_{i\in[N]} w''_i\left\langle (X_i-\overline{\mu})^{\otimes 2},\Sigma\right\rangle + O(\varepsilon)\cdot\|\mu(w'')-\mu(w)\|^2_{\mathcal{K}} && (5.20)\\
&\le O\left(\varepsilon\cdot\omega^2 + \frac{\varepsilon}{k}\log 1/\varepsilon\right) + O(\varepsilon)\cdot\|\mu(w'')-\mu(w)\|^2_{\mathcal{K}}
\end{aligned}$$

where (5.18) and (5.20) both follow from Fact 5.2.1, (5.19) follows from the earlier assumption that $\sum_{i\in[T]} w_i \le 3\varepsilon$ and the definition of $\|\cdot\|_{\mathcal{K}}$, and the last step follows by the third part of Lemma 5.4.8.

Now note that

$$\begin{aligned}
\|\mu(w'')-\mu(w)\|_{\mathcal{K}} &\le \|\mu(w'')-\overline{\mu}\|_{\mathcal{K}} + \|\mu(w)-\overline{\mu}\|_{\mathcal{K}} \\
&\le O\left(\frac{\sqrt{\log 1/\varepsilon}}{\sqrt{k}} + \omega\right) + \|\mu(w)-\overline{\mu}\|_{\mathcal{K}} \\
&\le O\left(\frac{\sqrt{\log 1/\varepsilon}}{\sqrt{k}} + \omega + \sqrt{\varepsilon\left(\|M(w)\|_{\mathcal{K}} + \omega^2 + \frac{\varepsilon}{k}\log 1/\varepsilon\right)}\right),
\end{aligned}$$

where the second step follows by the fourth part of Lemma 5.4.8 and the third step holds by Lemma 5.4.7. The desired bound follows. □

One consequence of this is that outside of the tails, the scores among good samples are small.

Corollary 5.4.12. *For all $i > T$, $\tau_i \leq O(\frac{1}{k}\log 1/\varepsilon + \varepsilon\|M(w)\|_{\mathcal{K}} + \omega^2)$.*

Proof. Note that

$$\sum_{i\in S_G\cap[T]} w_i = \sum_{i\in[T]} w_i - \sum_{i\in S_B\cap[T]} w_i \geq 2\varepsilon - \sum_{i\in S_B} w_i \geq \varepsilon,$$

so the claim follows from Lemma 5.4.11 and averaging. □

Next, we show that the deviation of the total scores of the good points from their expectation is negligible.

Lemma 5.4.13. $\sum_{i\in S_G} w_i\tau_i - \langle B(\mu(w)),\Sigma\rangle \leq O\left(\frac{\varepsilon}{k}\log 1/\varepsilon + \varepsilon\cdot\omega^2 + \varepsilon\cdot\|M(w)\|_{\mathcal{K}}\right)$.

Proof. Let w' be the weights given by $w'_i = w_i$ for $i \in S_G$ and $w'_i = 0$ otherwise. Then by Fact 5.2.1,

$$\begin{aligned}\sum_{i\in S_G} w_i\tau_i &= \sum_{i\in S_G} w_i\langle (X_i - \mu(w'))^{\otimes 2},\Sigma\rangle + \|w\|_1\cdot\langle(\mu(w)-\mu(w'))^{\otimes 2},\Sigma\rangle \\ &\leq \frac{1}{\sum_{i\in S_G} w_i}\left(\langle B(\mu(w')),\Sigma\rangle + O\left(\frac{\varepsilon}{k}\log 1/\varepsilon\right)\right) + \|\mu(w)-\mu(w')\|_{\mathcal{K}}^2\end{aligned}$$

where in the second step we used Fact 5.2.1, and in the third step we used Lemma 5.4.9 and the definition of $\|\cdot\|_{\mathcal{K}}$. To bound the $\|\mu(w)-\mu(w')\|_{\mathcal{K}}^2$ term, note that

$$\begin{aligned}\|\mu(w)-\mu(w')\|_{\mathcal{K}} &\leq \|\mu(w)-\overline{\mu}\|_{\mathcal{K}} + \|\mu(w')-\overline{\mu}\|_{\mathcal{K}} \\ &\leq \|\mu(w)-\overline{\mu}\|_{\mathcal{K}} + O\left(\frac{\varepsilon\sqrt{\log 1/\varepsilon}}{\sqrt{k}} + \varepsilon\cdot\omega\right) \\ &\leq O\left(\frac{\varepsilon\sqrt{\log 1/\varepsilon}}{\sqrt{k}} + \varepsilon\cdot\omega + \sqrt{\varepsilon\left(\|M(w)\|_{\mathcal{K}} + \omega^2 + \frac{\varepsilon}{k}\log 1/\varepsilon\right)}\right),\end{aligned}$$

where the second step follows by the fourth part of Lemma 5.4.8, and the third step follows by Lemma 5.4.7. Finally, by Corollary 5.3.5 we have that

$$\langle B(\mu(w')), \Sigma\rangle \leq \langle B(\mu(w)), \Sigma\rangle + \frac{3}{k}\|\mu(w') - \mu(w)\|_1 \leq \langle B(\mu(w)), \Sigma\rangle + O(\varepsilon/k),$$

where the last step follows by Fact 5.2.2. This completes the proof of the claim. □

We are now ready to complete the proof of Lemma 5.4.10. In light of Lemma 5.4.11, we wish to lower bound the right-hand side of (5.17).

Claim 5.4.14. *If $C > 0$ in the lower bound $\|M(w)\|_{\mathcal{K}} > C(\frac{\varepsilon}{k}\log 1/\varepsilon + \omega^2)$ is sufficiently large, then $\langle M(w), \Sigma^*\rangle$ must be positive.*

Proof. Let w' denote the weights given by $w'_i = w_i$ for $i \in S_G$ and $w'_i = 0$ otherwise. We have

$$\begin{aligned} M(w) &= \sum_{i\in[N]} w_i(X_i - \mu(w))^{\otimes 2} - B(\mu(w)) \\ &\succeq \sum_{i\in S_G} w'_i(X_i - \mu(w))^{\otimes 2} - B(\mu(w)) \\ &\succeq \sum_{i\in S_G} w'_i(X_i - \mu(w'))^{\otimes 2} - B(\mu(w)) \\ &= M(w') + B(\mu(w')) - B(\mu(w)) \end{aligned} \tag{5.21}$$

where the third step follows by Fact 5.2.1. Furthermore,

$$\|B(\mu(w')) - B(\mu(w))\|_{\mathcal{K}} \leq \frac{3}{k}\cdot\|\mu(w') - \mu(w)\|_1 \leq O(\varepsilon/k) \tag{5.22}$$

by Corollary 5.3.5 and Fact 5.2.2. Lastly, we must bound $\|M(w')\|_{\mathcal{K}}$. Letting $\hat{w}'$ denote the normalized version of w', we have that

$$\begin{aligned} \|M(w')\|_{\mathcal{K}} &\leq \|M(\hat{w}')\|_{\mathcal{K}} + \|M(w') - M(\hat{w}')\|_{\mathcal{K}} \\ &\leq \|M(\hat{w}')\|_{\mathcal{K}} + \|A(\hat{w}' - w', \overline{\mu})\|_{\mathcal{K}} \\ &\leq O\left(\frac{\varepsilon}{k}\log 1/\varepsilon + \omega^2\right), \end{aligned} \tag{5.23}$$

where the penultimate step follows by Fact 5.2.1 and the definition of the matrix $M(\cdot)$, and the last step follows by Lemma 5.4.9 and the third part of Lemma 5.4.8.

We conclude by (5.21), (5.22), and (5.23) that

$$\min_{\Sigma\in\mathcal{K}}\langle M(w),\Sigma\rangle \geq -O\left(\frac{\varepsilon}{k}\log 1/\varepsilon + \omega^2\right), \tag{5.24}$$

so we simply need to take C larger than the constant implicit in the right-hand side of (5.24) to ensure that $\langle M(w),\Sigma^*\rangle > 0$. □

By Claim 5.4.14 and the definition of the scores,

$$\sum_{i\in[N]} w_i\tau_i - \langle B(\mu(w)),\Sigma^*\rangle = \langle M(w),\Sigma^*\rangle \geq \|M(w)\|_{\mathcal{K}}.$$

This, together with Lemma 5.4.13, yields $\sum_{i\in S_B} w_i\tau_i \geq C'\|M(w)\|_{\mathcal{K}}$ for some $C' < C$ which we can take to be arbitrarily large. We want to show that this same sum, over only $S_B\cap[T]$, enjoys essentially the same bound. Indeed,

$$\begin{aligned}\sum_{i\in S_B\cap[T]} w_i\tau_i &\geq C'\|M(w)\|_{\mathcal{K}} - \sum_{i\in S_B\backslash[T]} w_i\tau_i \\ &\geq C'\|M(w)\|_{\mathcal{K}} - \left(\sum_{i\in S_B} w_i\right)\cdot O\left(\frac{1}{k}\log 1/\varepsilon + \omega^2 + \varepsilon\|M(w)\|_{\mathcal{K}}\right) \\ &\geq \overline{C}\cdot\|M(w)\|_{\mathcal{K}},\end{aligned}$$

for some arbitrarily large absolute constant $\overline{C}$, where the second step follows by Corollary 5.4.12, and the last by the assumption that $\|M(w)\|_{\mathcal{K}} > C\cdot(\frac{\varepsilon}{k}\log 1/\varepsilon + \omega^2)$. On the other hand, by this same assumption and by Lemma 5.4.11,

$$\sum_{i\in S_G\cap[T]} w_i\tau_i \leq O\left(\frac{\varepsilon}{k}\log 1/\varepsilon + \varepsilon\cdot\omega^2 + \varepsilon^2\|M(w)\|_{\mathcal{K}}\right) \leq \underline{C}\cdot\|M(w)\|_{\mathcal{K}},$$

where $\underline{C}$ can be taken to be smaller than $\overline{C}$. This proves (5.17) and thus Lemma 5.4.10. □

We can now combine Lemma 5.4.7 and Lemma 5.4.10 to get a proof of Theorem 5.4.1.

Proof of Theorem 5.4.1. Let $\hat{\mu}$ be the output of LearnWithFilter. By Lemma 4.5.4, it suffices to show that $\hat{\mu}$ satisfies $\|\hat{\mu}-\mu\|_{\mathcal{A}_{s(d+1)}} \le O(\omega + \frac{\varepsilon}{\sqrt{k}}\sqrt{\log 1/\varepsilon})$, or equivalently that for all $v \in \mathcal{V}^n_\ell$, where $\ell \triangleq 2s(d+1)$, we have that $\langle (\hat{\mu}-\mu)^{\otimes 2}, vv^\top\rangle^{1/2} \le O(\omega + \frac{\varepsilon}{\sqrt{k}}\sqrt{\log 1/\varepsilon})$. By Corollary 5.3.3, it is enough to show that $\|\hat{\mu}-\mu\|_{\mathcal{K}} \le O(\omega + \frac{\varepsilon}{\sqrt{k}}\sqrt{\log 1/\varepsilon})$. By Lemma 5.4.7 together with the termination condition of the main loop of LearnWithFilter, we just need to show that the algorithm terminates (in polynomial time) and that $w \in \mathcal{W}_{O(\varepsilon)}$.

But by induction and Lemma 5.4.10, every iteration of the loop removes more mass from the bad points than from the good points. Furthermore, by Lemma 5.4.2, the support of w goes down by at least one every time 1DFilter is run, so the loops terminates after at most N iterations, each of which can be implemented in polynomial time. At the end, at most an ε fraction of the total mass on S_G has been removed, so the final weights w satisfy $w \in \mathcal{W}_{2\varepsilon}$ as desired. □

5.5 Numerical Experiments

In this section we report on empirical evaluations of our algorithm on synthetic data. We compared our algorithm LearnWithFilter, the naive estimator which simply takes the empirical mean of all samples, the "oracle" algorithm which computes the empirical mean of the *uncorrupted samples*, and the threshold of $\varepsilon/\sqrt{k}$ which our theorems show that LearnWithFilter achieves, up to constant factors (in Figures 5-1 and 5-2, these are labeled "filter", "naive", "oracle", and $\varepsilon/\sqrt{k}$ respectively). Note that by definition, the oracle dominates the algorithms considered in the previous chapter and [JO19] for the *unstructured* case, as those algorithms search for a subset of the data and output the empirical mean of that subset. But as Theorem 5.4.1 predicts, LearnWithFilter should actually *outperform* the oracle in settings where the underlying distribution μ is *structured* and there are too few samples for the empirical mean of the uncorrupted points to concentrate sufficiently. In these experiments, we confirm this empirically.

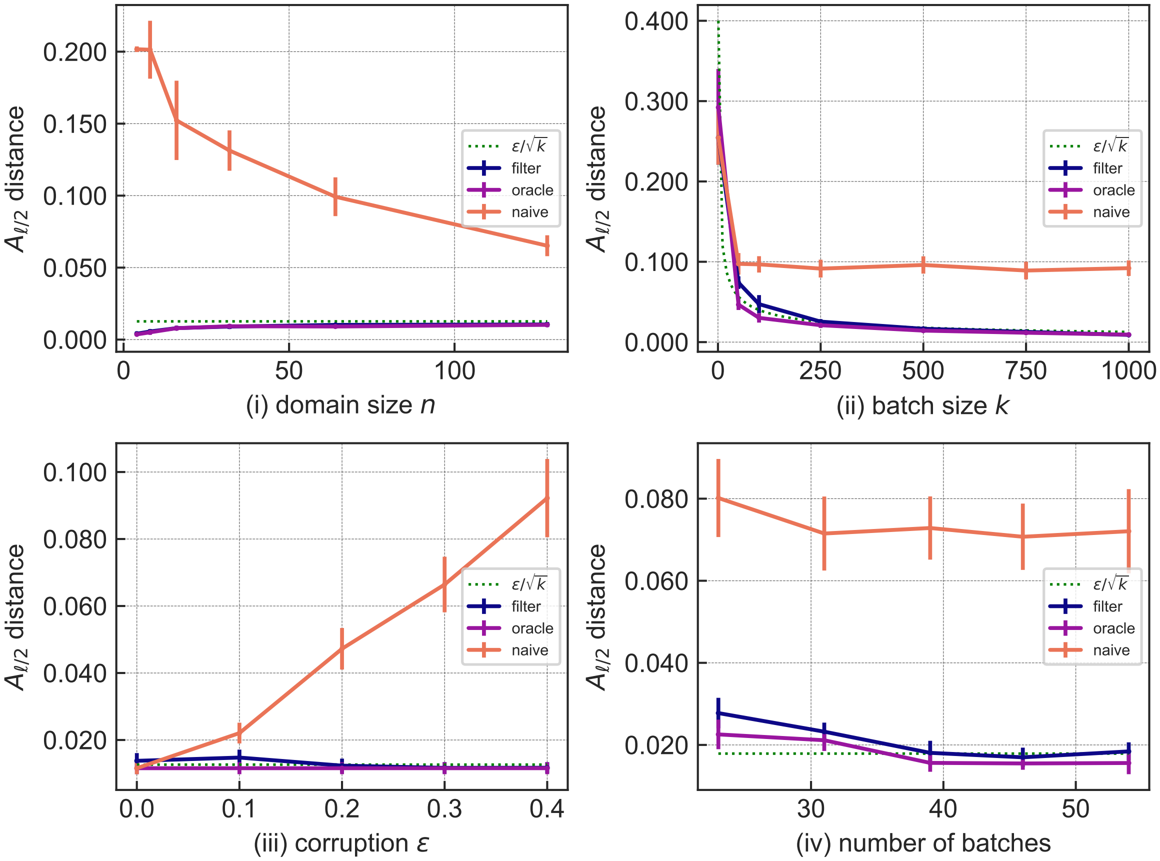


Figure 5-1: Experimental results for learning arbitrary distributions

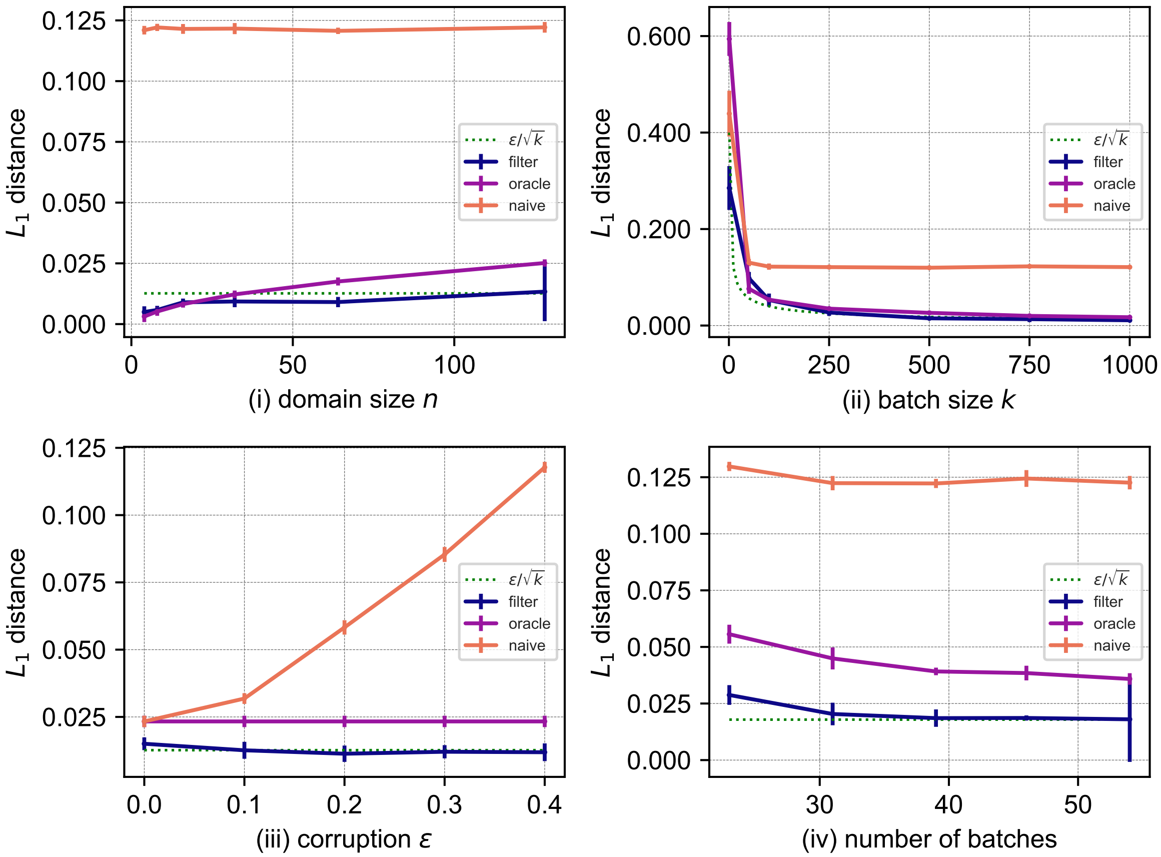


Figure 5-2: Experimental results for learning structured distributions

5.5.1 Experimental Design

Our experiments fall under two types: (A) those on learning an *arbitrary distribution* in $\mathcal{A}_{\ell/2}$ *norm* and B) those on learning a *structured distribution* in *total variation distance*. The purpose of experiments of type (A) will be to convey that LearnWithFilter can be used to learn from untrusted batches in $\mathcal{A}_{\ell/2}$ norm even for distributions which are not necessarily structured. The purpose of experiments of type (B) will be to demonstrate that LearnWithFilter can outperform the oracle for structured distributions.

Throughout, $\omega = 0$ and $\ell = 10$. While our algorithm can also be implemented for larger ℓ (as the size of the SDP we solve does not depend on ℓ), we choose $\ell = 5$ because it is small enough that the sample complexity savings of our algorithm are very pronounced, yet large enough that for the domain sizes n we work with, enumerating over $\mathcal{V}^n_\ell$ would be prohibitively expensive, justifying the need to use an SDP.

For experiments of type (A), we chose the true underlying distribution μ by sampling uniformly from $[0,1]^n$ and normalizing, and for experiments of type B), we chose μ by sampling a uniformly random piecewise constant function with $\ell = 5$ pieces.

Given μ and a prescribed parameter δ, the distribution from which the corrupted batches were drawn was taken to be $\mathrm{Mul}_k(\nu)$, where ν was constructed to satisfy $d_{\mathrm{TV}}(\mu,\nu) = \delta$ by adding $\frac{2\delta}{n}$ to the smallest entries of μ and subtracting $\frac{2\delta}{n}$ from the largest. Sometimes this does not give a probability distribution, in which case we resample μ. When k, ε, N are clear from context and we say that N ε-corrupted batches are drawn from the distribution specified by (μ,ν), we mean that $\lfloor(1-\varepsilon)N\rfloor$ samples are drawn from $\mathrm{Mul}_k(\mu)$ and $N - \lfloor(1-\varepsilon)N\rfloor$ from $\mathrm{Mul}_k(\nu)$.

As noted in [JO19], choosing δ too high makes it too easy to detect the corruptions in the data, while choosing δ too low means the naive estimator will already perform quite well. In light of this and the fact that the above process for generating ν only ensures that $d_{\mathrm{TV}}(\mu,\nu) = \delta$, whereas $\|\mu - \nu\|_{\mathcal{A}_\ell}$ might be much smaller, we chose δ for our experiments as follows. For experiments of type (A), we took $\delta = 0.5$ to ensure that the typical $\mathcal{A}_{\ell/2}$ distance between the empirical mean and the truth was still sufficiently large that the the naive estimator was not competitive. For experiments of type B) where we measure error in terms of total variation distance, we could afford to choose δ slightly smaller, namely $\delta = 0.3$.

We first describe the experiments of type (A). We examined the effect of varying one of the following four parameters at a time: domain size n, batch size k, corruption fraction ε, and total number of batches N. Each of the following four experiments was repeated for a total of ten trials.

(a) *Varying domain size* n: We fixed $\varepsilon = 0.4$, $k = 1000$, and $N = \lfloor \frac{\ell/\varepsilon^2}{1-\varepsilon} \rfloor$ to ensure $\lfloor \ell/\varepsilon^2 \rfloor$ samples from $\mathrm{Mul}_k(\mu)$. We chose such large k to ensure the gap between empirical mean and our algorithm was very noticable. In each trial and for each $n \in [4, 8, 16, 32, 64, 128]$, we randomly generated (μ, ν) via the above procedure, drew N ε-corrupted samples from distribution specified by (μ, ν). Note that while N is independent of n, the performance of our algorithm is comparable to that of the oracle.[3]

(b) *Varying batch size* k: We fixed $\varepsilon = 0.4$, $n = 64$, and $N = \frac{\ell/\varepsilon^2}{1-\varepsilon} \rfloor$. In each trial, we randomly generated (μ, ν) via the above procedure, and then for each value of $k \in [1, 50, 100, 250, 500, 750, 1000]$ we drew N samples from the distribution specified by (μ, ν). Note that while our algorithm's error and the oracle's error decay with k, the empirical mean's error remains fixed.

(c) *Varying corruption fraction* ε: We fixed $\varepsilon^* = 0.4$, $n = 64$, $k = 1000$, and $N = \lfloor \ell/\varepsilon^{*2} \rfloor$. In each trial, we randomly generated (μ, ν) via the above procedure and drew N samples from $\mathrm{Mul}(k, \mu)$. Then for each $\varepsilon \in [0.0, 0.1, 0.2, 0.3, 0.4]$, we augmented this with an additional $\lfloor \frac{\varepsilon N}{1-\varepsilon}$ samples from $\mathrm{Mul}(k, \nu)$. Note that while our algorithm's error remains close to $\varepsilon^*/\sqrt{k}$, the empirical mean's error increases linearly in ε.

(d) *Varying number of batches* N: We fixed $\varepsilon = 0.4$, $n = 128$, and $k = 500$. In each trial, we randomly generated (μ, ν) via the above procedure, and then for each $\rho \in [0.5, 0.75, 1, 1.25, 1.5]$, we drew $N = \lfloor \rho \cdot \ell/\varepsilon^2 \rfloor$ samples from the distribution specified by (μ, ν). Note that even with such a small number of samples, our algorithm can compete with the oracle. Also note that our error bottoms out at $\varepsilon/\sqrt{k}$ while the oracle's error goes beneath this threshold.

[3] The naive estimator's error is decreasing in n for an unrelated reason: as n increases, the above procedure for sampling (μ, ν) appears to skew towards μ for which the resulting perturbation ν is close in $\mathcal{A}_{\ell/2}$.

For type (B), we ran the exact same set of four experiments but over structured μ, with the key difference that after generating an estimate with LearnWithFilter, we post-processed it by rounding to a piecewise constant function via a simple dynamic program. We then compare the error of this piecewise constant estimator in *total variation distance* to that of the empirical mean of the whole dataset, and the empirical mean of the uncorrupted points.

As is evident from Figure 5-2, our algorithm outperforms even the oracle, as predicted by Theorem 5.4.1.

5.5.2 Implementation Details

The experiments were conducted on a MacBook Pro with 2.6 GHz Dual-Core Intel Core i5 processor and 8 GB of RAM. The experiments of type (A) respectively took 110m36.499s, 73m19.477s, 50m54.655s, and 536m39.212s to run. The experiments of type (B) respectively took 64m28.346s, 52m7.859s, 39m36.754s, and 362m50.742s to run. The discrepancy in runtimes between (A) and (B) can be explained by the fact that a number of unrelated processes were also running at the time of the former. The experiment of varying the number of batches N was the most expensive because we chose domain size $n = 128$ to accentuate the gap between our algorithm and the oracle. The abovementioned runtimes imply that over a domain of size 128, LearnWithFilter takes roughly 7-10 minutes.

For the implementation, we used the SCS solver in CVXPY for our semidefinite programs. In order to achieve reasonable runtimes, we needed to set the feasibility tolerance to $\texttt{1e} - 2$, and as a result the SDP solver would occasionally output matrices Σ which are moderately far from $\mathcal{K}$; in particular, one mode of failure that arose was that Σ might be non-PSD and give rise to negative scores in LearnWithFilter. We chose to address this mode of failure heuristically by terminating the algorithm whenever this happened and simply outputting the estimate for μ at that point in time. Of the 480 total trials that were run across all experiments, this happened 53 times. Another heuristic that we used was to terminate the algorithm as soon as $\|\Sigma\|_{\mathcal{K}}$ stopped increasing during a run of LearnWithFilter; this was primarily to have a stopping criterion that avoids the need to tune constant factors. As demonstrated by Figures 5-1 and 5-2, these heuristic decisions ultimately had negligible effect on the performance of our algorithm.

All code, data, and documentation can be found at https://github.com/secanth/federated.

5.6 Appendix: Concentration

In this section we prove Lemma 5.4.6, restated here for convenience:

Lemma 5.4.6 (Regularity of good samples)**.** *If U is a set of $\widetilde{\Omega}\left(\log(1/\delta)(\ell^2/\varepsilon^2)\cdot\log^3(n)\right)$ independent samples from $Mul_k(\mu_1),...,Mul_k(\mu_{|U|})$, then U is ε-good with probability at least $1-\delta$.*

5.6.1 Technical Ingredients

The key technical fact we use to get sample complexity that depend quadratically on ℓ is:

Lemma 5.6.1. *For every $0<\eta\le 1$, there exists a net $\mathcal{N}\subset\mathbb{R}^{n\times n}$ of size $O(n^3\ell^2\log^2 n/\eta)^{(\ell\log n+1)^2}$ of matrices such that for every $\Sigma\in\mathcal{K}$, there exists some $\tilde{\Sigma}=\sum_\nu\Sigma^*_\nu$ for $\Sigma^*_\nu\in\mathcal{N}$ such that the following holds: 1) $\|\Sigma-\tilde{\Sigma}\|_F\le\eta$, 2) $\sum_\nu\alpha_\nu\le 1$, and 3) $\|\Sigma^*_\nu\|_{\max}\le O(1)$.*

Note that this is a strengthening of a special case of Lemma 4.7.2 from the previous chapter. We defer the proof of Lemma 5.6.1 to Appendix 5.7.

For ε-goodness to hold, it will be crucial to establish the following sub-exponential tail bounds for the empirical covariance of a set of samples $X_1,\cdots,X_N\sim\mathrm{Mul}_k(\mu)$, as well as for $\|\hat{\mu}-\mu\|^2_{\mathcal{K}}$, where $\hat{\mu}$ is the empirical mean of those samples.

Lemma 5.6.2. *Let $\xi>0$ and let $\mathcal{N}\subset\mathbb{R}^{n\times n}$ be any finite set for which $\|\Sigma\|_{\max}\le O(1)$ for all $\Sigma\in\mathcal{N}$. Let $\mu_1,...,\mu_N,\overline{\mu}\in\Delta^n$ satisfy $\overline{\mu}\triangleq\frac{1}{N}\sum_{i=1}^N\mu_i$. Then for $X_i\sim Mul_k(\mu_i)$ for $i\in[N]$,*

$$Pr\left[\left|\left\langle\frac{1}{N}\sum_{i=1}^N(X_i-\mu_i)^{\otimes 2}-\mathbb{E}_{X\sim Mul_k(\mu_i)}\left[(X-\mu_i)^{\otimes 2}\right],\Sigma\right\rangle\right|>t\ \forall\ \Sigma\in\mathcal{N}\right]<2|\mathcal{N}|\exp\left(-\Omega\left(\frac{Nk^2t^2}{1+kt}\right)\right),$$

where the probability is over the samples $X_1,\cdots,X_N$.

Lemma 5.6.3. *Let $\xi > 0$ and let $\mathcal{N} \subset \mathbb{R}^{n\times n}$ be any finite set for which $\|\Sigma\|_{\max} \le O(1)$ for all $\Sigma \in \mathcal{N}$. For $X_i \sim Mul_k(\mu_i)$ for $i \in [N]$, $\hat{\mu} \triangleq \frac{1}{N}\sum_{i=1}^N X_i$, and $\overline{\mu} \triangleq \frac{1}{N}\sum_{i=1}^N \mu_i$,*

$$Pr\left[\left|\left\langle(\hat{\mu}-\mu)^{\otimes 2}, \Sigma\right\rangle - \mathbb{E}\left[\left\langle(\hat{\mu}-\mu)^{\otimes 2}, \Sigma\right\rangle\right]\right| > t\ \forall\ \Sigma \in \mathcal{N}\right] < 2|\mathcal{N}|\exp\left(-\Omega\left(\frac{N^2k^2t^2}{1+Nkt}\right)\right),$$

where the probability is over the samples $X_1, \cdots, X_N$.

Lemma 5.6.4. *Let $\xi > 0$ and let $\mathcal{N} \subset \mathbb{R}^{n\times n}$ be any finite set for which $\|\Sigma\|_{\max} \le O(1)$ for all $\Sigma \in \mathcal{N}$. Let $\mu_1, ..., \mu_N, \overline{\mu} \in \Delta^n$ satisfy $\|\mu_i - \overline{\mu}\|_1 \le \omega$ for all $i \in [N]$. For $X_i \sim Mul_k(\mu_i)$ for $i \in [N]$,*

$$Pr\left[\left|\frac{1}{N}\sum_{i=1}^N(\mu_i-\overline{\mu})^\top\Sigma(X_i-\mu_i)\right| > \omega\cdot t\ \forall\ \Sigma \in \mathcal{N}\right] < 2|\mathcal{N}|\exp\left(-\Omega\left(kNt^2\right)\right),$$

where the probability is over the samples $X_1, \cdots, X_N$.

Note that if $\mathcal{N}$ consisted solely of matrices of the form $vv^\top$ for $v \in \{\pm 1\}^n$, these lemmas would follow straightforwardly from standard binomial tail bounds. Instead, we only have entrywise bounds for the matrices in $\mathcal{N}$ and will therefore need to compute moment estimates from scratch in order to prove Lemmas 5.6.2 and 5.6.3. We defer the details of this to Appendix 5.8.

Lastly, we will need the following elementary consequence of Stirling's formula:

Fact 5.6.5. *For any $m \ge 1$, $\log\binom{m}{\varepsilon m} \le 2m\cdot\varepsilon\log 1/\varepsilon$.*

5.6.2 Proof of Lemma 5.4.6

We are now ready to prove that the four conditions for ε-goodness hold for a set U of independent draws from $\mathrm{Mul}_k(\mu_1), ..., \mathrm{Mul}_k(\mu_{|U|})$ respectively, of size

$$|U| = \widetilde{\Omega}\left(\log(1/\delta)(\ell^2/\varepsilon^2)\cdot\log^3(n)\right). \tag{5.25}$$

Proof of Lemma 5.4.6. As $\|\cdot\|_{\mathcal{K}}$ is defined as a supremum over $\mathcal{K}$, we will reduce controlling the infinitely many directions in $\mathcal{K}$ to controlling a finite net of such directions by invoking Lemma 5.6.1. Specifically, recall that for any $\Sigma \in \mathcal{K}$, by Lemma 5.6.1, there is some $\tilde{\Sigma} =$

$\sum_\nu \alpha_\nu \Sigma^*_\nu$ such that $\Sigma^*_\nu \in \mathcal{N}$ and $\|\Sigma - \tilde{\Sigma}\|_F \le \eta$.

(**Condition (I)**) By Lemma 5.6.3, with probability at least $1 - 2|\mathcal{N}|\exp\left(-\Omega(\frac{N^2k^2t^2}{1+Nkt})\right)$, we have that for all $\Sigma \in \mathcal{K}$,

$$\begin{aligned}
\left\langle (\mu(U) - \overline{\mu})^{\otimes 2}, \Sigma \right\rangle &\le \left\langle (\mu(U) - \overline{\mu})^{\otimes 2}, \tilde{\Sigma} \right\rangle + \|\mu(U) - \overline{\mu}\|_2^2 \cdot \|\Sigma - \tilde{\Sigma}\|_F \\
&\le \left\langle (\mu(U) - \overline{\mu})^{\otimes 2}, \tilde{\Sigma} \right\rangle + 2\eta \\
&= \sum_\nu \alpha_\nu \left\langle (\mu(U) - \overline{\mu})^{\otimes 2}, \Sigma^*_\nu \right\rangle + 2\eta \\
&\le \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\langle (X - \mu_i)^{\otimes 2}, \Sigma^*_\nu \right\rangle\right] + \sum_\nu \alpha_\nu \cdot t + 2\eta \\
&\le O(1/k|U|) + t + 2\eta, \qquad (5.26)
\end{aligned}$$

where the first step follows by Cauchy-Schwarz and triangle inequality, the second step follows by the trivial bound $\|\mu(U) - \mu_i\|_2^2 \le 2$ and the bound on $\|\Sigma - \tilde{\Sigma}\|_F$ guaranteed by Lemma 5.6.1, the fourth step holds with the claimed probability by Lemma 5.6.3 and the fact that $\|\Sigma^*_\nu\|_{\max} \le O(1)$ for all ν by the guarantees of Lemma 5.6.1, and the last step follows by the bound on $\sum \alpha_\nu$ by the guarantees of Lemma 5.6.1, as well as the moment bound in Lemma 5.8.2 applied to $r = 1$.

If $|U|$ satisfies (5.25) and $\eta, t = O(\frac{\varepsilon^2}{k}\log 1/\varepsilon)$, the first part of Condition (I) holds.

For the second part, by the steps leading to (5.26), a union bound over the $\binom{|U|}{\varepsilon|U|}$ subsets W and Fact 5.6.5, with probability at least

$$1 - 2\exp(2|U| \cdot \varepsilon \log 1/\varepsilon) \cdot |\mathcal{N}| \exp\left(-\Omega\left(\frac{\varepsilon^2|U|^2k^2t^2}{1 + \varepsilon|U|kt}\right)\right)$$

we have that $\|\mu(W) - \overline{\mu}_W\|_{\mathcal{K}}^2 \le O\left(\frac{1}{\varepsilon k|U|}\right) + t + 2\eta$ for all W. Note that $2\log 1/\varepsilon \le O\left(\frac{\varepsilon|U|^2k^2t^2}{1+\varepsilon|U|kt}\right)$ provided $t = \Omega\left(\frac{\log 1/\varepsilon}{k}\right)$, so if $|U|$ satisfies (5.25) and $\eta = O(\frac{\log 1/\varepsilon}{k})$, the second part of Condition (I) holds.

(**Condition (II)**) For the first part, let $\hat{\mathbf{M}} \triangleq M(\hat{w}(U), \{\mu_i\}_{i\in U})$. By Lemma 5.6.2, with probability at least $1 - 2|\mathcal{N}| \exp\left(-\Omega\left(\frac{|U|k^2t^2}{1+kt}\right)\right)$, we have that for all $\Sigma \in \mathcal{K}$,

$$\begin{aligned}
\langle \hat{\mathbf{M}}, \Sigma \rangle &\le \langle \hat{\mathbf{M}}, \tilde{\Sigma} \rangle + \|\hat{\mathbf{M}}\|_F \cdot \|\Sigma - \tilde{\Sigma}\|_F \\
&\le \langle \hat{\mathbf{M}}, \tilde{\Sigma} \rangle + 3\eta \\
&\le \sum_\nu \alpha_\nu \langle \hat{\mathbf{M}}, \Sigma^*_\nu \rangle + 3\eta \\
&\le \sum_\nu \alpha_\nu \cdot t + 3\eta \\
&\le t + 3\eta
\end{aligned} \tag{5.27}$$

where the first step follows by Cauchy-Schwarz and triangle inequality, and the second step follows by Lemma 5.2.3 and the bound on $\|\Sigma - \tilde{\Sigma}\|_F$ guaranteed by Lemma 5.6.1, the fourth step holds with the claimed probability by Lemma 5.6.2 and the fact that $\|\Sigma^*_\nu\|_{\max} \le O(1)$ for all ν by the guarantees of Lemma 5.6.1, and the last step follows by the bound on $\sum \alpha_\nu$ by the guarantees of Lemma 5.6.1.

If $|U|$ satisfies (5.25), $\eta = O\left(\frac{\varepsilon}{k}\log 1/\varepsilon\right)$, $t = O\left(\frac{\varepsilon}{k}\log 1/\varepsilon\right)$, the first part of Condition (II) holds.

For the second part, first note that it is slightly different from the first part because we do not subtract out $B(\overline{\mu})$, the reason being that $\|B(\overline{\mu})\|_{\mathcal{K}} \le O(1/k) = o(\frac{\log 1/\varepsilon}{k})$, so this term is negligible. By the steps leading to (5.27), a union bound over the $\binom{|U|}{\varepsilon|U|}$ subsets W, and Fact 5.6.5, with probability at least

$$1 - 2|\mathcal{N}| \exp(2\varepsilon|U|\log 1/\varepsilon) \cdot \exp\left(-\Omega\left(\frac{\varepsilon|U|k^2t^2}{1+kt}\right)\right),$$

we have that $\|M(\hat{w}(W), \{\mu_i\}_{i\in W})\|_{\mathcal{K}} \le t + 3\eta$ for all W. Note that $2\log 1/\varepsilon \le O\left(\frac{k^2t^2}{1+kt}\right)$ provided $t = \Omega\left(\frac{\log 1/\varepsilon}{k}\right)$, so if $|U|$ satisfies (5.25) and $\eta = O\left(\frac{\log 1/\varepsilon}{k}\right)$, the second part of Condition (II) holds.

(**Condition (III)**) First note that

$$B(\{\mu_i\}) - B(\overline{\mu}) = \frac{1}{|U|}\sum_{i\in U}\frac{1}{k}\left(\mathrm{diag}(\mu_i - \overline{\mu}) - (\mu_i^{\otimes 2} - \overline{\mu}^{\otimes 2})\right) = -\frac{1}{|U|}\sum_{i\in U}\frac{1}{k}(\mu_i^{\otimes 2} - \overline{\mu}^{\otimes 2}).$$

Also note that

$$\left\langle \Sigma, \frac{1}{|U|}\sum_{i\in U}(\mu_i^{\otimes 2} - \overline{\mu}^{\otimes 2})\right\rangle = \frac{1}{|U|}\sum_{i\in U}\left\langle(\mu_i - \overline{\mu})^{\otimes 2}, \Sigma\right\rangle \leq \max_i \|\mu_i - \overline{\mu}\|_1^2 \leq \omega^2,$$

where in the last step we used Fact 5.3.4. So $\|B(\{\mu_i\}) - B(\overline{\mu})\|_{\mathcal{K}} \leq \omega^2/k$.

It remains to bound $\|B(\hat{\mu}(U)) - B(\overline{\mu})\|_{\mathcal{K}}$. As we only need to show extremely mild concentration here, we will not make an effort to obtain tight bounds. Note that by (5.2),

$$|\langle \Sigma, B(\hat{\mu}(U)) - B(\overline{\mu})\rangle| \leq \frac{1}{k}\left|\langle \mathrm{diag}(\hat{\mu}(U) - \overline{\mu}), \Sigma\rangle\right| + \frac{1}{k}\left|\langle \hat{\mu}(U)^{\otimes 2} - \overline{\mu}^{\otimes 2}, \Sigma\rangle\right|. \tag{5.28}$$

We have

$$\begin{aligned}\langle \mathrm{diag}(\hat{\mu}(U) - \overline{\mu}), \Sigma\rangle &\leq \sum_\nu \alpha_\nu \langle \mathrm{diag}(\hat{\mu}(U) - \overline{\mu}), \Sigma^*_\nu\rangle + \|\Sigma - \tilde{\Sigma}\|_F \cdot \|\hat{\mu}(U) - \overline{\mu}\|_2 \\ &\leq \sum_\nu \alpha_\nu \langle \hat{\mu}(U) - \overline{\mu}, \mathrm{diag}(\Sigma^*_\nu)\rangle + O(\eta). \end{aligned} \tag{5.29}$$

Note that for any ν, $\langle \hat{\mu}(U) - \overline{\mu}, \mathrm{diag}(\Sigma^*_\nu)\rangle = \frac{1}{|U|}\sum_{i\in U} Z_i^\nu$ for $Z_i^\nu \triangleq \langle X_i - \mu_i, \mathrm{diag}(\Sigma^*_\nu)\rangle$. These are independent, mean-zero, $O(1)$-bounded random variables, so by Hoeffding's, for any fixed ν we have that $|\langle \hat{\mu}(U) - \overline{\mu}, \mathrm{diag}(\Sigma^*_\nu)\rangle| \leq t$ with probability at least $1 - 2\exp(-\Omega(|U|t^2))$. If we union bound over $\mathcal{N}$, then by taking $\eta, t = O(\varepsilon)$, and $|U|$ satisfying (5.25), (5.29) will be at most $O(\varepsilon)$.

We also have that

$$\begin{aligned}\left|\langle \hat{\mu}(U)^{\otimes 2} - \overline{\mu}^{\otimes 2}, \Sigma\rangle\right| &= \left|\langle (\hat{\mu}(U) - \overline{\mu})^{\otimes 2}, \Sigma\rangle - 2\overline{\mu}^\top \Sigma(\hat{\mu}(U) - \overline{\mu})\right| \\ &\leq O\left(\frac{\varepsilon^2 \log 1/\varepsilon}{k}\right) + 2\left|\overline{\mu}^\top \Sigma(\hat{\mu}(U) - \overline{\mu})\right|, \end{aligned} \tag{5.30}$$

where the second step follows by the first part of this lemma. For the other term, we have

$$\begin{aligned}\overline{\mu}^\top \Sigma(\hat{\mu}(U) - \overline{\mu}) &\leq \sum_\nu \alpha_\nu \overline{\mu}^\top \Sigma^*_\nu(\hat{\mu}(U) - \overline{\mu}) + \|\Sigma - \tilde{\Sigma}\|_F \cdot \|\overline{\mu}\|_2 \cdot \|\hat{\mu}(U) - \overline{\mu}\|_2 \\ &\leq \sum_\nu \alpha_\nu \overline{\mu}^\top \Sigma^*_\nu(\hat{\mu}(U) - \overline{\mu}) + O(\eta). \end{aligned} \tag{5.31}$$

For any ν, $\overline{\mu}^\top \Sigma^*_\nu(\hat{\mu}(U) - \overline{\mu}) = \frac{1}{|U|}\sum_{i\in U} W^\nu_i$ for $W^\nu_i \triangleq \overline{\mu}^\top \Sigma^*_\nu (X_i - \mu_i)$. These are independent, mean-zero, $O(1)$-bounded random variables, so by Hoeffding's, for any fixed ν, we have that $|\overline{\mu}^\top \Sigma(\hat{\mu}(U) - \overline{\mu})| \le t$ with probability at least $1 - 2\exp(-\Omega(|U|t^2))$. If we union bound over $\mathcal{N}$, then by taking $\eta, t = O(\varepsilon)$ and $|U|$ satisfying (5.25) again, (5.31) and thus (5.30) will be at most $O(\varepsilon)$.

By (5.28), we thus conclude that $\|B(\hat{\mu}(U)) - B(\overline{\mu})\|_{\mathcal{K}} \le O(\varepsilon/k)$ as claimed.

(**Condition (IV)**) By Lemma 5.6.4, with probability at least $1 - 2|\mathcal{N}|\exp\left(-\Omega\left(k|U|t^2\right)\right)$, we have that for all $\Sigma \in \mathcal{K}$,

$$\begin{aligned}
&\frac{1}{|U|}\sum_{i\in U}(\mu_i - \overline{\mu})^\top \Sigma (X_i - \mu_i) \\
&\quad\le \frac{1}{|U|}\sum_{i\in U}(\mu_i - \overline{\mu})^\top \tilde{\Sigma}(X_i - \mu_i) + \frac{1}{|U|}\sum_{i\in U}\|\Sigma - \tilde{\Sigma}\|_F \cdot \|\mu_i - \overline{\mu}\|_2 \cdot \|X_i - \mu_i\|_2 \\
&\quad\le \sum_\nu \alpha_\nu \cdot \frac{1}{|U|}\sum_{i\in U}(\mu_i - \overline{\mu})^\top \Sigma^*_\nu (X_i - \mu_i) + 2\omega\cdot\eta \\
&\quad\le \sum_\nu \alpha_\nu \cdot t + 2\omega \cdot \eta \\
&\quad\le \omega \cdot t + 2\omega\cdot\eta
\end{aligned} \tag{5.32}$$

where the first step follows by triangle inequality and Cauchy-Schwarz, the second step follows by the bound on $\|\Sigma - \tilde{\Sigma}\|_F$ guaranteed by Lemma 5.6.1 and the assumption that $\|\mu_i - \overline{\mu}\|_2 \le \omega$, and the third step holds with the claimed probability by Lemma 5.6.4 and the fact that $\|\Sigma^*_\nu\|_{\max} \le O(1)$ for all ν by Lemma 5.6.1, and the last step follows by the bound on $\sum \alpha_\nu$ by the guarantees of Lemma 5.6.1. If $|U|$ satisfies (5.25) and $\eta, t = O\left(\frac{\varepsilon\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right)$, the first part of Condition (IV) holds.

For the second part, by the steps leading to (5.32), a union bound over W, and Fact 5.6.5, with probability at least

$$1 - 2|\mathcal{N}|\exp(2\varepsilon|U|\log 1/\varepsilon)\cdot\exp\left(-\Omega\left(\varepsilon k|U|t^2\right)\right),$$

we have that $\frac{1}{|W|}\sum_{i\in W}(\mu_i - \overline{\mu})^\top\Sigma(X_i - \mu_i) \le \omega\cdot t + 2\omega\cdot\eta$ for all W.

Note that $2\log 1/\varepsilon \le O(kt^2)$ provided $t = \Omega\left(\frac{\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right)$, so if $|U|$ satisfies (5.25) and

$\eta = O\left(\frac{\sqrt{\log 1/\varepsilon}}{\sqrt{k}}\right)$, the second part of Condition (IV) holds. □

5.7 Appendix: Netting Over $\mathcal{K}$

In this section we prove Lemma 5.6.1, restated here for convenience:

Lemma 5.6.1. *For every $0 < \eta \le 1$, there exists a net $\mathcal{N} \subset \mathbb{R}^{n\times n}$ of size $O(n^3\ell^2\log^2 n/\eta)^{(\ell\log n+1)^2}$ of matrices such that for every $\Sigma \in \mathcal{K}$, there exists some $\tilde{\Sigma} = \sum_\nu \Sigma^*_\nu$ for $\Sigma^*_\nu \in \mathcal{N}$ such that the following holds: 1) $\|\Sigma - \tilde{\Sigma}\|_F \le \eta$, 2) $\sum_\nu \alpha_\nu \le 1$, and 3) $\|\Sigma^*_\nu\|_{\max} \le O(1)$.*

As alluded to in Remark 5.3.2 and Appendix 5.6, we will use the extra Constraints 3 and 4 in the definition of $\mathcal{K}$ to tighten the proof of Lemma 4.7.2 to obtain Lemma 5.6.1 above.

The following well-known trick will be useful.

Lemma 5.7.1 ("Shelling"). *If $v \in \mathbb{R}^m$ satisfies $\|v\|_2 \le C$ and $\|v\|_1 = C\cdot\sqrt{k}$, then there exist k-sparse vectors $v[1], ..., v[m/k]$ with disjoint supports for which 1) $v = \sum_{i=1}^{m/k} v[i]$, 2) $\sum_{i=1}^{m/k}\|v[i]\|_2 \le 2C$, and 3) $\sum_{i=1}^{m/k}\|v[i]\|_\infty \le \frac{1}{k}\|v\|_1 + \|v\|_\infty$.*

Proof. Assume without loss of generality that $C = 1$. Letting $B_1 \subset [m]$ be the indices of the k largest entries of v in absolute value, B_2 those of the next k largest, etc., we can write $[m] = B_1 \sqcup \cdots \sqcup B_{m/k}$. For $i \in [m/k]$, define $v[i] \in \mathbb{R}^m$ to be the restriction of v to the coordinates indexed by B_i. For any i and $j \in B_i$, $|v_j| \le \frac{1}{k}\|v[i-1]\|_1$. This immediately implies that

$$\sum_{i=1}^{m/k}\|v[i]\|_\infty \le \|v\|_\infty + \frac{1}{k}\sum_{i=1}^{m/k}\|v[i]\|_1,$$

yielding 3) above. Likewise, it implies that

$$\|v[i]\|_2^2 = \sum_{j\in B_i} v_j^2 \le k\cdot\frac{1}{k^2}\cdot\|v[i-1]\|_1^2 = \frac{1}{k}\|v[i-1]\|_1^2.$$

So $\|v[i]\|_2 \le \|v[i-1]\|_1/\sqrt{k}$ and thus

$$\sum_{i=1}^{m/k}\|v[i]\|_2 \le \|v[1]\|_2 + \frac{1}{\sqrt{k}}\|v\|_1 \le 2,$$

giving 2) above. □

By rescaling the entries of v in Lemma 5.7.1, we immediately get the following extension to Haar-weighted norms:

Corollary 5.7.2. *If $v \in \mathbb{R}^m$ satisfies $\|v\|_{2;\mathbf{h}} \le C$ and $\|v\|_{1;\mathbf{h}} = C \cdot \sqrt{k}$, then there exist k-sparse vectors $v_1, ..., v_{m/k}$ with disjoint supports for which 1) $v = \sum_{i=1}^{m/k} v_i$, 2) $\sum_{i=1}^{m/k} \|v_i\|_{2;\mathbf{h}} \le 2C$, and 3) $\sum_{i=1}^{m/k} \|v[i]\|_{\infty;\mathbf{h}} \le \frac{1}{k}\|v\|_{1;\mathbf{h}} + \|v\|_{\infty;\mathbf{h}}$.*

We remark that whereas in the proof of Lemma 4.7.2, shelling was applied to the unweighted L_1, L_2 norms, and the only L_2 information used about $v \in \mathcal{V}^n_\ell$ was that $\|v\|_2^2 = n$, in the sequel we will shell under the Haar-weighted norms and use the refined bounds on the Haar-weighted norms given by Constraints 3 and 4 from Definition 5.3.1. This will be crucial to getting a net of size exponential in ℓ^2 rather than just $\mathrm{poly}(\ell)$.

We now complete the proof of Lemma 5.6.1.

Proof of Lemma 5.6.1. Let $s = \ell \log n + 1$, and let $m = \log n$. Let $\mathcal{N}'$ be an $O\left(\frac{\eta}{n \cdot s^2}\right)$-net in Frobenius norm for all s^2-sparse $n \times n$ matrices of unit Frobenius norm. Because $\mathbb{S}^{s^2-1}$ has an $O\left(\frac{\eta}{n \cdot s^2}\right)$-net in L_2 norm of size $O(n \cdot s^2/\eta)^{s^2}$, by a union bound we have that

$$|\mathcal{N}'| \le \binom{n^2}{s^2} \cdot O(n \cdot s^2/\eta)^{s^2} = O(n^3 \ell^2 \log^2 n/\eta)^{s^2}$$

Take any $\Sigma \in \mathcal{K}$ and consider $\mathbf{L} \triangleq H \Sigma H^\top$. By Constraints 2, 3, 4 in Definition 5.3.1,

$$\|\mathbf{L}\|_{1,1;\mathbf{h}} \le s^2, \qquad \|\mathbf{L}\|_{F;\mathbf{h}}^2 \le s^2, \qquad \text{and} \qquad \|\mathbf{L}\|_{\max;\mathbf{h}} \le 1. \tag{5.33}$$

We can use the first two of these and apply Corollary 5.7.2 to the n^2-dimensional vector $\mathbf{L}$ to conclude that $\mathbf{L} = \sum_j \mathbf{L}^j$ for some matrices $\{\mathbf{L}^j\}_j$ of sparsity at most s^2 and for which $\sum_j \|\mathbf{L}^j\|_{F;\mathbf{h}} \le 2s^2$ and $\sum_j \|\mathbf{L}^j\|_{\max;\mathbf{h}} \le \frac{1}{s^2}\|\mathbf{L}^j\|_{1,1;\mathbf{h}} + \|\mathbf{L}^j\|_{\max;\mathbf{h}}$.

By definition of the Haar-weighted Frobenius norm, $\|\mathbf{L}^j\|_F \le n \cdot \|\mathbf{L}^j\|_{F,\mu}$, so

$$\sum_j \|\mathbf{L}^j\|_F \le O(n \cdot s^2).$$

For each $\mathbf{L}^j$, there is some $(\mathbf{L}')^j \in \mathcal{N}'$ such that for $\tilde{\mathbf{L}}^j \triangleq \|\mathbf{L}^j\|_F \cdot (\mathbf{L}')^j$,

$$\|\mathbf{L}^j - \tilde{\mathbf{L}}^j\|_F \le O\left(\frac{\eta}{n \cdot s^2}\right) \|\mathbf{L}^j\|_F. \tag{5.34}$$

We conclude that if we define $\tilde{\mathbf{L}} \triangleq \sum_j \tilde{\mathbf{L}}^j$, then $\|\mathbf{L} - \tilde{\mathbf{L}}\|_F \le \eta$.

Now let $\mathcal{N} \triangleq H^{-1}\mathcal{N}^{-1}(H^{-1})^\top$. As $\Sigma = H^{-1}\mathbf{L}(H^{-1})^\top$ and H^{-1} is an isometry, if we define $\tilde{\Sigma}^j \triangleq H^{-1}\tilde{\mathbf{L}}^j(H^{-1})^\top$ and $\tilde{\Sigma} \triangleq \sum_j \tilde{\Sigma}^j$, then we likewise get that $\|\Sigma - \tilde{\Sigma}\|_F \le \eta$, and clearly $\tilde{\Sigma}^j \in \mathbb{P}\mathcal{N}$ for every j, concluding the proof of part 1) of the lemma.

For each $\tilde{\Sigma}^j$, define

$$\alpha_j \triangleq \|\mathbf{L}^j\|_{\max;\mathbf{h}}/2 \tag{5.35}$$

and define $\Sigma^j_* \triangleq \tilde{\Sigma}^j/\alpha_j$ so that $\tilde{\Sigma} = \sum_{j,\sigma,\tau} \alpha_j \cdot \Sigma^j_*$. Note that by part 3) of Corollary 5.7.2 and (5.33),

$$\begin{aligned}\sum_j \alpha_j &= \frac{1}{2}\sum_j \|\mathbf{L}^j\|_{\max;\mathbf{h}} \\ &\le \frac{1}{2}\frac{1}{s^2}\|\mathbf{L}\|_{1,1;\mathbf{h}} + \|\mathbf{L}\|_{\max;\mathbf{h}} \le 1\end{aligned}$$

where in the last step we used the fact that $\|\mathbf{L}\|_{1,1;\mathbf{h}} \le s^2$ and $\|\mathbf{L}[\sigma,\tau]\|_{\max;\mathbf{h}} \le 1$. This concludes the proof of part 2) of the lemma.

Finally, we need to bound $\|\Sigma^j_*\|_{\max}$. Note first that for any matrix $\mathbf{J}$ supported only on a submatrix consisting of entries of $\mathbf{L}$ from the rows i (resp. columns j) for which $i \in T_\sigma$ (resp. $j \in T_\tau$), we have that

$$\|H^{-1}\mathbf{J}(H^{-1})^\top\|_{\max} = 2^{-(m-\sigma)/2} \cdot 2^{-(m-\tau)/2} \cdot \|\mathbf{J}\|_{\max} = \frac{2^{(\sigma+\tau)/2}}{n}\|\mathbf{J}\|_{\max}$$

because the Haar wavelets $\{\psi_{\sigma,j}\}_j$ (resp. $\{\psi_{\tau,j}\}_j$) have disjoint supports and L_∞ norm $2^{-(m-\sigma)/2}$ (resp. $2^{-(m-\tau)/2}$). For general $\mathbf{J}$, by decomposing $\mathbf{J}$ into such submatrices, call them $\mathbf{J}[\sigma,\tau]$, we get by triangle inequality that

$$\|H^{-1}\mathbf{J}(H^{-1})^\top\|_{\max} \le \sum_{\sigma,\tau} \frac{2^{(\sigma+\tau)/2}}{n}\|\mathbf{J}[\sigma,\tau]\|_{\max} \le \|\mathbf{J}\|_{\max}. \tag{5.36}$$

By applying this to $\mathbf{J} = \tilde{\Sigma}^j$, we get

$$\begin{aligned}
\|\tilde{\Sigma}^j\|_{\max} &\leq \left(\|H^{-1}\mathbf{L}^j(H^{-1})^\top\|_{\max} + \|H^{-1}\left(\mathbf{L}^j - \tilde{\mathbf{L}}^j\right)(H^{-1})^\top\|_{\max}\right) \\
&\leq \|\mathbf{L}^j\|_{\max} + \|\mathbf{L}^j - \tilde{\mathbf{L}}^j\|_{\max} \\
&\leq \|\mathbf{L}^j\|_{\max} + \|\mathbf{L}^j - \tilde{\mathbf{L}}^j\|_F \\
&\leq \|\mathbf{L}^j\|_{\max} + O\left(\frac{\eta}{n \cdot s^2}\right)\|\mathbf{L}^j\|_F \\
&\leq \|\mathbf{L}^j\|_{\max} \cdot (1 + O(\eta/n)) \\
&\leq 2 \cdot \|\mathbf{L}^j\|_{\max},
\end{aligned}$$

where the first inequality is triangle inequality, the second inequality follows by (5.36), the third inequality follows from monotonicity of L_p norms, the fourth inequality follows from (5.34), and the fifth inequality follows from the fact that $\mathbf{L}^j$ is s^2 sparse.

Recalling (5.35) and the definition of $\Sigma_*^{\sigma,\tau;j}$, we conclude that $\|\Sigma_*^{\sigma,\tau;j}\|_{\max} \leq O(1)$ as claimed. □

5.8 Appendix: Sub-Exponential Tail Bounds From Section 5.6

In this section, we provide proofs for Lemmas 5.6.2, 5.6.3, and 5.6.4, restated here for convenience.

Lemma 5.6.2. *Let $\xi > 0$ and let $\mathcal{N} \subset \mathbb{R}^{n\times n}$ be any finite set for which $\|\Sigma\|_{\max} \leq O(1)$ for all $\Sigma \in \mathcal{N}$. Let $\mu_1, ..., \mu_N, \overline{\mu} \in \Delta^n$ satisfy $\overline{\mu} \triangleq \frac{1}{N}\sum_{i=1}^N \mu_i$. Then for $X_i \sim Mul_k(\mu_i)$ for $i \in [N]$,*

$$Pr\left[\left|\left\langle \frac{1}{N}\sum_{i=1}^N (X_i - \mu_i)^{\otimes 2} - \mathbb{E}_{X\sim Mul_k(\mu_i)}\left[(X - \mu_i)^{\otimes 2}\right], \Sigma\right\rangle\right| > t \ \forall\ \Sigma \in \mathcal{N}\right] < 2|\mathcal{N}|\exp\left(-\Omega\left(\frac{Nk^2t^2}{1+kt}\right)\right),$$

where the probability is over the samples $X_1, \cdots, X_N$.

Lemma 5.6.3. *Let $\xi > 0$ and let $\mathcal{N} \subset \mathbb{R}^{n\times n}$ be any finite set for which $\|\Sigma\|_{\max} \leq O(1)$ for* *all $\Sigma \in \mathcal{N}$. For $X_i \sim Mul_k(\mu_i)$ for $i \in [N]$, $\hat{\mu} \triangleq \frac{1}{N}\sum_{i=1}^N X_i$, and $\overline{\mu} \triangleq \frac{1}{N}\sum_{i=1}^N \mu_i$,*

$$Pr\left[\left|\left\langle(\hat{\mu}-\mu)^{\otimes 2}, \Sigma\right\rangle - \mathbb{E}\left[\left\langle(\hat{\mu}-\mu)^{\otimes 2}, \Sigma\right\rangle\right]\right| > t \ \forall\ \Sigma \in \mathcal{N}\right] < 2|\mathcal{N}| \exp\left(-\Omega\left(\frac{N^2k^2t^2}{1+Nkt}\right)\right),$$

where the probability is over the samples $X_1, \cdots, X_N$.

Lemma 5.6.4. *Let $\xi > 0$ and let $\mathcal{N} \subset \mathbb{R}^{n\times n}$ be any finite set for which $\|\Sigma\|_{\max} \le O(1)$ for all $\Sigma \in \mathcal{N}$. Let $\mu_1, ..., \mu_N, \overline{\mu} \in \Delta^n$ satisfy $\|\mu_i - \overline{\mu}\|_1 \le \omega$ for all $i \in [N]$. For $X_i \sim Mul_k(\mu_i)$ for $i \in [N]$,*

$$Pr\left[\left|\frac{1}{N}\sum_{i=1}^N (\mu_i - \overline{\mu})^\top \Sigma (X_i - \mu_i)\right| > \omega \cdot t \ \forall\ \Sigma \in \mathcal{N}\right] < 2|\mathcal{N}| \exp\left(-\Omega\left(kNt^2\right)\right),$$

where the probability is over the samples $X_1, \cdots, X_N$.

We remark that if we restricted our attention to test matrices of the form $\Sigma = vv^\top$ for $v \in \{\pm 1\}^n$, these lemmas would follow straightforwardly from Bernstein's and the sub-Gaussianity of binomial distributions.

We will need the following well-known combinatorial fact, a proof of which we include for completeness in Section 5.8.1

Fact 5.8.1. *For any $m, r \in \mathbf{Z}$, there are at most $O(m)^r \cdot r!$ tuples $(i_1, ..., i_{2r}) \in [m]^t$ for which every element of $[m]$ occurs an even (possibly zero) number of times.*

Central to the proofs of Lemmas 5.6.2 and 5.6.3 is the following sub-exponential moment bound. We remark that this moment bound would be an immediate consequence of McDiarmid's if Σ not only satisfied $\|\Sigma\|_{\max}$ but was also psd, but because the matrices arising from shelling need not be psd, it turns out to be unavoidable that we must prove this moment bound from scratch.

In this section, given $\mu \in \Delta^n$, let $\mathcal{D}_\mu$ denote the distribution over standard basis vectors $\{e_i\}$ of $\mathbb{R}^n$ where for any $i \in [n]$, e_i has probability mass equal to the i-th entry of μ.

Lemma 5.8.2. *Let $\Sigma \in \mathbb{R}^{n\times n}$ have entries bounded in absolute value by $O(1)$, and for $\mu_1, ..., \mu_m, \overline{\mu} \in \Delta^n$, let $\overline{\mu} \triangleq \frac{1}{m}\sum_{i=1}^m \mu_i$. If $Y_1, ..., Y_m$ are independent draws from $\mathcal{D}_{\mu_i}$ respectively, and $\hat{\mu} \triangleq \frac{1}{m}\sum_{i=1}^m Y_i$, then for every $r \ge 1$, $\mathbb{E}\left[\left((\hat{\mu}-\overline{\mu})^\top \Sigma (\hat{\overline{\mu}} - \mu)\right)^r\right] \le \Omega(m)^{-r} \cdot r!$.*

Proof. Without loss of generality, suppose Σ has entries bounded in absolute value by 1. For $i, i' \in [m]$, define $Z_{i,i'} \triangleq (Y_i - \mu_i)^\top \Sigma (Y_{i'} - \mu_{i'})$. Note that because $\|Y_i - \mu_i\|_1 \le 2$ with probability 1 for all $i \in [m]$, and the entries of Σ are bounded in absolute value by 1, $|Z_{i,i'}| \le 4$ with probability 1 for all $i, i' \in [m]$. We can write $\mathbb{E}\left[\left((\hat{\mu} - \overline{\mu})^\top \Sigma (\hat{\mu} - \overline{\mu})\right)^r\right]$ as

$$\frac{1}{m^{2r}} \mathbb{E}\left[\left(\sum_{i,i' \in [m]} Z_{i,i'}\right)^r\right] = \frac{1}{m^{2r}} \sum_{(i_1,i'_1),\dots,(i_r,i'_r)} \mathbb{E}\left[\prod_{j=1}^r Z_{i_j,i'_j}\right]. \tag{5.37}$$

Now that if there exists some index $i \in [m]$ which occurs an odd number of times among $i_1, i'_1, \dots, i_r, i'_r$, then by the fact that the tensor $\mathbb{E}\left[(Y_i - \mu_i)^{\otimes a}\right]$ is identically zero for odd a, we have that $\mathbb{E}\left[\prod_{j=1}^r Z_{i_j,i'_j}\right]$. So the nonzero summands on the right-hand side of (5.37) correspond to indices $\{(i_j, i'_j)\}_{j\in[r]}$ which must satisfy that every index appearing among $i_1, i'_1, \dots, i_r, i'_r$ appears an even number of times. By Fact 5.8.1, there are $O(m)^r \cdot r!$ such tuples.

Finally, by the fact that $|Z_{i,i'}| \le 4$ with probability 1 for all $i, i' \in [M]$, each monomial $\mathbb{E}\left[\prod_{j=1}^r Z_{i_j,i'_j}\right]$ is upper bounded by 4^r. We conclude that $\mathbb{E}\left[\left((\hat{\mu} - \overline{\mu})^\top \Sigma (\hat{\mu} - \overline{\mu})\right)^r\right] \le \frac{1}{m^{2r}} \cdot O(m)^r \cdot r! \cdot 4^r$, from which the claim follows. □

Similarly, a crucial ingredient to the proof of Lemma 5.6.4 is the following moment bound.

Lemma 5.8.3. *Let $\Sigma \in \mathbb{R}^{n\times n}$ have entries bounded in absolute value by $O(1)$, and suppose $\mu_1, \dots, \mu_m, \overline{\mu} \in \Delta^n$ satisfy $\|\mu_i - \mu_m\|_1 \le \omega$ for all $i \in [m]$. Then for every $r \in \mathbf{Z}$, $\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^m (\mu_i - \overline{\mu})^\top \Sigma (Y_i - \mu_i)\right)^r\right]$ is 0 if r is odd and at most $O(r\omega^2/m)^{r/2}$ otherwise.*

Proof. It is clear that the r-th moment is zero when r is odd. Henceforth, write r as $2r$. Without loss of generality, suppose Σ has entries bounded in absolute value by 1. For $i \in [m]$, define $Z_i \triangleq (\mu_i - \overline{\mu})^\top \Sigma (Y_i - \mu_i)$. Note that because $\|Y_i - \mu_i\|_1 \le 2$ with probability 1 for all $i \in [m]$, and the entries of Σ are bounded in absolute value by 1, $|Z_i| \le 2\omega$ with probability 1 for all $i \in [m]$. We can write $\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^m (\mu_i - \overline{\mu})^\top \Sigma (Y_i - \mu_i)\right)^{2r}\right]$ as

$$\frac{1}{m^r} \mathbb{E}\left[\left(\sum_{i \in [m]} Z_i\right)^r\right] = \frac{1}{m^r} \sum_{i_1,\dots,i_{2r}} \mathbb{E}\left[\prod_{j=1}^{2r} Z_{i_j}\right].$$

As in the proof of Lemma 5.8.2, the only nonzero summands correspond to tuples $(i_1, ..., i_{2r})$ such that every element of $[m]$ appears an even (possibly zero) number of times. By Fact 5.8.1, there are at most $O(m)^r \cdot r!$ such tuples, from which we can complete the proof. □

Lemmas 5.6.2 and 5.6.3 will now follow as consequences of Lemma 5.8.2 and the following standard tail bound for random variables with sub-exponential moments:

Fact 5.8.4. *Let $Z_1, ..., Z_m$ be random variables for which there exists a constant $\nu > 0$ such that $\mathbb{E}[Z_i^r] \le \frac{1}{2}\nu^r \cdot r!$ for all integers $r \ge 1$ and $i \in [m]$. Then*

$$Pr\left[\left|\frac{1}{m}\sum_{i=1}^m Z_i - \mathbb{E}[Z]\right| > t\right] \le 2e^{-\Omega\left(\frac{mt^2}{\nu^2+\nu t}\right)}.$$

Similarly, Lemma 5.6.4 will follow as a consequence of Lemma 5.8.3 and the following standard tail bound for random variables with sub-Gaussian moments:

Fact 5.8.5. *Let $Z_1, ..., Z_m$ be random variables for which there exists a constant $\nu > 0$ such that $\mathbb{E}[Z_i^r] \le (r \cdot \nu^2)^{r/2}$ for all integers $r \ge 1$ and $i \in [m]$. Then*

$$Pr\left[\left|\frac{1}{m}\sum_{i=1}^m Z_i - \mathbb{E}[Z]\right| > t\right] \le 2e^{-\Omega\left(mt^2/\nu^2\right)}.$$

Proof of Lemma 5.6.2. This follows by taking $m = k$ in Lemma 5.8.2 and $m = N$ in Fact 5.8.4 and noting that for any $\Sigma \in \mathcal{N}$, $\|\Sigma\|_{\max} \le O(1)$ by Lemma 5.6.1. □

Proof of Lemma 5.6.3. This follows by taking $m = kN$ in Lemma 5.8.2 and $m = 1$ in Fact 5.8.4 and noting that for any $\Sigma \in \mathcal{N}$, $\|\Sigma\|_{\max} \le O(1)$ by Lemma 5.6.1. □

Proof of Lemma 5.6.4. This follows by taking $m = k$ in Lemma 5.8.3 and $m = N$ in Fact 5.8.5 and noting that for any $\Sigma \in \mathcal{N}$, $\|\Sigma\|_{\max} \le O(1)$ by Lemma 5.6.1. □

5.8.1 Proof of Fact 5.8.1

Proof. To count the number N^* of such tuples $(i_1, ..., i_{2r})$, for every $1 \le s \le r$ let N_s denote the number of tuples $\beta \in \{2, 4..., 2r\}^s$ for which $\sum_{i=1}^s \beta_i = 2r$. By balls-and-bins, $N_s = \binom{r+s-1}{r} \le (\frac{3es}{2r})^r$. Now note that to enumerate N^*, we can 1) choose the number

$1 \leq s \leq \min(m,r)$ of unique indices among $\{i_j\}$, 2) choose a subset S of $[m]$ of size s, 3) choose one of the N_s tuples β, and 4) choose one of the $\binom{2r}{\beta_1,\ldots,\beta_s}$ ways of assigning index S_1 to β_1 indices in $\{i_j\}$, S_2 to β_2 indices, etc. For convenience, let $r' \triangleq \min(m,r)$. We get an upper bound of

$$\begin{aligned}
N^* &\leq \sum_{s=1}^{\min(m,r)} \binom{m}{s} \cdot N_s \cdot \binom{2r}{\beta_1, \ldots, \beta_s} \\
&\leq \sum_{s=1}^{\min(m,r)} \frac{m^s}{s!} \left(\frac{3es}{2r}\right)^r \cdot (2s)! \\
&\leq \frac{m^{r'}}{(r')!} \cdot r' \cdot \left(\frac{3er'}{2r}\right)^r \cdot (2r')! \\
&\leq \frac{m^r}{(r)!} \cdot r \cdot (3e/2)^r \cdot (2r)! \\
&= m^r \cdot r \cdot (3e/2)^r \cdot \binom{2r}{r} \cdot r! \\
&\leq O(m)^r \cdot r!,
\end{aligned}$$

where in the second step we used basic bounds on binomial and multinomial coefficients together with the above bound on N_s, in the third step we used the fact that the summands are increasing in s, and in the fourth step we used this fact along with the fact that $r' \leq r$ by definition. □

Chapter 6

Huber-Contaminated Regression and Contextual Bandits

6.1 Introduction

In this chapter, we turn to a classic question in the field of robust statistics, namely robust regression, as well as a number of robust analogues of related problems in online learning. While we have recently seen considerable progress in algorithmic robust statistics [DKK$^+$19a, LRV16, DKK$^+$17, CSV17, KKM18, DKK$^+$19b, HL18, KSS18, BK20, Kan20, DHKK20] that has yielded a number of exciting further applications to robust regression [KKM18, BP20, ZJS20, CAT$^+$20] and robust stochastic optimization [DKK$^+$19b], a shortcoming that all these works share is that they are based on assumptions that the uncorrupted data is somehow evenly spread out. These assumptions can either come about by explicitly assuming a generative model, like a Gaussian [DKK$^+$19a] or a mixture of Gaussians [BK20, Kan20, DHKK20], or through a deterministic condition like hypercontractivity [KKM18] or certifiable sub-Guassianity [HL18, KSS18].

Still, as we discussed in Section 1.1.2, there is a widespread need for provably robust learning algorithms even in settings where these types of "evenly spread out" assumptions are just not appropriate. This is particularly the case in the context of online prediction [CBL06] which operates in a setting where the input data is ever-changing and potentially even adversarially chosen. This flexibility allows it to capture challenging dynamic settings, as arise in reinforcement learning, where our learning algorithm interacts with the world around it and its decisions may in turn influence the next prediction task it is expected to solve. In this work we take an important first step towards answering a much broader question:

> *Are there provably robust learning algorithms that can tolerate adversarial corruptions even for challenging high-dimensional and distribution-free online prediction tasks?*

We will work in the Huber contamination model [Hub64]. We will study two classic online learning problems: online linear regression with squared loss and linear contextual bandits. In unsupervised learning settings, the Huber contamination model posits that each random sample we get has an η probability of coming from an arbitrary noise distribution chosen by an adversary instead of from our model. In our setting we will allow the feedback in each round to be arbitrarily corrupted with η probability, and otherwise is subject to the usual stochastic noise.

As we noted in the discussion immediately proceeding Definition 1.2.9, it turns out that for our problems the key challenge is to disentangle the effect of the *dynamic range* of predictions vs. the effect of the *noise level* on the overall regret guarantee. In particular, consider the basic linear regression problem where $(x_t)_{t=1}^T$ is the input sequence of covariate vectors[1] and our goal is to robustly predict the response y_t. Without adversarial corruptions, we assume the responses are generated according to the following well-specified model:

$$y_t = \langle w^*, x_t \rangle + \xi_t$$

where w^* is unknown and ξ_t is the noise, and our goal is to predict the clean, noiseless response $\langle w^*, x_t \rangle$ accurately. This problem is straightforward to solve with variants of Ordinary Least Squares [AW01, Vov01] even in the online setting. Now, consider what happens when we allow a random η fraction of the responses y_t to be adversarially corrupted, and our goal is to predict the *clean/uncorrupted responses* $\langle w^*, x_t \rangle$ accurately. Let R be the dynamic range

[1] In this paper, we will study the general case where these vectors are chosen adversarially and adaptively and the predictions are made online, but the importance of distinguishing dynamic range vs. noise level we discuss is relevant already in the basic (offline) setting.

of the true optimal predictions, so $|\langle w^*, x_t \rangle| \leq R$, and let σ^2 be the variance of ξ_t. When σ^2 is comparable to R^2, then the problem is relatively easy as there is (information-theoretically) not much that can be learned about w^* in the first place. See the left panel of Figure 1-1 for an illustration.

In contrast we will be interested in the setting where σ^2 is much smaller than R^2 (recall Figure 1-1). It turns out that existing approaches break down in the sense that they pay an extra factor of R or R^2 in the clean prediction error (resp. clean regret). Moreover getting around this dependence is a serious obstacle for the usual techniques: we show that regression using any convex surrogate (including Huber loss and L_1 loss) must pay this price (see Theorem 6.9.1). Thus our main question is:

> *Is it algorithmically possible, in the presence of adversarial corruptions, to achieve average clean prediction error (resp. average clean regret) that is independent of R?*

We answer this question in the affirmative for both online regression with squared loss and linear contextual bandits. Our algorithms succeed where convex surrogates fail, and are based on a novel alternating minimization scheme that interleaves OLS with carefully designed reweighting schemes found through SDPs.

Finally we emphasize that the issue of R^2 vs. σ^2 dependence is quite relevant in modern reinforcement learning. In particular, there are many sequential tasks where at each step the variance in the losses/rewards is much smaller than the dynamic range. This can happen naturally when there are some catastrophic states that we must avoid, but at no point is the outcome of playing an action in a given state all that uncertain – e.g. when manipulating a robotic arm, some actions can require the application of orders of magnitude more torque. Thus our work may be viewed as a stepping stone towards achieving stronger and more meaningful robustness guarantees in reinforcement learning more broadly.

6.1.1 Our Results

In this section, we present our main results for both linear regression and contextual bandits in the Huber contamination model. We go on to discuss related work (e.g. robust linear regression under distributional assumptions) in Section 6.3 below.

Distribution-free offline linear regression with Huber contamination. We begin by discussing our results in the simplest setting we consider, which is the classical *offline* linear regression model with a Huber contamination adversary. In the clean version of this model, an arbitrary set of covariates $x_1, \dots, x_n$ is fixed and clean responses are generated by

$$y_t = \langle w^*, x_t \rangle + \xi_t \tag{6.1}$$

for some mean zero noise ξ_t; for example, if $\xi_t \sim N(0, \sigma^2)$ then $y_t \sim N(\langle w^*, x_t \rangle, \sigma^2)$. In the Huber contamination model, we relax the assumptions to a total variation distance ball around the generative model. In particular, using the coupling interpretation of total variation distance, this translates into the assumption that with probability $1-\eta$ the response y_t is generated by (6.1) above, and with probability η the response y_t is sampled from an adversarially chosen noise distribution, which we allow to depend on all other randomness in the problem. In this setting, we obtain the following strong result (and for a fairly simple algorithm, see Technical Overview):

Theorem 6.1.1 (Informal version of Theorem 6.5.13 and Theorem 6.6.1)**.** *Suppose that $\eta < 0.499$ is an upper bound on the contamination level, and suppose for some $\sigma \geq 0$ that for all $1 \leq t \leq n$, $\|x_t\| \leq 1$ and the noise ξ_t is conditionally mean-zero and σ^2-subgaussian. Suppose also that $\|w^*\| \leq R$. Then if $\eta = 0$ or $n \gtrsim \log(\min(n, d))/\eta$, there exists a polynomial time algorithm outputting w satisfying the clean squared loss guarantee*

$$\sqrt{\frac{1}{n}\sum_{t=1}^{n} \langle w^* - w, x_t \rangle^2} \lesssim \eta\sigma\sqrt{\log(1/\eta)} + \eta^{1/8} R^{1/2} \sigma^{1/2} (\eta\sqrt{\log(1/\eta)})^{1/4} \sqrt[8]{\frac{\log(\min(n,d))}{n}}$$
$$+ \eta^{1/4} R \sqrt[4]{\frac{\log(\min(n,d))}{n}} + \min\left\{\sigma\sqrt{d/n}, (R\sigma)^{1/2}\sqrt[4]{1/n}\right\}$$

with high probability.

Note that all but the first term are $o(1)$ as $n \to \infty$. On the other hand, when $\eta = 0$ only the last term remains and our result simplifies to standard (minimax optimal) guarantees for

Ordinary Least Squares and Ridge regression, see e.g. [Kee10, RH17, SSBD14]. Our result obtains the optimal dependence on η up to the $\sqrt{\log(1/\eta)}$ factor, because the information-theoretic lower bound is $\Omega(\eta\sigma)$:

Proposition 6.1.2. *For any $0 \leq \eta < 1/2$, any algorithm for Huber-contaminated regression with Gaussian noise must incur clean square loss $\frac{1}{n}\sum_{t=1}^{n}\langle w^* - w, x_t\rangle^2$ at least $\Omega(\eta^2\sigma^2)$.*

This follows by embedding the 1-dimensional robust mean estimation problem in a straightforward way — see Example 6.4.3.

We also show the other aspects of the bound (lower bound on n, and the presence of additional "middle terms") are required — see Example 6.5.9. Our results generalize naturally to the setting with heavy-tailed noise, even without second moments, and achieve the optimal dependence on η in those settings too. We defer the detailed statement of these variants to Section 6.5.

Impossibility of strengthening the adversary. Before proceeding to the more sophisticated online settings we consider, we emphasize the impossibility of strengthening the adversary even in the basic model above. First, we consider the version of this problem where the adversary is allowed to corrupt an *arbitrary* η fraction of responses, as opposed to corrupting responses in random locations. In this case, the problem is trivially impossible even in 1-dimension. If $1 - \eta$ fraction of x_i are zero and η fraction are 1, $w^* = \pm R$, and the adversary corrupts an arbitrary η fraction of responses, it's information-theoretically impossible to tell if $w^* = R$ or $w^* = -R$. Thus, we have the following lower bound:

Proposition 6.1.3 (Impossibility with adversarial corruption locations)**.** *In the linear regression model where an adversary corrupts an* arbitrary *η fraction of responses y_t, any algorithm must suffer clean squared loss $\frac{1}{n}\sum_{t=1}^{n}\langle w^* - w, x_t\rangle^2$ at least $\Omega(\eta R^2)$.*

We note variants of this example have already appeared previously in the literature, see e.g. Lemma 6.1 in [KKM18] or Theorem D.1 in [CAT$^+$20]. Similarly, we can consider a strengthened adversary which still corrupts in random locations, but is allowed to change the covariate x_t as well as the response y_t. For essentially the same reason (the adversary can change covariates x_t from 0 to 1 and label them with negated responses $y_t = \mp R$), it again becomes impossible to tell whether $w^* = R$ or $w^* = -R$ and so we have a strong impossibility result:

Proposition 6.1.4 (Impossibility with corrupted covariates)**.** *In the linear regression model where an adversary corrupts an* random *η fraction of covariate and response pairs (x_t, y_t), any algorithm must suffer clean squared loss $\frac{1}{n}\sum_{t=1}^{n}\langle w^* - w, x_t\rangle^2$ at least $\Omega(\eta R^2)$.*

Finally, we consider the "breakdown point" assumption $\eta < 1/2$. (We wrote $\eta < 0.499$ above only to simplify the statement.) If $\eta = 1/2$, a special case of our model is a balanced mixture of linear regressions where half of the responses are generated according to linear model $\langle w_1, x_t\rangle + \xi_t$ and the other half are generated according to a different linear model $\langle w_2, x_t\rangle + \xi_t$. By symmetry, it's impossible to know which of w_1, w_2 is the ground truth linear model, so a clean loss guarantee as in Theorem 6.1.1 is information-theoretically impossible. In fact, in this setting even list recovery, i.e. outputting both w_1 and w_2, is computationally hard [YCS14] and this holds even if $\sigma = 0$.

Online linear regression with Huber contamination. Next, we consider an *online* version of the linear regression model from before. In this case, the algorithm faces two additional complications compared to before:

1. (Online prediction.) The algorithm is forced to output a prediction $\hat{y}_t$ given only x_t and the information from previous rounds $(x_1, y_1), \dots, (x_{t-1}, y_{t-1})$, instead of being able to predict based on all of the data.

2. (Adaptive covariates.) Instead of having the covariates $x_1, \dots, x_T$ fixed in advance, i.e. chosen obliviously, the covariate x_t is chosen *adaptively* by the adversary, based on all information from rounds 1 to $t-1$. In particular, the algorithm's choices may affect the future inputs it receives.

Nevertheless, we are able to give a version of our algorithm which deals with both of these issues. The statement below is for the finite-dimensional setting, but we also give a version of the result with no dependence on d (Theorem 6.7.4), appropriate for the setting of kernel regression. As above, it has an optimal dependence on η up to the log factor. In all online settings, we use T for the total number of rounds/covariates to distinguish from the offline setting where we use n.

Theorem 6.1.5 (Robust online regression, informal version of Theorem 6.7.2)**.** *In the setting of Huber-Contaminated Online Regression (see Definition 6.4.1) with subgaussian noise, $\|x_t\| \leq 1$ for all t and $\|w^*\| \leq R$, for any fixed $\eta < 0.499$, there exists an algorithm which runs in time* $\mathrm{poly}(n, d)$ *and outputs online predictions $\hat{y}_t$ which satisfy the following clean square loss regret bound with high probability:*

$$\mathrm{Reg}_{\mathsf{HSq}}(T) = \sum_{t=1}^{T} (\langle w^*, x_t \rangle - \hat{y}_t)^2 \lesssim \sigma^2 \eta^2 \log(1/\eta) T + \mathrm{poly}(R, \sigma, d, \eta) \cdot o(T).$$

Online contextual bandits with Huber contamination. Finally, by combining our online linear regression result with a recent reduction from the contextual bandits literature ([FR20], see Appendix 6.10), we obtain a result for contextual bandits with adaptive contexts and Huber-contaminated losses/rewards. We note that other reductions can probably be applied in the special case of stochastic contexts, e.g. [SLX20], but for simplicity we only state a result in the more general setting with adaptive contexts. First, we describe the interaction model for each round t:

1. Nature chooses context $z_t = (z_{ta})_{a \in \mathcal{A}}$, possibly adversarially based on the transcript from previous rounds. Here $\mathcal{A}$ with $K \triangleq |\mathcal{A}|$ is the space of possible actions.

2. Learner chooses action a_t from $\mathcal{A}$.

3. A $\mathrm{Ber}(\eta)$ coin γ_t is flipped to decide whether this round is corrupted.

4. If $\gamma_t = 0$, i.e. the round is not corrupted, the learner sees loss $\ell_t^*(a_t) \triangleq \langle z_{ta}, w^* \rangle + \xi_t$ where ξ_t is mean-zero noise.

5. If $\gamma_t = 1$, i.e. the round is corrupted, the learner sees an arbitrary loss $\ell_t(a_t)$ chosen by an adversary based on z_t, a_t, and the transcript from the previous rounds.

In this model, the goal is to minimize the *clean regret*, that is, to compete with the best policy π in hindsight as measured by the *true uncorrupted losses.* We obtain the following guarantee.

Theorem 6.1.6 (Robust contextual bandits, informal version of Theorems 6.8.1 and 6.8.2)**.** *In the setting of Huber-Contaminated Contextual Bandits (see Definition 6.4.4) with σ^2-subgussian noise ξ_t, for any fixed $\eta < 0.499$, there is an algorithm which runs in polynomial time and selects actions a_t which satisfy the following clean regret bound with high probability:*

$$\mathrm{Reg}_{\mathsf{HCB}}(T) = \sup_{\pi} \mathbb{E}\left[\sum_{t=1}^{T}(\ell_t^*(a_t) - \ell_t^*(\pi(z_t)))\right] \lesssim \left(\sigma\eta\sqrt{K\log(1/\eta)}\right)T + \mathrm{poly}(R, K, \eta, \sigma)\cdot o(T)$$

where the supremum ranges over all (non-adaptive) policies π, see Preliminaries.

An impossibility result: failure of convex M-estimators. It may appear surprising that our algorithms for dealing with Huber contamination, even in the simplest linear regression setting, do not use an established approach like Huber regression or L_1/LAD (Least Absolute Deviation) regression — classical approaches which have been studied for decades, and in the case of LAD, even as far back as the 1700s [Bos57]. This is because there are fundamental reasons that *neither* of these approaches can match our strong guarantees in the distribution-free setting. In fact, we prove a lower bound showing the failure of any M-estimator based on a convex loss function:

Theorem 6.1.7 (Lower bound against convex M-estimators, informal version of Theorem 6.9.1)**.** *There is an instance of Huber-contaminated linear regression where the covariates x_t are drawn i.i.d. from a distribution, for which no vector w obtained by minimizing a convex loss with respect to the Huber-contaminated distribution over (x, y)'s can achieve square loss better than $\Omega(\eta^3 R\sigma)$ on the true distribution.*

6.1.2 Roadmap

In Section 6.2, we give an overview of the main techniques in our approach. In Section 6.3, we discuss related work in more detail. In Section 6.4 we record some useful technical facts we use from the literature and state slightly more general versions of the models which we consider. In Section 6.5, we give an alternating minimization algorithm for solving the offline case of Huber-contaminated linear regression. In Section 6.6, we give a sum-of-squares algorithm to handle the case of high contamination rate; combined with the result of the previous section, we obtain Theorem 6.1.1. In Section 6.7, we give a generic recipe for converting our fixed-design guarantees into online ones, thereby proving Theorem 6.1.5. In Section 6.8 we apply the reduction of [FR20] to our regression results to obtain our main result for contextual bandits, Theorem 6.1.6. Lastly, in Section 6.9, we prove our lower bound, Theorem 6.1.7. In Appendix 6.10 we verify that the reduction in [FR20] applies to our Huber-contaminated setting.

6.2 Technical Overview

By a slight modification of the proof of Theorem 5 in [FR20], we can reduce the problem of achieving low clean regret in the contextual bandits setting of Definition 6.4.4 to that of producing an oracle for Hubert-contaminated online regression which gets low clean square loss regret. In this section, we overview the main ingredients for producing such an oracle.

There are two main steps: 1) designing an algorithm for fixed-design Huber-contaminated regression that achieves low square loss, and 2) a generic online-to-offline reduction based on cutting plane methods/online gradient descent.

6.2.1 Huber-Contaminated Fixed-Design Regression

We start with the offline/fixed-design setting, where we are given an arbitrary fixed set of covariates $x_1, \ldots, x_n \in \mathbb{R}^d$ and for the indices t for which y_t was not corrupted, $y_t = \langle w^*, x_t \rangle + \xi_t$ for some independent noise $\xi_t \sim \mathcal{D}$. The exact assumption on the noise is not so important for the argument, since our algorithm is robust to Huber contamination: given an analysis for bounded noise, all the other versions of the results follow more or less by a straightforward truncation argument, treating heavy-tail events as outliers.

Spectrally Regularized Alternating Minimization. Similar to existing approaches in the robust statistics literature, our starting point is to formulate an optimization problem that searches for a regressor w and a "structured" subset $S \subset [n]$ of size $(1 - O(\eta))n$ over which the clean square loss of w is minimized, i.e.

$$w, S = \underset{\substack{w,S: \\ S \text{ large and ``structured''}}}{\operatorname{argmin}} \frac{1}{n}\sum_{t\in S}(y_t - \langle w, x_t\rangle)^2. \tag{6.2}$$

The subset S should satisfy certain structural properties that the set of uncorrupted points $S^* \subseteq [n]$ would collectively satisfy and that can be used to *certify* that the regressor we use is close to w^*. Before we describe how the structural property that we use fundamentally differs from the ones exploited in prior works on robust regression, we first discuss our approach to optimizing the nonconvex objective (6.2). What we do is use a version of a standard heuristic, alternating minimization:

- Given a candidate regressor w, we consider the optimization problem

  $$\min_S \frac{1}{n}\sum_{t\in S}(y_t - \langle w, x_t\rangle)^2.$$

  We relax the set of $(1 - O(\eta))n$-sized "structured" subsets S to the set of $[0,1]$-valued "structured" weights $\{a_t\}_{t\in[n]}$ over the dataset satisfying $\sum_t a_t = 1 - O(\eta)$, and it will be apparent from our definition of "structured" below that this can be formulated as a basic SDP.

- Given a candidate set of weights $\{a_t\}_{t\in[n]}$, we solve the *convex* optimization problem

  $$\min_w \frac{1}{n}\sum_{t\in S} a_t(y_t - \langle w, x_t\rangle)^2.$$

By repeatedly alternating between these two steps, we arrive at an approximate first-order stationary point $(w, \{a_t\})$: more precisely, one for which $\{a_t\}$ is optimal given w and for which

$$\frac{1}{n}\sum_{t\in[n]} a_t(y_t - \langle w, x_t\rangle)\langle x_t, v - w\rangle \le o(1) \tag{6.3}$$

for all v of bounded norm (Lemma 6.5.7). Of course, this stationary point does not have to be a global optimum of the objective function. Nevertheless, our analysis shows that any stationary point of our objective has strong statistical guarantees (Section 6.5.4). To show this, we can decompose the left-hand side of (6.3) for the choice $v = w^*$ into two quantities: 1) the contribution from the uncorrupted points, indexed by some subset $T \subset [n]$, and 2) the contribution from the corrupted ones, indexed by $[n]\backslash T$.

In 1), we can pull out the contribution from the quantity $\frac{1}{n}\sum_{t\in T}\langle x_t, w^* - w\rangle^2$, which corresponds to the clean square loss achieved by the regressor w we have found and turns out to be the dominant term. To upper bound the rest of 1) and 2), the key technical challenge is respectively to control the error incurred from failing to place nonzero weight a_t on some of the points $t \in T$, and from placing nonzero weight a_t on some of the points $t \notin T$. To bound both sources of error, we end up needing to control the quantity

$$\frac{1}{n}\sum_{t\in T}(1 - a_t)\langle x_t, w^* - w\rangle^2. \tag{6.4}$$

The way in which we do so marks the key distinction between our approach and that of previous works on robust regression.

In prior works (see Section 6.3 below), this is the place where one could insist that the weights $\{a_t\}$ are structured in the sense that along every univariate projection, the empirical moments of the dataset reweighted by $\{a_t\}$ are k-hypercontractive for some $k \geq 4$, in which case we could use Holder's to upper bound (6.4). This is not applicable in the general case, where $x_1, \ldots, x_n$ are arbitrary bounded vectors, so a reweighting with hypercontractive empirical moments may not even exist. Instead, our approach is to insist that $\{a_t\}$ must *sub-sample the empirical covariance*, i.e. that

$$\frac{1}{n}\sum_{t\in[n]} a_t x_t x_t^\top \succeq (1-\eta)\frac{1}{n}\sum_{t\in[n]} x_t x_t^\top - o(1)\cdot \mathrm{Id} \tag{6.5}$$

The intuition for this constraint is that because the points that get corrupted in the Huber contamination setting form a *random* subset of the data, the ideal reweighting $\{a^*_t\}$ given by placing uniform mass on the true set of uncorrupted points would satisfy this constraint with high probability by standard matrix concentration. So for any $\{a_t\}$ which sub-samples the empirical covariance, ignoring the low-order term in (6.5), we can thus upper bound the quantity (6.4) by $\eta\sum_{t\in[n]}\langle w^* - w, x_t\rangle^2$. This is negligible compared to the aforementioned dominant term, allowing us to complete the proof that (6.3) suffices to ensure that w incurs low clean square loss.

Optimal breakdown point via Sum of Squares. It turns out that the above approach fails for η larger than $1/3$. Consider a scenario where $1/3$ of the data has been corrupted to come from a different linear model; in this case, there is a spurious local minima in which one takes w in (6.2) to be the linear model generating the corrupted data and S to consist of the corrupted data and a random half of the uncorrupted data (see Remark 6.5.3 for further details).

To circumvent this issue, we appeal to a different algorithm when $1/3 \le \eta < 1/2$. Our starting point is the observation that another way of circumventing the nonconvexity of (6.2) is by considering the natural degree-4 sum-of-squares (SoS) relaxation of (6.2). It turns out that an analysis similar to the one for our alternating minimization algorithm suffices to show that the pseudoexpectation one gets out of solving this relaxation achieves low clean square loss. At a high level, the reason is that one can extract from the former analysis a simple proof in the degree-4 SoS proof system that for w and S satisfying the constraints imposed by the SoS program and optimizing the objective of (6.2), w achieves low clean square loss. The key difference that allows us to circumvent the bad loss landscape of (6.2) when η is large is that the SoS relaxation is guaranteed to produce a lower bound on the original (unrelaxed) problem (6.2), whereas the objective value achieved by an arbitrary stationary point need not.

Other extensions. Using existing generalization bounds [SST10], we give natural and fairly sharp versions of our results for the stochastic/random-design setting. The analysis we outlined works with heavy-tailed noise in L_q for any $q > 1$ and achieves the optimal dependence on η in this setting. If we only use the estimator described above, the sample complexity of our estimator with small confidence parameter δ is not as good with heavy-tailed noise as with subgaussian noise; we show how to improve the sample complexity when $q \ge 2$ by combining our estimator with a simple median-of-means approach from the heavy-tailed regression literature [HS16, M$^+$15].

6.2.2 Online-to-Offline Reduction

We now explain how to use the guarantee of the previous section to get an algorithm for online regression. At a high level, the idea is to use the fixed-design guarantee above to design a *separation oracle* between whatever bad predictor we might be using at a particular time step, and the small ball $\mathcal{B}$ of good predictors w around w^*, any of which would incur sufficiently low regret over any possible sequence of samples. This reduction has a similar spirit to the "halving" algorithm from online learning [SS$^+$11], and efficient variants for halfspace learning based on the ellipsoid algorithm [YJY09, TK08].

Concretely, suppose inductively we have seen samples $(x_1, y_1), \ldots, (x_n, y_n)$ thus far and have used some vector w to predict in the last m steps where we were given $(x_{n-m+1}, y_{n-m+1}), \ldots, (x_n, y_n)$. Let $\mathbf{\Sigma}$ be the average of $x_i x_i^T$ over the last m steps. One of two things could be true.

It could be that in these last m steps, w actually performed well, that is, $\|w - w^*\|_{\mathbf{\Sigma}}^2$ is small, either because $w \in \mathcal{B}$ or because $x_{n-m+1}, \ldots, x_n$ mostly lie in the slab of space where w and w^* yield similar predictions. Either way, because the prediction error under w has been small so far, there is no need to update to a new predictor just yet.

Alternatively, if $\|w - w^*\|_{\mathbf{\Sigma}}^2$ is large, then the gradient of the function $w \mapsto \|w - w^*\|_{\mathbf{\Sigma}}^2$ would give a separating hyperplane between w and $\mathcal{B}$. Of course, the issue with this is that we don't know w^*. To get around this, recall from the fixed-design guarantee that if we ran the alternating minimization algorithm above on the data $(x_{n-m+1}, y_{n-m+1}), \ldots, (x_n, y_n)$ (assuming m is large enough that things concentrate sufficiently well), then the resulting vector $\widetilde{w}$ is close to w^* under $\|\cdot\|_{\mathbf{\Sigma}}$. So to check whether $\|w - w^*\|_{\mathbf{\Sigma}}^2$ is large, by triangle inequality we can simply check whether $\|w - \widetilde{w}\|_{\mathbf{\Sigma}}^2$ is large! If so, the gradient of $w \mapsto \|w - \widetilde{w}\|_{\mathbf{\Sigma}}^2$ gives us a separating hyperplane that we can actually compute.

To summarize, the contrapositive of this tells us that if we don't form a separating hyperplane in a given step, then we know $\|w - w^*\|_{\mathbf{\Sigma}}^2$ is small and we are content to continue using w. Conversely, if we do form a separating hyperplane, we know we won't cut $\mathcal{B}$. This is because every point in $\mathcal{B}$ is, by design, close to w^* under any norm $\|\cdot\|_{\mathbf{\Sigma}}$ defined by the empirical covariance $\mathbf{\Sigma}$ of a sequence of samples.

With these two facts in hand, we can safely run a cutting plane algorithm like ellipsoid or Vaidya's method to update our predictor every time we find a separating hyperplane and ensure that after a bounded number of updates, we find a predictor that will achieve low regret on subsequent steps.

Handling the high-dimensional case. The above approach does not work when the dimension is unbounded, e.g. in kernelized settings, because the guarantees of cutting plane methods are inherently dimension-dependent. We now describe an alternative approach based on wrapping online gradient descent around our guarantee for Huber-contaminated fixed-design regression.

Instead of using Vaidya's algorithm to update the vector w that we predict with whenever the separation oracle returns $\nabla\varphi_t(w)$, we can imagine updating w by simply stepping in the direction of $-\nabla\varphi_t(w)$. The key challenge is to bound the number of times V we get a hyperplane from the separation oracle and have to make such a step, because as long as we don't receive any new hyperplanes, the predictions we make will incur low square loss. For this, we can appeal to the the fundamental regret bound for online gradient descent [Zin03]. Specifically, if we receive a sequence of convex losses $\varphi_1, \dots, \varphi_V$ and play a sequence of inputs $w_1, ..., w_V$ where w_{t+1} is given by taking a gradient step with respect to φ_t from w_t, then the cumulative loss $\sum \varphi_t(w_t)$ incurred only exceeds $\sum \varphi_t(w^*)$ for any single move w^* by an $O(\sqrt{V})$ term (see Theorem 6.7.3). But because the separation oracle is called only when $\varphi_t(w_t) \approx \varphi_t - \varphi_t(w^*)$ is large, this immediately implies that V is bounded.

6.2.3 Lower Bound for Convex Losses

At a high level, the intuition for why convex losses fails is this: in order for the algorithm to be robust to outliers, the loss needs to look roughly like the L_1 loss (e.g. the Huber loss looks like the L_1 loss except in a ball near the origin). However, the L_1 loss $\mathbb{E}[|Y - \langle w, X\rangle|]$ is much less sensitive to making errors for X lying in rare areas of the space than the usual L_2/squared loss $\mathbb{E}[(Y - \langle w, X\rangle)^2]$. In order to take advantage of this, we construct a 1-dimensional example with $1 - \Theta(\eta/R)$ fraction of the covariate distribution a delta mass at $1/R$ and the remainder a delta mass at -1. For simplicity, we take the noise variance $\sigma = 1$. By having the adversary corrupt the response for the much more common portion of the data at $1/R$, the L_1 regression is tricked into making an order ηR error on the rare portion of the data, which causes a squared loss of $\Omega((\eta/R)\eta^2 R^2) = \Omega(\eta^3 R)$. By appropriately generalizing this argument, we rule out the success of all convex losses.

6.3 Related Work

Robust regression, when both the covariates and responses are corrupted As discussed in Section 6.1, our work is closely tied to the long line of recent work on designing efficient algorithms for robust statistics in high dimensions. We refer to [Li18b,Ste18,DK19] for comprehensive surveys of this literature and focus here on the results related to regression [KKM18,BP20,ZJS20,PJL20,DKK$^+$19b,PSB$^+$20,DKS19,CAT$^+$20]. These works are for the stochastic setting where the covariates are drawn i.i.d. from some distribution $\mathcal{D}_x$ but work in a corruption model where the adversary can arbitrarily alter any η fraction of the responses *and* the corresponding covariates. All of these results operate under the assumption that the underlying distribution $\mathcal{D}_x$ is either Gaussian or at least 4-hypercontractive. This is not merely an issue of convenience: in the absence of such assumptions, it is impossible to do anything even in one dimension under this corruption model. We recall the following example from the Results section above:

Example 6.3.1. *Let $d = 1$ and $\varepsilon = 0$, and suppose $w^* = R$. Suppose the distribution over covariates is Ber(η), i.e. it has $1 - \eta$ mass at 0 and η mass at 1. Suppose the adversary corrupts an η fraction of the pairs $(0, 0)$ to be $(1, -R)$. Then it is impossible for the learner to distinguish whether $w^* = R$ or $w^* = -R$.*

We note that variants of this example have already appeared previously in the literature, see e.g. Lemma 6.1 in [KKM18] or Theorem D.1 in [CAT$^+$20]. This does not contradict prior results which make distributional assumptions, because they consider the case where η is small: when $\eta = o(1)$, *Ber*(η) is no longer $O(1)$-hypercontractive as its fourth moment is ηR^4 while the square of its second moment is $\eta^2 R^4$. In summary: when there exist rare features in the data, or when the corruption fraction η is large, it is simply not information-theoretically possible to handle corruption in the covariates.

We also note that the work of [PJL20] shows that, at least in some cases, the covariate corruption can be handled separately from the response corruption by first running a standard filtering method on the covariates, and second running a method robust to response outliers (in their case, Huber regression) on the remaining data. This suggests that handling covariate corruption (when it is possible) and response corruption may be largely orthogonal problems. Finally, one commonality with our work and much of the previous literature is the use of Sum of Squares programming (for us, only needed near the breakdown point 1/2); however, we use a fairly simple degree-4 SoS program, as opposed to prior work (e.g. [KKM18, BP20]) where the SoS degree and sample complexity need to be large in order to take advantage of stronger regularity assumptions.

Robust regression, when just the responses are corrupted A milder corruption model which has received significant attention in the statistics literature is the setting where a fraction, either randomly or adversarily chosen, of the *responses* are corrupted, while the covariates are left intact. One popular approach for regression in this setting is M-estimation [L+17, ZBFL18], originally introduced by Huber [Hub64], in which one minimizes a loss function with suitable robustness properties. Common choices of loss function include the L_1 loss and the Huber loss. In addition to the earlier asymptotic results for this approach [BJK78, Hub73, Pol91], by now numerous works have obtained non-asymptotic guarantees for M-estimation under a variety of models for how the responses are corrupted, but predominantly under the assumption that the design is sub-Gaussian or similarly structured [KP18, DT19, SF20, dNS20]. Notably, in [DT19, SF20] it was shown that in the setting of sparse linear regression with Huber-contaminated responses, M-estimation with (ℓ_1-regularized) Huber loss is nearly minimax-optimal when the noise distribution $\mathcal{D}$ and the covariates are i.i.d. Gaussian.

One exception, and perhaps the result closest in spirit to our results for regression, is that of [Chi20]. One consequence of the results in this work is that in the random-design setting of Definition 6.4.1, that is when the covariates are drawn i.i.d. from some distribution $\mathcal{D}_x$, then if the function class (equivalently, covariate distribution) is hypercontractive in the sense that for any $w \in \mathcal{W}$, $\mathbb{E}_{\mathcal{D}_x}[\langle w - w^*, x\rangle^p]^{2/p} \le \mathbb{E}_{\mathcal{D}_x}[\langle w - w^*, x\rangle^2]$ for some $p > 2$, and if the noise distribution $\mathcal{D}$ satisfies suitable conditions, then M-estimation with Huber loss achieves the information-theoretically optimal error of $\Theta(\sigma^2\eta^2)$ in squared loss. It is also possible to modify their proof to show that the same algorithm would yield the information-theoretically optimal error of $\Theta(\sigma\eta)$ in a different metric, the L_1 *loss*, without the hypercontractivity condition. An L_1 guarantee is much weaker than the usual L_2 (i.e. squared loss) guarantee: for example, it is too weak to give anything interesting for the contextual bandits application.

In fact, as we show in Theorem 6.9.1, M-estimation with Huber loss, and more generally minimization of *any* convex surrogate loss, will not achieve squared loss $\Theta(\sigma^2\eta^2)$ in general when the function class/covariate distribution fails to satisfy this hypercontractivity condition. Instead, we show such estimators must pay squared loss at least $\Omega(\sigma R\eta^3)$. We also mention that to our knowledge, the only work that has explicitly considered *online* regression with corruptions is [PF20], where they considered Gaussian covariates and a random fraction of responses are corrupted by an *oblivious* shift. Additionally, another notable line of work to mention in the literature on regression with contaminated responses stems from using hard thresholding [BJK15, BJKK17, SBRJ19], though these works work also make strong regularity assumptions on the covariates.

Lastly, we mention that in the context of *classification*, there have been a number of recent works giving new algorithmic results for corruption models where the binary labels are corrupted by some process that is halfway between purely stochastic and purely adversarial. For instance, [DGT19, CKMY20, DKTZ20] focus on the *Massart noise model* which can essentially be viewed as a setting where an adversary can only control a random fraction of the labels, but can change them in an arbitrary way. This can be thought of as the classification version of the Huber-contaminated regression problem that we consider in the present work, and the former two results work in the setting without distributional assumptions. We also note that the recent work of [DKK+20] considers the stronger model of *Tsybakov noise* and obtains polynomial-time algorithms under distributional assumptions.

Robustness for bandits There have been a number of notions of robustness proposed in the bandits literature. A classic notion is that of adversarial bandits, a setting where one would like to prove regret bounds even when the rewards are chosen adversarially [ACBFS02]. Many papers have worked to identify ways of interpolating between fully adversarial rewards and stochastically generated ones, including the line of work on "best of both worlds" results [BS12, SS14a, AC16, SL17] as well as an interesting model of bandits with adversarial corruptions introduced by [LMPL18] and subsequently studied by [GKT19]. The latter is a setting of multi-armed bandits where rewards are generated stochastically but then perturbed by an adaptive adversary with a fixed budget of how much he can move the rewards in any given sample path. *We stress that the setting of adversarial bandits is orthogonal to the thrust of the present work, where the goal is to get small clean regret.* For example, while the adversarial nature of the rewards makes the former quite challenging, it is still possible to achieve sublinear regret for adversarial bandits, whereas in our setting, one cannot do better than $\Omega(\eta^2\sigma^2 T)$.

Other notions of robustness that have been considered include the standard notion of misspecification (e.g. [FR20, NO20]) as in Definition 6.4.4, as well as the notion of heavy-tailed reward distributions [BCBL13]. The setting of Huber-contaminated rewards that we study was previously studied in the multi-armed case by [KPK19, ABM19]. [KPK19] also studied Huber-contaminated linear contextual bandits when the contexts are Gaussian or collectively satisfy some RSC-like condition. Even in this distribution-specific setting, their analysis loses a factor of R. A recent work [AGKS21] also studied the Gaussian context case of Huber-contaminated linear contextual bandits and improved over [KPK19]; however their result also suffers from a dependence on R. Lastly, we mention the work of [SS14a,ZS19] who considered a different corruption model for the multi-armed case where the contaminations cannot reduce the "gap," i.e. the difference between the reward of the best arm and that of any other arm, by more than a constant factor in any time step.

6.4 Preliminaries

6.4.1 Formal Description of Models

For technical reasons which will appear naturally in the analysis, it is useful for us to consider the general *misspecified* model where $\varepsilon \geq 0$ is a misspecification parameter that accommodates deviation between the true prediction rule and the best linear model. However, the reader should feel free to consider the usual well-specified setting $\varepsilon = 0$ when reading the results.

Robust Offline Regression. Our analysis in the oblivious setting allows the corruption adversary to depend arbitrarily on the randomness in the problem, as in e.g. [Chi20]. This is different from in the online setting, where it's important that all of the randomness respects the filtration corresponding to time. To be clear, we define the offline model explicitly here.

1. Covariates $x_1, \dots, x_n$ are arbitrary fixed vectors in the unit ball of $\mathbb{R}^d$, i.e. they are chosen obliviously.

2. For every t from 1 to n, a $Ber(\eta)$ coin is flipped to determine if round t is corrupted or not. Let a_t^* be the indicator for whether round t was uncorrupted, i.e. $a_t^* = 1$ when the round is *not* corrupted and this occurs with probability $1 - \eta$.

3. For every uncorrupted round, we observe y_t given by

$$y_t = y_t^* + \xi_t, \qquad y_t^* = \langle w^*, x_t \rangle + \varepsilon_t$$

   where w^* is the true regressor and $\|w^*\| \leq R$, and ξ_t is independently sampled from the noise distribution $\mathcal{D}$ and $|\varepsilon_t| \leq \varepsilon$ is the misspecification. The misspecification ε_t can be chosen in a completely adversarial fashion: formally, it is a random variable depending arbitrarily on all other randomness in the setup (e.g. it can depend arbitrarily on the noise and the coin flips from all rounds).

4. For every corrupted round, y_t is chosen freely by the adversary. Again, we assume nothing about y_t – it can depend arbitrarily on all other randomness in the problem.

Robust Online Regression. We begin by introducing the setup for the online linear regression problem, which is closely related to the linear contextual bandits problem we introduce later. Online regression itself is one of the fundamental problems in online learning that has been extensively studied in the uncontaminated setting, see e.g. [Vov01, AW01, CBL06].

Definition 6.4.1 (Huber-Contaminated Online Regression). *Fix Huber contamination rate $\eta \in (0, 1/2)$, misspecification bound ε, noise distribution $\mathcal{D}$, and unknown weight vector w^*. In each round $t \in [T]$:*

1. *Nature chooses input $x_t \in \mathbb{R}^d$, possibly adversarially based on the transcript from previous rounds.*

2. *Learner chooses prediction $\widehat{y}_t$.*

3. *A Ber(η) coin is flipped to decide whether this round is corrupted.*

4. *If the round is not corrupted, sample ξ_t independently from $\mathcal{D}$. The learner sees $y_t \triangleq y_t^* + \xi_t$, where $y_t^* \triangleq \langle w^*, x_t \rangle + \varepsilon_t$ for some quantity $\varepsilon_t(x_t)$ satisfying $|\varepsilon_t(x_t)| \leq \varepsilon$.*

5. *If the round is corrupted, the learner sees an arbitrary y_t chosen by an adversary based on x_t and the transcript from the previous rounds.*

The goal of the learner, given any x_t in round t (and the transcript from the previous rounds), is to choose a prediction $\widehat{y}_t$ such that with high probability over the choice of Ber(η) coins, and for any (possibly adaptively chosen) sequence of feature vectors $\{x_1, \ldots, x_T\}$ in the above model, the quantity

$$\mathrm{Reg}_{\mathsf{HSq}}(T) = \sum_{t=1}^{T} (\widehat{y}_t - y_t^*)^2. \tag{6.6}$$

is small. We say that A achieves clean square loss regret $\mathrm{Reg}_{\mathsf{HSq}}(T)$. *Note that* $\mathrm{Reg}_{\mathsf{HSq}}$ *is a random variable depending on the randomness of the Ber(η) coins, the randomness of the noise ξ_t, any stochasticity in the choice of the inputs x_t, and the randomness of the learner and adversary. We will establish high-probability bounds on this random variable.*

Remark 6.4.2 (Clean vs Dirty Loss). *It is very important to note that the goal for robust statistics is to minimize the* clean square loss $\sum_{t=1}^{T} (\hat{y}_t - y_t^*)^2$ *and not the "dirty" square loss $\sum_{t=1}^{T} (\hat{y}_t - y_t)^2$ where y_t is potentially corrupted. If our goal was to try to fit the corruptions, as in agnostic learning, then using Ordinary Least Squares would be a good approach for this regression problem.*

On the other hand, there is no importance difference between optimizing the noisy clean square loss $\sum_{t=1}^{T}(\hat{y}_t-(y_t^*+\xi_t))^2$ *and the clean square loss as defined above. Because the noise is by definition independent of* $\hat{y}_t, y_t^*$*, we know that in expectation* $\mathbb{E}[\sum_{t=1}^{T}(\hat{y}_t-(y_t^*+\xi_t))^2] = \mathbb{E}[\sum_{t=1}^{T}(\hat{y}_t-y_t^*)^2]+\sigma^2 T$ *and so the additive term coming from the noise doesn't depend on the prediction sequence* $\hat{y}_t$*.*

Connection to robust mean estimation Note that regression with Huber contaminations is at least as hard as the problem of mean estimation under Huber contaminations, implying that achieving sublinear regret for Huber-contaminated online regression is impossible:

Example 6.4.3. *Let* $d = 1$ *and* $\varepsilon = 0$*, and suppose* $w^* = R$ *and* $\mathcal{D} = \mathcal{N}(0, \sigma^2)$*. Suppose we only ever see* $x_t = 1$*, so that we always have* $y_t^* = R$*. Then each uncorrupted* y_t *is simply an independent draw from* $\mathcal{N}(R, \sigma^2)$*, so the question of producing a good predictor* $\widehat{y}$ *in this special case is* equivalent *to that of estimating the mean of a univariate Gaussian with variance* σ^2 *under the Huber contamination model. It is known that one cannot do this to error better than* $\Omega(\eta\sigma)$ *(see [DKK+18]). More generally, if we only assume* $\mathcal{D}$ *has hypercontractive moments up to degree* k*, one can devise distributions* $\mathcal{D}$ *for which one cannot do better than error* $\Omega(\eta^{1-1/k}\sigma)$ *(see e.g. Fact 2 from [HL19]).*

Robust Contextual Bandits. We study the following robust version of contextual bandits, first introduced in [KPK19]. We first state the general form of the contextual bandits model (for an abstract regression function f), then specialize to the linear case.

Definition 6.4.4 (Huber-Contaminated Contextual Bandits)**.** *Let* $\mathcal{Z}$ *be an arbitrary state space, and let* $\mathcal{A}$ *be an action space of size* K*. Fix Huber contamination rate* $\eta \in (0, 1/2)$*, misspecification rate* ε*, and unknown function* $f : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$*. Ahead of time, an oblivious adversary chooses distributions* $\Pr_{\ell_t^*}[\cdot|z_t]$ *over loss functions* $\ell_t^* : \mathcal{A} \to [0, R]$ *for all possible contexts* z_t *and all time steps* $t \in [T]$*. We assume the conditional means of the loss distributions are* realized up to misspecification ε *by* f*, i.e. for all* t, z, a*,*

$$\mathop{\mathbb{E}}_{\ell_t^*}[\ell_t^*(a)|z_t = z] = f(z, a) + \varepsilon_t(z, a), \qquad |\varepsilon_t(z, a)| \le \varepsilon. \tag{6.7}$$

Let ξ_t be the random variable which, conditioned on $z_t = z$, takes on the value

$$\xi_t \triangleq \ell_t^*(a) - f(z,a) - \varepsilon_t(z,a),$$

and define noise parameter σ *by* $\sigma^2 \triangleq \sup_{z,t} \mathbb{E}[\xi_t^2 | z_t = z]$. *In each round $t \in [T]$:*

1. *Nature chooses z_t, possibly adversarially based on the transcript from previous rounds.*
2. *Learner chooses action $a_t \in \mathcal{A}$.*
3. *A Ber(η) coin γ_t is flipped to decide whether this round is corrupted.*
4. *If $\gamma_t = 0$, i.e. the round is not corrupted, the learner sees loss $\ell_t^*(a_t)$, where ℓ_t^* is drawn independently from the distribution* $\Pr_{\ell_t^*}[\cdot|z_t]$.
5. *If $\gamma_t = 1$, i.e. the round is corrupted, the learner sees an arbitrary loss $\ell_t(a_t)$ chosen by an adversary based on z_t, a_t, and the transcript from the previous rounds.*

The goal of the learner in the adversarial setting is to compete with the best policy in hindsight as measured by the clean losses ℓ_t^* *incurred in every round, that is to select a sequence of actions $a_1, \ldots, a_T$ for which*

$$\widetilde{\mathrm{Reg}}_{\mathsf{HCB}}(T) = \sup_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \left(\ell_t^*(a_t) - \ell_t^*(\pi(z_t))\right)\right], \tag{6.8}$$

is small, where the supremum ranges over all (non-adaptive) policies $\pi : \mathcal{X} \to \mathcal{A}$ and the expectation is over the randomness of the Ber(η) coins, the randomness of the rewards, any stochasticity in the choice of contexts, and the randomness of the learner. We say that such a learner achieves clean pseudo-regret $\widetilde{\mathrm{Reg}}_{\mathsf{HCB}}(T)$.

In the special case where $\varepsilon = 0$, we will consider the quantity

$$\mathrm{Reg}_{\mathsf{HCB}}(T) = \sum_{t=1}^{T} \left(\ell_t^*(a_t) - \ell_t^*(\pi^*(z_t))\right)$$

where $\pi^*(z) \triangleq \arg\max_a f(z,a)$. *Note that this is a random variable in the same things defining the expectation in (6.8). We say that a learner achieves* clean regret $\mathrm{Reg}_{\mathsf{HCB}}(T)$.

We will establish high-probability bounds on $\mathrm{Reg}_{\mathsf{HCB}}$.

Definition 6.4.5 (Huber-Contaminated Linear Contextual Bandits)**.** *This is the special case of Definition 6.4.4 where the regression function* $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ *is linear in the following sense. The context space* $\mathcal{X}$ *is a Hilbert space and each context is of the form* $z_t = (z_{t1}, \ldots, z_{tK})$, *i.e. there is a separate context vector for each arm. Then we assume that*

$$f(z, a) = \langle z_{ta}, w^* \rangle$$

for some vector $w^* \in \mathbb{R}^d$.

Without adversarial corruptions this is the familiar linear contextual bandits problem, which has a wide range of applications precisely because in many settings the context is an important component of the prediction task. For example, in online advertising the choice of which ad to display ought to depend on information about the webpage that the ad will be displayed on as well as any information we have about the user we are displaying it to, which can be encoded as a high-dimensional vector. In healthcare, when we want to choose between various treatment options again we want to adapt to the relevant context such as the patient history. For additional applications, see the survey [BR19].

However in many of these settings it is natural to imagine that some of the feedback we receive departs in arbitrary ways from the model. This could happen in online advertising due to clickfraud, particularly when malware takes over a user's account. It could happen in healthcare in the context of drug trials, particularly ones that measure some real valued variable, when there are testing errors or confounding variables that are difficult to model. For all these and many more reasons it is natural to wonder if there could be algorithms for contextual bandits with stronger robustness guarantees.

Remark 6.4.6. *We note that in some papers on contextual bandits, the range of the loss functions is normalized to* $[0, 1]$ *for convenience. The scale-invariant quantity which we want to avoid dependence on is the ratio* R/σ.

Remark 6.4.7. *As we will rely on a formal connection between contextual bandits and online regression illuminated in [FR20], it will be helpful to situate our definitions in their context. In particular, when* $\eta = 0$, *Definition 6.4.4 specializes to Assumption 4 of [FR20], and an* *algorithm for Definition 6.4.1 achieving clean square loss regret at most* $\mathrm{Reg}_{\mathsf{HSq}}(T)$ *would satisfy Assumption 2b of [FR20] in the realizable case with* ε*-misspecification.*

Model Assumptions. We adopt the following standard normalization convention for the covariates and weight vector.

Assumption 5. *In the regression setting (Definition 6.4.1), for any round* t*,* $\|x_t\| \leq 1$ *almost surely,* $\|w^*\| \leq R$*. Correspondingly, in the contextual bandits setting (Definition 6.4.5) we assume* $\|z_{ta}\| \leq 1$ *for all* a *and* $\|w^*\| \leq R$*.*

To simplify the statement of bounds we assume in all statements that $\varepsilon, \sigma = O(R)$. The last assumption can be removed at the cost of longer Theorem statements (e.g. writing $R+\sigma$ instead of R); this scaling captures the interesting setting for the bounds, because if $\varepsilon \gg R$ then the responses are arbitrary, and if $\sigma \gg R$ then no interesting robustness guarantee is possible, as explained earlier — the trivial guarantee of Ordinary Least Squares in this setting is already close to optimal.

We now formally describe the (weak) assumptions on the noise under which we can perform our analysis.

Definition 6.4.8 (Weak L_q Space)**.** *Suppose* X *is a real-valued random variable and* $q \geq 1$*. We define the* weak L_q *or* $L_{q,\infty}$ quasinorm *of* ξ *to be*

$$\|X\|_{q,\infty} \triangleq \sup_{\lambda > 0} \lambda \cdot \left|\Pr[|X| > \lambda]^{1/q}\right|$$

so that $\Pr[|X| > \lambda] \leq \|X\|_{q,\infty}^q / \lambda^q$*. When* $q = \infty$*, we define* $\|X\|_{\infty,\infty} = \inf\{\lambda > 0 : \Pr[|X| \geq \lambda] = 0\}$ *to be the same as the* L_∞ *norm. We say that* X *is in weak* L_q *or* $L_{q,\infty}$ *space if* $\|X\|_{q,\infty} < \infty$*.*

From Markov's inequality, one has that $\Pr[|X| > \lambda] \leq \mathbb{E}[|X|^q]/\lambda^q$ which shows that $\|X\|_{q,\infty} \leq \|X\|_q$.

Assumption 2. *We assume the noise* $\xi \sim \mathcal{D}$ *is mean zero and that for some* $q > 1$*,*

$$\sigma_q \triangleq \|\xi\|_{q,\infty} < \infty.$$

6.4.2 Technical Preliminaries

Here we collect miscellaneous technical facts that will be useful in later sections. Throughout this paper we use standard notation for inequalities up to constants; for example, $a \lesssim b$ and $a = O(b)$ both denote an inequality true up to an absolute constant, and occasionally we use $C > 0$ to denote a universal constant which can change from line to line. Given a matrix M, we let $\|M\|$ denote the operator norm of M. Given a positive semidefinite matrix $\mathbf{\Sigma}$, we define the *Mahalanobis* norm by

$$\|x\|_{\mathbf{\Sigma}}^2 := \|\mathbf{\Sigma}^{1/2}x\|^2 = \langle x, \mathbf{\Sigma}x\rangle.$$

Truncation Lemma. In our algorithm and analysis, we handle heavy-tailed noise using a truncation argument; this somewhat parallels the use of truncation arguments in large deviation theory, see e.g. [FN71]. The following Lemma shows that random variables with tail bounds behave reasonably under truncation, in the sense that their means do not move drastically.

Lemma 6.4.9. *Suppose that X is a mean-zero random variable and $\sigma_q \triangleq \|X\|_{q,\infty} < \infty$. Then for any $s > 0$,*

$$|\mathbb{E}[X\,\mathbb{1}[|X| < s]]| \le \frac{q}{q-1} \cdot \frac{\sigma_q^q}{s^{k-1}}.$$

Proof. We know

$$0 = \mathbb{E}[X] = \mathbb{E}[X\,\mathbb{1}[|X| < s]] + \mathbb{E}[X\,\mathbb{1}[|X| \ge s]]$$

so using the identity $\mathbb{E}[Z] = \int_0^\infty \Pr[Z > y]dy$ for nonnegative random variable Z (Lemma 1.2.1 of [Ver18]), we have

$$\begin{aligned}
|\mathbb{E}[X\,\mathbb{1}[|X| < s]]| = |\mathbb{E}[X\,\mathbb{1}[|X| \ge s]]| &\le \mathbb{E}[|X|\,\mathbb{1}[|X| \ge s)]] \\
&= \int_0^\infty \Pr[|X|\,\mathbb{1}[|X| \ge s] > y]dy \\
&= s\Pr[|X| \ge s] + \int_s^\infty \Pr[|X| > y]dy \\
&\le \frac{\sigma_q^q}{s^{q-1}} + \int_s^\infty \frac{\sigma_q^q}{y^k}dy
\end{aligned}$$

$$= \sigma_q^q \left(\frac{1}{s^{q-1}} + \frac{1}{(q-1)s^{q-1}} \right) = \frac{q}{q-1} \cdot \frac{\sigma_q^q}{s^{k-1}}$$

where in the last inequality, we used the definition of $L_{q,\infty}$. □

6.5 Alternating Minimization for Offline Regression

In this section, we prove our main results for regression in the usual offline setting. After giving some setup and stating the main offline result in Section 6.5.1, in Section 6.5.2 we give a full description of our alternating minimization-based algorithm. In Section 6.5.3 we show that it converges to an approximate stationary point. In Section 6.5.4 we show that this suffices to obtain our claimed error guarantees, and also give improved rates in the case of subgaussian noise. In Section 6.5.5 we show how our fixed-design guarantee can yield strong results in the stochastic setting often considered in statistical learning. Finally, in Section 6.5.6, we give improved rates when the noise is in L_q for $q \geq 2$ by boosting via a high-dimensional median.

6.5.1 Setup and Main Result

We will state and prove results for two closely related settings: (1) the usual setting in linear regression where the covariates x_t are fixed arbitrary vectors (i.e. chosen obliviously), and (2) the model which is relevant for our online applications, where the covariates x_t are generated sequentially and adaptively, so they can depend on e.g. the realization of the noise in previous rounds. The second setting is the proper offline version of the Huber-Contaminated Online Regression Problem as defined in Definition 6.4.1.

We briefly recall some of the relevant notation. Let a_t^* be the indicator for whether round t was uncorrupted, i.e. $a_t^* = 1$ when the round is *not* corrupted and this occurs with probability $1 - \eta$. Recall from (6.6) that for every $t \in [n]$ corresponding to a round which is not corrupted, we observe y_t given by

$$y_t = y_t^* + \xi_t, \qquad y_t^* = \langle w^*, x_t \rangle + \varepsilon_t$$

where w^* is the true regressor and $\|w^*\| \le R$, and ξ_t is independently sampled from the noise distribution $\mathcal{D}$, and $|\varepsilon_t| \le \varepsilon$ is the misspecification. On the other hand, on corrupted rounds y_t is chosen freely by the adversary. For convenience, define

$$\mathbf{\Sigma}_n \triangleq \frac{1}{n}\sum_{t=1}^{n} x_t x_t^\top$$

Let u^* be the best norm R linear predictor of the uncorrupted and unnoised data, that is,

$$u^* \triangleq \arg\min_{u:\|u\|\le R} \frac{1}{n}\sum_t (y_t^* - \langle u, x_t\rangle)^2 \tag{6.9}$$

and let $\delta_t \triangleq y_t^* - \langle u^*, x_t\rangle$. By definition of u^*, we have that

$$\frac{1}{n}\sum_t \delta_t^2 \le \frac{1}{n}\sum_t \varepsilon_t^2 \le \varepsilon^2 \tag{6.10}$$

almost surely; in fact, the conclusion of (6.10) is all we need about the misspecification model and w^*, ε_t play no further role in this section.

Our goal will be to output $\widehat{w}$ such that the MSE (Mean Squared Error) with respect to the true responses is as small as possible; since u^* is the optimal linear predictor, this is the same (by the Pythagorean Theorem) as asking for $\|\widehat{w} - u^*\|_{\mathbf{\Sigma}_n}^2$ is small. When there is no misspecification, this is equivalent to recovering w^* up to small error in $\mathbf{\Sigma}_n$ norm. When there is misspecification, it is easy to see that if $\|\widehat{w} - u^*\|$ is small, then $\|\widehat{w} - w^*\|_{\mathbf{\Sigma}_n}$ is also small, up to an extra $O(\varepsilon)$ term from the triangle inequality. The algorithm achieving our goal is SCRAM (**S**pe**C**trally **R**egularized **A**lternating **M**inimization, defined in Algorithm 15 and analyzed in Theorem 6.5.1).

In the following Theorem, the constants in the guarantee must deteriorate slightly as we approach the breakdown point $\eta = 1/3$ of this estimator, so we introduce a parameter β which tracks the distance to 1/3; as long as we are strictly bounded away from this point, β is a $\Theta(1)$ quantity and can be ignored. As explained in Remark 6.5.3, this breakdown point is optimal for SCRAM, but in Section 6.6 we will give a more powerful version of this estimator based on sum-of-squares programming which achieves optimal breakdown point

1/2.

Theorem 6.5.1. *Suppose that $\eta < 1/3$ is an upper bound on the contamination level, define*

$$\beta \triangleq (1/3 - \eta)^2 \tag{6.11}$$

and suppose for some $q \in (1, \infty], \sigma_q \geq 0$ and all t that

$$\|\xi_t\|_{q,\infty} \leq \sigma_q$$

in the sense of Assumption 2. Then provided

$$\eta \cdot n \gtrsim \log(\min(n, d)/\delta),$$

we can take $\alpha = \Theta\left(\sqrt{\frac{\eta \log(d/\delta)}{n}}\right)$ and $\overline{\eta} = \eta + \Theta(\eta\sqrt{\beta})$ such that the output w of SCRAM *with poly$(R/\sigma, \log(2/\delta), d, n)$ many steps satisfies for oblivious covariates the bound*

$$\begin{aligned}\beta^{1+1/q}\|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim{}& \frac{q}{q-1}\eta^{1-1/q}\sigma_q + \eta^{1/2}\varepsilon + \eta^{1/8}R^{1/2}(\varepsilon + \frac{q}{q-1}\eta^{1/2-1/q}\sigma_q)^{1/2}\sqrt[8]{\frac{\log(\min(n,d)/\delta)}{n}} \\ &+ \eta^{1/4}R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}} + \eta^{-1/q}\min\left\{\sigma\sqrt{\frac{d + \log(2/\delta)}{n}}, (R\sigma)^{1/2}\sqrt[4]{\frac{\log(2/\delta)}{n}}\right\}\end{aligned}$$

with probability at least $1 - \delta$. In the more general case of adaptive covariates, it satisfies the bound

$$\begin{aligned}\beta^{1+1/q}\|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim{}& \frac{q}{q-1}\eta^{1-1/q}\sigma_q + \eta^{1/2}\varepsilon + \eta^{1/8}R^{1/2}(\varepsilon + \frac{q}{q-1}\eta^{1/2-1/q}\sigma_q)^{1/2}\sqrt[8]{\frac{\log(\min(n,d)/\delta)}{n}} \\ &+ \eta^{1/4}R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}} + \eta^{-1/q}(R\sigma)^{1/2}\sqrt[4]{\frac{\log(2/\delta)}{n}}\end{aligned}$$

i.e. the same bound except the last term was changed.

Remark 6.5.2 (Oracle Inequality Interpretation)**.** *As mentioned before, the only guarantee on the misspecification we need is (6.10). This means that for any $\varepsilon^2 \geq \frac{1}{n}\sum_t \delta_t^2$, i.e. any*

$\varepsilon > 0$ such that (6.10) is true almost surely, we have

$$\frac{1}{n}\sum_t (y_t^* - \langle \hat{w}, x_t \rangle)^2 \lesssim \varepsilon^2 + \|u^* - \hat{w}\|_{\mathbf{\Sigma}_n}^2$$

which combined with Theorem 6.5.1 makes formal that $\langle \hat{w}, x_t \rangle$ is the best linear model of y_t^* up to a small error term. This kind of bound for an estimator in the presence of misspecification is known as an oracle inequality *[Tsy08], since $\hat{w}$ competes with the oracle fit u^*.*

Remark 6.5.3 (Breakdown point and landscape)**.** *The breakdown point of $\eta = 1/3$ is optimal for this estimator based on local search. This breakdown point is optimal even if $X \sim N(0, I)$ and the true generative model is a noiseless mixture of two linear regressions $w_1 \neq w_2$ with corresponding weights $1/3, 2/3$, so we view w_1 as contamination. In this setting $\sigma_q = 0$ so an estimator achieving the optimal $O(\sigma_q)$ rate gets error $o(1)$. However, the pair (w_1, a_1) is a bad local minimum where the weight vector a_1 keeps all of the data points from w_1 and keeps each point labeled by w_2 with probability $1/2$. In Section 6.6 we show how to overcome the bad landscape for $\eta \in [1/3, 1/2)$, achieving the optimal $O(\sigma_q)$ error guarantee, using more powerful optimization tools (the Sum of Squares hierarchy) and a new analysis.*

Remark 6.5.4 (Small η regime)**.** *If the true contamination level is very small, e.g. $\eta = 0$, then applying Theorem 6.5.1 with a larger value of η will optimize the upper bound.*

When the noise is L_q for $q \geq 2$, we show how to improve the last term on the right-hand side of Theorem 6.5.1 to avoid an $\eta^{-1/q}$ dependence in the last term on the right-hand side, see Theorem 6.5.18.

6.5.2 Algorithm Specification

The algorithm used in Theorem 6.5.1 is based upon finding first-order stationary points of the following nonconvex problem.

Program 1. *Define variables $w, a_1, \ldots, a_n$ and consider the optimization problem with pa-*

rameters $\overline{\eta}, \alpha, R \geq 0$ given by

$$\begin{aligned}
\min_{w} \ \mathop{min}_{a_1,\ldots,a_n} \quad & \frac{1}{n}\sum_{t=1}^{n} a_t(y_t - \langle w, x_t\rangle)^2 \\
s.t. \quad & 0 \leq a_t \leq 1 \quad \forall t \in [n] \\
& \sum_t a_t \geq (1 - \overline{\eta} - \alpha)n \\
& \frac{1}{n}\sum_t (1 - a_t) x_t x_t^\top \preceq \overline{\eta}\mathbf{\Sigma}_n + \alpha \cdot Id \\
& \|w\| \leq R
\end{aligned}$$

where $\|w\|$ denotes the Euclidean norm of w.

The overall objective

$$L(w, a) := \frac{1}{n}\sum_t a_t(y_t - \langle w, x_t\rangle)^2$$

is *biconvex*, i.e. convex individually in the variables a and the variables w, but not jointly convex. Since it is a nonconvex problem, we cannot guarantee to find the true global minimum of this optimization problem. One of the most common heuristics for biconvex problems is to perform alternating minimization, which will output an approximate first order stationary point. Fortunately, we prove in our setting that this suffices and all approximate first order stationary points satisfy the desired statistical guarantee. As one half of the alternating minimization procedure, we observe that minimizing a for fixed w is a simple SDP (semidefinite program):

Program 2. *For fixed vector w, define variables $a_1, \ldots, a_n$ and define the optimization problem $\boldsymbol{SDP}_w$ with additional parameters $\overline{\eta}, \alpha \geq 0$ given by*

$$\begin{aligned}
\mathop{min}_{a_1,\ldots,a_n} \quad & \frac{1}{n}\sum_{t=1}^{n} a_t(y_t - \langle w, x_t\rangle)^2 \\
s.t. \quad & 0 \leq a_t \leq 1 \quad \forall t \in [n] \\
& \sum_t a_t \geq (1 - \overline{\eta} - \alpha)n \\
& \frac{1}{n}\sum_t (1 - a_t) x_t x_t^\top \preceq \overline{\eta}\mathbf{\Sigma}_n + \alpha \cdot Id.
\end{aligned}$$

Note that this corresponds to Program 1 for a fixed choice of w.

```
Algorithm 15: SCRAM($D, \varepsilon_{\mathsf{OPT}}$)
  Input: Dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
  Output: Approximate first-order critical point of Program 1 (see Lemma 6.5.7)
  Let $w^{(1)} = 0$.
  for $s = 1$ to $\infty$ do
      Let $a^{(s)}$ be the minimizer of Program 2 with $w = w^{(s)}$.
      Let $w^{(s+1)}$ be the minimizer of $L(w, a^{(s)}) = \sum_t a_t^{(s)}(y_t - \langle w, x_t \rangle)^2$ over all $w$ with
        $\|w\| \le R$.
      if $L(w^{(s+1)}, a^{(s)}) > L(w^{(s)}, a^{(s)}) + \varepsilon_{\mathsf{OPT}}$ then
          return $w^{(s)}, a^{(s)}$.
```

6.5.3 Optimization Analysis

For the analysis we need the following simple Taylor expansion inequality used to analyze gradient descent on smooth functions:

Lemma 6.5.5 (Standard, see e.g. [Bub14]). *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth in the sense that $\|\nabla^2 f\|_{OP} \le 2L$. Then*

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + L\|y - x\|^2.$$

From this we get the following Descent Lemma on the ball:

Lemma 6.5.6. *Suppose that f is L-smooth and x, y are vectors in $\mathbb{R}^d$ such that $\|x\|, \|y\| \le R$ and $\langle \nabla f(x), x - y \rangle \ge \Delta > 0$. Then there exists a point z which is a convex combination of x, y such that*

$$f(z) \le f(x) - \frac{\Delta^2}{16LR^2}.$$

Proof. We consider points of the form $z_\lambda := (1 - \lambda)x + \lambda y$ which by convexity lie in the radius R ball. Observe by Lemma 6.5.5 that

$$f(z_\lambda) \le f(x) - \lambda\Delta + 4LR^2\lambda^2$$

since $\|x - x_\lambda\| \le \lambda\|x\| + \lambda\|y\| \le 2\lambda R$. The upper bound is optimized by $\lambda = \frac{\Delta}{8LR^2}$ and plugging in gives the result. □

Lemma 6.5.7. SCRAM *with* $\varepsilon_{\mathsf{OPT}} = \varepsilon_{grad}^2/4R^2$ *outputs vector* w *and weights* $a_1, \ldots, a_n$ *satisfying the constraints of Program 1 such that:*

1. *(Partial optimality) The variables* a *are global minimizers of* $\boldsymbol{SDP}_w$ *(Program 2).*

2. *(First order stationarity)*

$$\frac{1}{n}\sum_t a_t(y_t - \langle w, x_t\rangle)\langle x_t, v - w\rangle \leq \varepsilon_{grad} \tag{6.12}$$

for all v *with* $\|v\| \leq R$.

Furthermore, the expected number of iterations in the main loop is at most $O((R^2+\sigma^2)R^2/\varepsilon_{grad}^2)$.

Proof. By definition $a^{(s)}$ is the minimizer of the $\mathbf{SDP}_{w^{(s)}}$ so the first property is satisfied by construction. We now prove the second property. Observe that the objective $L(w, a^{(s)})$ is 1-smooth in w and

$$\nabla_w L(w, a^{(s)}) = -\frac{2}{n}\sum_t (y_t - \langle w, x_t\rangle)x_t. \tag{6.13}$$

Therefore by Lemma 6.5.6 and the fact that $w^{(s+1)}$ is the optimizer for fixed $a^{(s)}$, we know that if there exists v with $\langle \nabla_w L(w^{(s)}, a^{(s)}), w^{(s)} - v\rangle \geq \Delta > 0$

$$L(w^{(s+1)}, a^{(s)}) \leq L(w^{(s)}, a^{(s)}) - \frac{\Delta^2}{16R^2}.$$

By the contrapositive, if the decrease in objective value when moving from $w^{(s)}$ to $w^{(s+1)}$ is less than $\varepsilon_{\mathsf{OPT}}$, then it implies that

$$\langle \nabla_w L(w^{(s)}, a^{(s)}), w^{(s)} - v\rangle \leq 4R\sqrt{\varepsilon_{\mathsf{OPT}}}$$

for all v in the unit ball. Hence by (6.13) taking $\varepsilon_{\mathsf{OPT}} = \varepsilon_{grad}^2/4R^2$ gives the stated guarantee.

Finally, we bound the number of iterations needed. Every time the loop is repeated, the objective value $L(w, a)$ decreases by at least $\varepsilon_{\mathsf{OPT}}$ and clearly $L(w, a) \geq 0$. Therefore the total number of iterations can be upper bounded by $L(0, a^{(1)})/\varepsilon_{\mathsf{OPT}}$. By considering the (possibly suboptimal solution) $a_t = a_t^*$ to the first SDP, we see that the expected value of $L(0, a^{(1)})$ is at most $R^2 + \sigma^2$. Therefore the expected total number of iterations is at most $(R^2 + \sigma^2)/\varepsilon_{\mathsf{OPT}}$. □

6.5.4 All Stationary Points are Good

It remains to show why condition (6.12) implies the desired error guarantee. To establish the general guarantee of Theorem 6.5.1, it's sufficient to reduce to the case where the noise ξ_t is bounded, unless we care about the precise sample complexity. For this reason, we start with this setting (Section 6.5.4), show how to reduce the $L_{q,\infty}$ setting of Theorem 6.5.1 to the bounded case, and then discuss how to tailor the analysis to get refined guarantees for subgaussian noise in Section 6.5.4. Later in Section 6.5.6, we give an improved version of Theorem 6.5.1 when the noise $\{\xi_t\}$ is L_q for $q \geq 2$, see Theorem 6.5.18.

Bounded Noise Analysis

In the bounded case we establish the following result:

Theorem 6.5.8 (SCRAM Guarantee with Bounded Noise). *Suppose that $\eta < 1/3$, define β as in (6.11), and suppose for some $\sigma \geq 0$ that for all t,*

$$|\xi_t| \leq \sigma \tag{6.14}$$

almost surely. Then if $\eta = 0$ or

$$n \gtrsim \log(\min(n, d)/\delta)/\eta, \tag{6.15}$$

taking $\alpha = \Theta\left(\sqrt{\frac{\eta \log(\min(n,d)/\delta)}{n}}\right)$ and $\overline{\eta} = \eta$, the output w of SCRAM *with poly$(R/\sigma, \log(2/\delta), d, n)$ many steps satisfies for oblivious covariates the bound*

$$\beta \|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim \eta\sigma + \eta^{1/2}\varepsilon + \eta^{1/8} R^{1/2} (\eta^{1/2}\sigma + \varepsilon)^{1/2} \sqrt[8]{\frac{\log(\min(n,d)/\delta)}{n}} + \eta^{1/4} R \sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}}$$
$$+ \min\left\{\sigma\sqrt{\frac{d + \log(2/\delta)}{n}}, (R\sigma)^{1/2} \sqrt[4]{\frac{\log(2/\delta)}{n}}\right\}$$

with probability at least $1 - \delta$. In the more general case of adaptive covariates, it satisfies the bound

$$\beta \|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim \eta\sigma + \eta^{1/2}\varepsilon + \eta^{1/8} R^{1/2} (\eta^{1/2}\sigma + \varepsilon)^{1/2} \sqrt[8]{\frac{\log(\min(n,d)/\delta)}{n}} + \eta^{1/4} R \sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}}$$
$$+ (R\sigma)^{1/2} \sqrt[4]{\frac{\log(2/\delta)}{n}},$$

i.e. the same bound except the second line was changed.

Example 6.5.9 (Lower bound when $\sigma = \varepsilon = 0$). *Consider the special case with $\sigma = \varepsilon = 0$ with oblivious contexts. Observe that when $\sigma = \varepsilon = 0$ the only nonzero term in the upper bound is $\eta^{1/4} R \sqrt[4]{\frac{\log(n/\delta)}{n}}$. Now consider the setting where the clean regression model with $d = n$ is given by $Y^* = w^* \in \mathbb{R}^n$, and we consider the η-contaminated version of this model with $\delta = 1/n$ and $\eta = \log(n/\delta)/n,$. The number of contaminated coordinates of Y will be close to $\eta n = \Theta(\log(n/\delta))$, and for each of those coordinates i, the algorithm observes no information about w^*_i. Considering letting $w^* = \pm R e_j$ for an arbitrary $j \in [n]$, then the probability coordinate j is missed is $\Theta(\eta) = \Theta(\log(n/\delta)/n) = \omega(\delta)$ and on this event the algorithm must pay a cost in squared loss $\|u^* - w\|^2_{\mathbf{\Sigma}_n}$ of $R^2/n = \frac{1}{\log(n/\delta)} R^2 \eta^{1/2} \sqrt{\frac{\log(n/\delta)}{n}}$, matching the upper bound up to the log factor.*

This example also shows the necessity of (6.15) when $\eta \neq 0$: without this lower bound, we could take $\eta = 1/n^{1+\gamma}$ for some $\gamma > 0$, $\delta = 0.1/n^{1+\gamma}$ and we would conclude by the same argument that $R^2/n \lesssim R^2\eta^{1/2}\sqrt{\log(n/\delta)/n} = \Theta(R^2\sqrt{\log(n)}/n^{1+\gamma/2})$ which is false.

Given this result, Theorem 6.5.1 follows by slightly increasing the value of η, so that heavy tail events are counted as contamination; we have to be slightly careful when the noise is asymmetric, because truncating can also induce also a small amount of misspecification, but it does not affect the final bound.

Proof of Theorem 6.5.1. We prove this Theorem by reducing to Theorem 6.5.8. We consider the effect of treating all clean responses with $|\xi| \geq M\sigma_q$ for some $M \geq 1$ as contamination, increasing the effective η to $\overline{\eta} = \eta + \sqrt{\beta}\eta/2$ and making the noise bounded. Recall from the definition that

$$\Pr[|\xi| \geq M\sigma_q] \leq \frac{1}{M^q}$$

so by solving $\beta^{1/2}\eta/2 \geq 1/M^q$ we find that setting

$$M = (\beta^{1/2}\eta)^{-1/q}$$

ensures the total contamination level is at most $\eta + \sqrt{\beta}\eta/2 = \overline{\eta}$ as desired. Applying Lemma 6.4.9 shows that this reduction this causes an additional misspecification cost of

$$\frac{q\sigma_q}{(q-1)M^{q-1}} = \frac{q}{q-1}\sigma(\beta^{1/2}\eta)^{1-1/q}.$$

Now plugging into the conclusion of Theorem 6.5.8 with $\sigma_\infty = M\sigma$, $\overline{\eta}$, and $\varepsilon' = \varepsilon + \Theta(\frac{q}{q-1}\sigma(\beta^{1/2}\eta)^{1-1/q})$ gives, as long as

$$\eta \cdot n \gtrsim \log(\min(n,d)/\delta)$$

a bound of the form

$$\begin{aligned}\beta\|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim \eta M\sigma + \eta^{1/2}\varepsilon' + \eta^{1/8}R^{1/2}(\eta^{1/2}\sigma M + \varepsilon')^{1/2}\sqrt[8]{\frac{\log(\min(n,d)/\delta)}{n}} + \eta^{1/4}R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}} \\ + M\min\left\{\sigma\sqrt{\frac{d+\log(2/\delta)}{n}}, (R\sigma)^{1/2}\sqrt[4]{\frac{\log(2/\delta)}{n}}\right\}\end{aligned}$$

where the first term is bounded as

$$\eta M\sigma \lesssim \beta^{-1/2q}\eta^{1-1/q}\sigma_q$$

the second term is bounded as

$$\eta^{1/2}\varepsilon' \lesssim \eta^{1/2}\varepsilon + \frac{q}{q-1}\beta^{-1/2q}\eta^{3/2-1/q}\sigma_q$$

and the third term is bounded by observing

$$\eta^{1/2}\sigma M + \varepsilon' \lesssim \beta^{-1/2q}\eta^{1/2-1/q}\sigma_q + \varepsilon + \frac{q}{q-1}\beta^{-1/2q}\eta^{1-1/q}\sigma_q \lesssim \varepsilon + \frac{q}{q-1}\beta^{-1/2q}\eta^{1/2-1/q}\sigma_q$$

and the last term is bounded by plugging in M. Combining these bounds and upper bounding gives

$$\begin{aligned}\beta^{1+1/q}\|u^*-w\|_{\boldsymbol{\Sigma}_n} \lesssim \frac{q}{q-1}\eta^{1-1/q}\sigma_q + \eta^{1/2}\varepsilon + \eta^{1/8}R^{1/2}(\varepsilon + \frac{q}{q-1}\eta^{1/2-1/q}\sigma_q)^{1/2}\sqrt[8]{\frac{\log(\min(n,d)/\delta)}{n}} \\ + \eta^{1/4}R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}} + \eta^{-1/q}\min\left\{\sigma\sqrt{\frac{d+\log(2/\delta)}{n}}, (R\sigma)^{1/2}\sqrt[4]{\frac{\log(2/\delta)}{n}}\right\}\end{aligned}$$

which is the result in the oblivious setting. Dropping one of the terms in the min gives the adaptive setting result. □

We will now prove Theorem 6.5.8, so for the remainder of this section we proceed under assumption (6.14). In Lemma 6.5.11 we establish deterministic regularity conditions which hold with high probability. First, in Lemma 6.5.10 we prove a version of a standard maximal inequality used in the analysis of Ordinary Least Squares (see e.g. [RH17]), which shows that the norm of the noise vector shrinks when projecting onto a lower-dimensional subspace.

Lemma 6.5.10. *Suppose that $\xi_1, \dots, \xi_n$ is a martingale difference sequence with $|\xi_t| \le \sigma$ almost surely for all t. Suppose that V is a subspace of dimension d, $P_V : n \times n$ is the projection map onto V, and $\xi = (\xi_1, \dots, \xi_n)$. Then*

$$\|P_V\xi\| \lesssim \sigma\sqrt{d + \log(2/\delta)}$$

with probability at least $1-\delta$.

Proof. For $v \in V$ with $\|v\| = 1$, define $Z_v = \langle v, \xi\rangle = \sum_i v_i\xi_i$ which is a martingale. Since $|v_i\xi_i| \le \sigma M|v_i|$ almost surely and $\sum_i v_i^2 = 1$, it follows from Azuma-Hoeffding inequality (Fact 1.3.21) that

$$\Pr[|Z_v| \ge t] \le \exp\left(-\frac{Ct^2}{\sigma^2}\right)$$

By a well-known chaining argument over the sphere (Exercise 4.4.2 of [Ver18]), we can upper bound

$$\|P_V\xi\| = \max_{\|v\|=1} Z_v \le 2\max_{v\in\mathcal{N}} Z_v$$

where $\mathcal{N}$ is a 1/2-net of the unit sphere in V. Standard covering number bounds (e.g.

Corollary 4.2.13 of [Ver18]) let us take $|\mathcal{N}| \le 6^d$. Therefore by the union bound

$$\Pr\left[\max_{\|v\|=1} Z_v \ge t\right] \le 6^d \exp\left(-\frac{Ct^2}{\sigma^2}\right).$$

Taking $t = \Theta(\sigma\sqrt{(d + \log(2/\delta)})$ gives the result. □

Lemma 6.5.11. *For any $\alpha \in (0, \eta)$, suppose*

$$n \gtrsim \frac{\eta \log(\min(n, d)/\delta)}{\alpha^2} \tag{6.16}$$

For any sequence of $x_1, ..., x_n$ chosen during the process in Definition 6.4.1, we have that with probability at least $1 - \delta$ over the randomness of the Ber(η) coins generating $a_1^*, ..., a_n^*$, the following event holds. Let $\mathbf{\Sigma}' \triangleq \frac{1}{n}\sum_t a_t^* x_t x_t^\top$. Then:

1. $\frac{1}{n}\sum_{t=1}^n a_t^* \ge 1 - \eta - \alpha$.
2. $\left|\frac{1}{n}\sum_{t=1}^n a_t^* \xi_t \langle x_t, v\rangle\right| \le \sigma\lambda\|v\|_{\mathbf{\Sigma}'} + \sigma\lambda'\|v\|$ *for all v where:*
   (a) *In the special case of obliviously chosen covariates x_t: $\lambda \triangleq \Theta\left(\sqrt{\frac{d+\log(2/\delta)}{n}}\right)$ and $\lambda' \triangleq 0$.*
   (b) *In the general case of adaptive chosen covariates x_t: $\lambda \triangleq 0$ and $\lambda' \triangleq \Theta\left(\sqrt{\frac{\log(2/\delta)}{n}}\right)$*
3. $\mathbf{\Sigma}' \succeq (1 - \eta)\mathbf{\Sigma}_n - \alpha \cdot Id$.

Proof. We start with part 1. We have $\mathbb{E}[\frac{1}{n}\sum_{t=1}^n a_t^*] = 1 - \eta$ and using that the variance of *Ber*(p) is $p(1-p)$ we have $\mathbb{V}[a_t^*] \le \eta$. Then by Bernstein's inequality (Fact 1.3.22) we know that

$$\Pr[\frac{1}{n}\sum_{t=1}^n a_t^* \ge 1 - \eta - \alpha] \le \exp\left(-\frac{Cn\alpha^2}{\frac{1}{n}\sum \mathbb{V}[a_t^*] + \alpha}\right) \le \exp\left(-\frac{Cn\alpha^2}{\eta + \alpha}\right)$$

so we find $\frac{1}{n}\sum_{t=1}^n a_t' \ge 1 - \eta - \alpha$ with probability $1 - \delta$, provided $n = \Omega\left(\frac{\eta}{\alpha^2}\log(1/\delta)\right)$.

For part 2 (a), let $T \subseteq [n]$ denote the set of indices t for which $a_t^* = 1$; we now treat x and T as fixed and consider only ξ.

$$\frac{1}{n}\sum_{t=1}^n a_t^* \xi_t \langle x_t, v\rangle = \frac{1}{n}\langle (X')^T \xi, v\rangle = \frac{1}{n}\langle P_V \xi, (X')v\rangle \le \frac{1}{\sqrt{n}}\|P_V \xi\|\|v\|_{\mathbf{\Sigma}'}$$

where $X' : n \times d$ has rows $a_1^* x_1, \ldots, a_n^* x_n$, P_V is the projection onto subspace V and V is the column span of X', the last step applies Cauchy-Schwarz and the definition of $\mathbf{\Sigma}'$. Finally, the result follows by bounding $P_V \xi$ using Lemma 6.5.10.

For part 2(b), observe by Cauchy-Schwarz

$$\frac{1}{n}\sum_{t=1}^{n} a_t^* \xi_t \langle x_t, v\rangle = \left\langle \frac{1}{n}\sum_{t=1}^{n} a_t^* \xi_t x_t, v \right\rangle \le \left\| \frac{1}{n}\sum_{t=1}^{n} a_t^* \xi_t x_t \right\| \|v\|$$

and the sum inside the absolute value is a vector-valued martingale with step size at most σ, so the result follows from Theorem 1.3.23.

We now show part 3. We can apply the matrix Freedman inequality in the form of Corollary 1.3.27 to the matrix martingale difference sequence

$$(a_1^* - (1-\eta)) \cdot x_1 x_1^\top, (a_2^* - (1-\eta)) \cdot x_2 x_2^\top, \ldots, (a_t^* - (1-\eta)) \cdot x_t x_t^\top,$$

which satisfies $\mathbb{E}[(a_t^* - (1-\eta))^2 (x_t x_t^T)^2 | \mathcal{F}_{t-1}] \preceq \eta$ to get

$$\Pr\left[\left\| \frac{1}{n}\sum_{t=1}^{n} a_t^* \cdot x_t x_t^\top - \frac{(1-\eta)}{n}\sum_{t=1}^{n} x_t x_t^\top \right\| \ge \alpha\right] \le d_2(\alpha) \exp\left(\frac{-Cn\alpha^2}{\eta + \alpha}\right)$$

where the probability is over the randomness of the martingale, and from Corollary 1.3.27 we recall $f(x) = \min(1, x)$ hence

$$d_2(\alpha) = \operatorname{Tr} f(\alpha \sum_t \mathbb{E}[(a_t^* - (1-\eta))^2 (x_t x_t^T)^2]/\eta) \le \operatorname{Tr} f(\alpha \sum_t \mathbb{E}[x_t x_t^T]) \le \min\{d, \alpha n\}.$$

Using that $\alpha < \eta < 1$ by assumption, we conclude that as long as $n = \Omega(\frac{\eta \log(\min(n,d)/\delta)}{\alpha^2})$, then

$$\frac{1}{n}\sum_{t=1}^{n} a_t^* x_t x_t^\top \succeq (1-\eta)\mathbf{\Sigma}_n - \alpha \cdot \mathrm{Id},$$

from which part 3 follows. □

We are now ready to prove Theorem 6.5.8. We present the deterministic argument in Lemma 6.5.12 below, then show how combining it with the previous Lemma establishes the result.

Lemma 6.5.12. *Suppose that:*

1. *$x_1, \dots, x_n \in \mathbb{R}^d$, $a_1^*, \dots, a_n^* \in \{0, 1\}$, and for $t = 1, \dots, n$ we have a sequence y_t^* such that u^*, δ_t defined by (6.9) satisfies (6.10).*

2. *$y_1, \dots, y_n \in \mathbb{R}$ satisfy*
$$y_t = y_t^* + \xi_t$$
whenever $a_t^* = 1$ and $|\xi_t| \le \sigma$ as in (6.14).

3. *The conclusions of Theorem 6.5.8 are satisfied with parameters $\eta, \lambda, \lambda', \alpha$. The parameter β is defined in terms of η by (6.11).*

4. *$w \in \mathbb{R}^d$ and $a_1, \dots, a_n \in [0, 1]$ are feasible for Program 1 and satisfy the conclusion of Lemma 6.5.7, i.e. partial optimality and ε_{grad}-approximate first order stationarity.*

Then, the following conclusion holds:

$$\beta\|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim \eta\sigma + \eta^{1/2}\varepsilon + \sigma\lambda + \left(\varepsilon_{grad}^{1/2} + (R\sigma\lambda')^{1/2} + (R^2\alpha)^{1/4}\left(\sqrt{\eta^{1/2}\sigma + \varepsilon} + (R^2\alpha)^{1/4}\right)\right).$$

Proof of Theorem 6.5.8. Let w and $a_1, \dots, a_n$ be given by Lemma 6.5.7. Let T denote the subset of $t \in [n]$ for which $a_t^* = 1$, i.e. T is the set of rounds which are uncorrupted. We apply the first order optimality condition (6.12) with $v = u^*$ to get that

$$\frac{1}{n}\sum_t a_t(y_t - \langle w, x_t\rangle)\langle x_t, u^* - w\rangle \le \varepsilon_{grad}. \tag{6.17}$$

We will lower bound the left-hand side of (6.17) by considering the contribution from T and $[n] \setminus T$.

Contribution from T. For the former, we have

$$\begin{aligned}&\frac{1}{n}\sum_{t \in T} a_t(y_t - \langle w, x_t\rangle)\langle x_t, u^* - w\rangle \\ &\quad = \frac{1}{n}\sum_{t \in T} a_t(\delta_t + \xi_t + \langle u^* - w, x_t\rangle)\langle x_t, u^* - w\rangle\end{aligned}$$

$$= \frac{1}{n}\sum_{t\in T}\big[\underbrace{a_t\langle x_t, u^* - w\rangle^2}_{\textcircled{1}} + \underbrace{\xi_t\langle x_t, u^* - w\rangle}_{\textcircled{2}} - \underbrace{(1-a_t)\xi_t\langle x_t, u^* - w\rangle}_{\textcircled{3}} + \underbrace{a_t\delta_t\langle x_t, u^* - w\rangle}_{\textcircled{4}}\big] \tag{6.18}$$

We control all four terms separately, $\textcircled{1}$ being the dominant term. Define $\mathbf{\Sigma}'$ as in Lemma 6.5.11.

For $\textcircled{1}$, we write $a_t = 1 - (1 - a_t)$ and use Lemma 6.5.11 to get

$$\begin{aligned}\frac{1}{n}\sum_{t\in T} a_t\langle x_t, u^* - w\rangle^2 &= \frac{1}{n}\sum_{t\in T}\langle x_t, u^* - w\rangle^2 - \frac{1}{n}\sum_{t\in T}(1-a_t)\langle x_t, u^* - w\rangle^2 \\ &= \|u^* - w\|_{\mathbf{\Sigma}'}^2 - \frac{1}{n}\sum_{t\in T}(1-a_t)\langle x_t, u^* - w\rangle^2 \\ &\geq (1-\eta)\|u^* - w\|_{\mathbf{\Sigma}_n}^2 - \frac{1}{n}\sum_{t\in T}(1-a_t)\langle x_t, u^* - w\rangle^2 - O(\alpha R^2) \\ &\geq (1-2\eta)\|u^* - w\|_{\mathbf{\Sigma}_n}^2 - O(\alpha R^2),\end{aligned}$$

where in the last step we expanded the sum from $i \in T$ to $i \in [n]$ and then used the last constraint in Program 2.

For $\textcircled{2}$, note that

$$\frac{1}{n}\sum_{t\in T}\xi_t\langle x_t, u^* - w\rangle \leq O\left(\|u^* - w\|_{\mathbf{\Sigma}_n}\sigma\lambda + R\sigma\lambda'\right)$$

by Part 2 of Lemma 6.5.11, the fact that $\mathbf{\Sigma}' \preceq \mathbf{\Sigma}_n$, and $\|u^* - w\| \leq 2R$.

For $\textcircled{3}$, we have that

$$\frac{1}{n}\sum_{t\in T}(1-a_t)\xi_t\langle x_t, u^* - w\rangle \leq \left(\frac{1}{n}\sum_{t\in T}(1-a_t)\langle x_t, u^* - w\rangle^2\right)^{1/2}\left(\frac{1}{n}\sum_{t\in T}(1-a_t)\xi_t^2\right)^{1/2}. \tag{6.19}$$

By the last constraint in Program 2, we can upper bound the first factor on the right-hand side by $\sqrt{\eta\|u^* - w\|_{\mathbf{\Sigma}_n}^2 + \alpha\|u^* - w\|_2^2} \leq \eta^{1/2}\|u^* - w\|_{\mathbf{\Sigma}_n} + \sqrt{\alpha}R$. For the second factor, we can upper bound it by Holder's inequality as (recalling $\alpha \leq \eta$) we have

$$\left(\frac{1}{n}\sum_{t\in T}(1-a_t)\xi_t^2\right)^{1/2} \leq \sqrt{\eta}\sigma \tag{6.20}$$

so overall we get a bound on (6.19) of $\left(\eta^{1/2}\|u^* - w\|_{\mathbf{\Sigma}_n} + \sqrt{\alpha}R\right)\cdot\eta^{1/2}\sigma$.

Finally, for ④, note that first-order optimality of u^* implies that $\frac{1}{n}\sum_t \delta_t\langle x_t, u^* - w\rangle = 0$. So we can write

$$
\begin{aligned}
&\frac{1}{n}\sum_{t\in T} a_t\delta_t\langle x_t, u^* - w\rangle \\
&\quad= \frac{1}{n}\sum_{t\in[n]}(1-a_t)\delta_t\langle x_t, u^* - w\rangle - \frac{1}{n}\sum_{t\notin T} a_t\delta_t\langle x_t, u^* - w\rangle. \\
&\quad\le \left(\frac{1}{n}\sum_{t\in[n]}(1-a_t)^2\langle x_t, u^* - w\rangle^2\right)^{1/2}\left(\frac{1}{n}\sum_{t\in[n]}\delta_t^2\right)^{1/2} + \left(\frac{1}{n}\sum_{t\notin T}a_t^2\langle x_t, u^* - w\rangle^2\right)^{1/2}\left(\frac{1}{n}\sum_{t\notin T}\delta_t^2\right)^{1/2} \\
&\quad\le \left(\frac{1}{n}\sum_{t\in[n]}\delta_t^2\right)^{1/2}\cdot\left(\eta^{1/2}\|u^* - w\|_{\mathbf{\Sigma}_n} + \left(\eta\|u^* - w\|^2_{\mathbf{\Sigma}_n} + \alpha\|u^* - w\|_2^2\right)^{1/2}\right),
\end{aligned}
$$

where in the last step we used the fact that $(1-a_t)^2 \le 1 - a_t$ and $a_t^2 \le 1$ by the first constraint in Program 2, as well as the third constraint in Program 2 and Part 3.

Using (6.10) to upper bound the first parenthesized term, we conclude that

$$
\frac{1}{n}\sum_{t\in T} a_t\delta_t\langle x_t, u^* - w\rangle \le O(\varepsilon\cdot(\eta^{1/2}\|u^* - w\|_{\mathbf{\Sigma}_n} + \sqrt{\alpha}R)).
$$

Having controlled ①,②,③,④, from (6.18) we can therefore lower bound $\frac{1}{n}\sum_{t\in T} a_t(y_t - \langle w, x_t\rangle)\langle x_t, u^* - w\rangle$ by

$$
\begin{aligned}
&(1-2\eta)\|u^* - w\|^2_{\mathbf{\Sigma}_n} \\
&\quad - O\Bigg(\|u^* - w\|_{\mathbf{\Sigma}_n}\left(\sigma\lambda + \eta\sigma + \varepsilon\eta^{1/2}\right) + \alpha R^2 + R\sigma\lambda' + \sqrt{\alpha}R\eta^{1/2}\sigma\Bigg).
\end{aligned}
\tag{6.21}
$$

Contribution from $[n]\setminus T$. It remains to control the contribution to the left-hand side of (6.17) coming from the corrupted summands indexed by $[n]\setminus T$, which we do by upper bounding the term in absolute value. By Cauchy-Schwarz and $a_t^2 \leq a_t$,

$$\left|\frac{1}{n}\sum_{t \notin T} a_t(y_t - \langle w, x_t\rangle)\langle x_t, u^* - w\rangle\right| \leq \left(\frac{1}{n}\sum_{t\notin T} a_t(y_t - \langle w, x_t\rangle)^2\right)^{1/2}\left(\frac{1}{n}\sum_{t\notin T} a_t\langle x_t, u^* - w\rangle^2\right)^{1/2}$$

$$\leq \left(\frac{1}{n}\sum_{t\notin T} a_t(y_t - \langle w, x_t\rangle)^2\right)^{1/2}\left(\eta^{1/2}\|u^* - w\|_{\mathbf{\Sigma}_n} + \sqrt{\alpha}R\right) \tag{6.22}$$

where in the second step we used the fact that $a_t \in [0,1]$ along with Part 3 of Lemma 6.5.11. As for the first factor on the right-hand side, by the fact that $\{a_t\}$ were chosen in Program 2 to minimize $\frac{1}{n}\sum_{t\in[n]} a_t(y_t - \langle w, x_t\rangle)^2$, we have that

$$\frac{1}{n}\sum_{t\in[n]} a_t(y_t - \langle w, x_t\rangle)^2 \leq \frac{1}{n}\sum_{t\in[n]} a_t^*(y_t - \langle w, x_t\rangle)^2 = \frac{1}{n}\sum_{t\in T}(y_t - \langle w, x_t\rangle)^2,$$

hence rearranging gives

$$\begin{aligned}
&\frac{1}{n}\sum_{t\notin T} a_t(y_t - \langle w, x_t\rangle)^2 \\
&\quad\leq \frac{1}{n}\sum_{t\in T}(y_t - \langle w, x_t\rangle)^2 - \frac{1}{n}\sum_{t\in T}^{n} a_t(y_t - \langle w, x_t\rangle)^2 \\
&\quad= \frac{1}{n}\sum_{t\in T}(1 - a_t)(y_t - \langle w, x_t\rangle)^2 \\
&\quad= \frac{1}{n}\sum_{t\in T}(1 - a_t)(\langle u^* - w, x_t\rangle + \delta_t + \xi_t)^2 \\
&\quad\leq \frac{2 + 1/\beta}{n}\sum_{t\in T}(1 - a_t)\xi_t^2 + \frac{2 + 1/\beta}{n}\sum_{t\in T}(1 - a_t)\delta_t^2 + \frac{1 + 2\beta}{n}\sum_{t\in T}(1 - a_t)\langle u^* - w, x_t\rangle^2
\end{aligned}$$

where in the second-to-last step we used Cauchy-Schwarz to show

$$(a + b + c)^2 \leq (2 + 1/\beta)(a^2 + b^2 + c^2\beta) = (2 + 1/\beta)(a^2 + b^2) + (1 + 2\beta)c^2.$$

We continue and see

$$\frac{1}{n}\sum_{t\notin T} a_t(y_t - \langle w, x_t\rangle)^2 \le (1+2\beta)\eta\|u^* - w\|_{\mathbf{\Sigma}_n}^2 + O\left(\frac{1}{\beta}\eta\sigma^2 + \frac{1}{\beta}\varepsilon^2 + \alpha R^2\right)$$

where in the last step we used Holder's inequality and (6.14), (6.10), and the last constraint in Program 2 with $\|u^* - w\| \le 2R$.

So by (6.22) we can upper bound $\frac{1}{n}\sum_{t\notin T} a_t(y_t - \langle w, x_t\rangle)\langle x_t, u^* - w\rangle$ by

$$\begin{aligned}&\left((1+2\beta)^{1/2}\eta^{1/2}\|u^* - w\|_{\mathbf{\Sigma}_n} + O\left(\beta^{-1/2}\eta^{1/2}\sigma + \beta^{-1/2}\varepsilon + \alpha^{1/2}R\right)\right)\left(\eta^{1/2}\|u^* - w\|_{\mathbf{\Sigma}_n} + \sqrt{\alpha}R\right)\\ &= (1+2\beta)^{1/2}\eta\|u^* - w\|_{\mathbf{\Sigma}_n}^2 + O(\beta^{-1/2}\eta\sigma + \beta^{-1/2}\eta^{1/2}\varepsilon + \alpha^{1/2}\eta^{1/2}R)\|u^* - w\|_{\mathbf{\Sigma}_n} + \mathcal{E}, \quad (6.23)\end{aligned}$$

where

$$\mathcal{E} \triangleq \sqrt{\alpha}R \cdot O(\beta^{-1/2}(\eta^{1/2}\sigma + \varepsilon) + \sqrt{\alpha}R)$$

captures all the error terms that vanish as $\alpha \to 0$.

Combining. Putting the bounds on $\frac{1}{n}\sum_{t\in T} a_t(y_t - \langle w, x_t\rangle)\langle x_t, u^* - w\rangle$ and $\frac{1}{n}\sum_{t\notin T} a_t(y_t - \langle w, x_t\rangle)\langle x_t, u^* - w\rangle$ by (6.21) and (6.23) together with (6.12), we conclude that

$$(1 - 3\eta - \sqrt{2\beta}\cdot\eta))\|u^* - w\|_{\mathbf{\Sigma}_n}^2 \le O\left(\beta^{-1/2}\eta\sigma + \beta^{-1/2}\eta^{1/2}\varepsilon + \alpha^{1/2}R + \sigma\lambda\right)\|u^* - w\|_{\mathbf{\Sigma}_n} + \mathcal{E}',$$

where $\mathcal{E}' \triangleq \varepsilon_{grad} + R\sigma\lambda' + O(\mathcal{E})$. We do case analysis based on which of the two terms on the rhs of the above bound dominates:

1. In the first case, the first term is at least as large as $\mathcal{E}'$. Then the bound simplifies to

$$(1 - 3\eta - \sqrt{2\beta}\cdot\eta)\|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim \beta^{-1/2}\eta\sigma + \beta^{-1/2}\eta^{1/2}\varepsilon + \alpha^{1/2}R + \sigma\lambda$$

2. Otherwise, $\mathcal{E}'$ is larger than the first term. Then taking a square root the bound can be simplified to

$$(1 - 3\eta - \sqrt{2\beta}\cdot\eta)\|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim \varepsilon_{grad}^{1/2} + (R\sigma\lambda')^{1/2} + \mathcal{E}^{1/2}.$$

In either case, since $\alpha^{1/2}R = O(\mathcal{E}^{1/2})$ we see the inequality

$$(1 - 3\eta - \sqrt{2\beta} \cdot \eta)\|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim \beta^{-1/2}\eta\sigma + \beta^{-1/2}\eta^{1/2}\varepsilon + \sigma\lambda + \varepsilon_{grad}^{1/2} + (R\sigma\lambda')^{1/2} + \mathcal{E}^{1/2}$$

holds. Since $\beta = (1/3 - \eta)^2$ and $\eta < 1/3$ we know

$$(1 - 3\eta - \sqrt{2\beta}\eta) \geq 3\sqrt{\beta} - \sqrt{2\beta} = \Theta(\sqrt{\beta})$$

so we get a final bound of

$$\begin{aligned}
&\|u^* - w\|_{\mathbf{\Sigma}_n} \\
&\lesssim \beta^{-1}\eta\sigma + \beta^{-1}\eta^{1/2}\varepsilon + \beta^{-1/2}\sigma\lambda + \beta^{-1/2}(\varepsilon_{grad}^{1/2} + (R\sigma\lambda')^{1/2} + \mathcal{E}^{1/2}) \\
&\lesssim \beta^{-1}\eta\sigma + \beta^{-1}\eta^{1/2}\varepsilon + \beta^{-1/2}\sigma\lambda + \beta^{-1/2}(\varepsilon_{grad}^{1/2} + (R\sigma\lambda')^{1/2} + (R^2\alpha)^{1/4}(\beta^{-1/4}\sqrt{\eta^{1/2}\sigma + \varepsilon} + (R^2\alpha)^{1/4})).
\end{aligned}$$

Using $\beta < 1$ to upper bound all of the powers of β by β^{-1} gives the result. □

Now combining our claims proves Theorem 6.5.8:

Proof of Theorem 6.5.8. **Oblivious covariates.** By Lemma 6.5.12 and Lemma 6.5.11 we know the output w of Lemma 6.5.7 with $\varepsilon_{grad} = O(\sigma^2\lambda^2) = O(\sigma^2\frac{d+\log(2/\delta)}{n})$ satisfies

$$\beta\|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim \eta\sigma + \eta^{1/2}\varepsilon + \sigma\sqrt{\frac{d + \log(2/\delta)}{n}} + (R^2\alpha)^{1/4}(\sqrt{\eta^{1/2}\sigma + \varepsilon} + (R^2\alpha)^{1/4})$$

with probability at least $1 - \delta$, as long as $\alpha < \eta$ and (6.16) holds:

$$n \gtrsim \frac{\eta \log(\min(n, d)/\delta)}{\alpha^2}.$$

Based on this we take $\alpha = \Theta\left(\sqrt{\frac{\eta\log(\min(n,d)/\delta)}{n}}\right)$ and require

$$n \gtrsim \log(\min(n, d)/\delta)/\eta$$

so that $\alpha < \eta$. Then we can write the error bound as

$$\beta\|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim \eta\sigma + \eta^{1/2}\varepsilon + \eta^{1/8}R^{1/2}\sqrt[8]{\frac{(\eta^{1/2}\sigma + \varepsilon)^4 \log(\min(n,d)/\delta)}{n}} \\ + \eta^{1/4}R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}} + \sigma\sqrt{\frac{d + \log(2/\delta)}{n}}.$$

Adaptive covariates. The only change is that the term $\sigma\lambda$ disappears and the term

$$(R\sigma\lambda')^{1/2} = (R\sigma)^{1/2}\sqrt[4]{\frac{\log(2/\delta)}{n}}$$

appears, which gives

$$\beta\|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim \eta\sigma + \eta^{1/2}\varepsilon + \eta^{1/8}R^{1/2}\sqrt[8]{\frac{(\eta^{1/2}\sigma + \varepsilon)^4 \log(\min(n,d)/\delta)}{n}} \\ + \eta^{1/4}R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}} + (R\sigma)^{1/2}\sqrt[4]{\frac{\log(2/\delta)}{n}}.$$

Since this bound also applies in the special case of oblivious covariates, we get the stated result. □

Subgaussian noise

In this section we consider the case where the noise is subgaussian. Subgaussian random variables are in L_q for every q, so we could analyze them using our previous result (taking $q = \log(1/\eta)$), but since subgaussian noise behaves similar to bounded noise, we can optimize the argument by avoiding truncation. This yields the following result, which in the uncontaminated $\eta = 0$ setting with oblivious covariates, recovers the same (minimax optimal) rate achieved by Ordinary Least Squares/Ridge Regression and gracefully degrades with increasing η.

Theorem 6.5.13 (SCRAM Guarantee with Subgaussian Noise). *Suppose that $\eta < 1/3$ is an upper bound on the contamination level, define β as in* (6.11)*, and suppose for some $\sigma \geq 0$*

that for all t the noise ξ_t is σ^2-subgaussian. Then if $\eta = 0$ or

$$n \gtrsim \log(\min(n,d)/\delta)/\eta,$$

$\alpha = \Theta\left(\sqrt{\frac{\eta \log(d/\delta)}{n}}\right)$ *and* $\overline{\eta} = \eta$*, the output* w *of* SCRAM *with poly*$(R/\sigma, \log(2/\delta), d, n)$ *many steps satisfies for oblivious covariates the bound*

$$\beta \|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim c_{\delta,\eta,n}\eta\sigma + \eta^{1/2}\varepsilon + \eta^{1/8}R^{1/2}(\sqrt{c_{\delta,\eta,n}\eta}\sigma + \varepsilon)^{1/2}\sqrt[8]{\frac{\log(\min(n,d)/\delta)}{n}} \\ + \eta^{1/4}R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}} + \min\left\{\sigma\sqrt{\frac{d + \log(2/\delta)}{n}}, (R\sigma)^{1/2}\sqrt[4]{\frac{\log(2/\delta)}{n}}\right\}$$

with probability at least $1 - \delta$*, where*

$$c_{\delta,\eta,n} \triangleq \sqrt{\log(1/\eta)} \exp\left(\max\left(1, \frac{\log\log(1/\delta) \cdot \log(1/\eta)}{2\log(n)}\right)\right) \tag{6.24}$$

captures a logarithmic term which is $O(\sqrt{\log(1/\eta)})$ *assuming* $\log n \geq (1/100)\log\log(1/\delta)\log(1/\eta)$*. In the more general case of adaptive covariates,* SCRAM *satisfies the bound*

$$\beta \|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim c_{\delta,\eta,n}\eta\sigma + \eta^{1/2}\varepsilon + \eta^{1/8}R^{1/2}(\sqrt{c_{\delta,\eta,n}\eta}\sigma + \varepsilon)^{1/2}\sqrt[8]{\frac{\log(\min(n,d)/\delta)}{n}} \\ + \eta^{1/4}R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}} + (R\sigma)^{1/2}\sqrt[4]{\frac{\log(2/\delta)}{n}}$$

i.e. the same bound except the last term was changed.

Proof. The proof is the same as Theorem 6.5.8 with a few modifications which we describe now. The main difference is in the use of Holder's inequality to bound terms including noise, e.g. (6.20). In this case, since ξ_t is no longer bounded we use for $q = \min(2\log(n)/\log\log(1/\delta), \log(1/\eta))$ that by Holder's inequality

$$\left(\frac{1}{n}\sum_{t \in T}(1 - a_t)\xi_t^2\right)^{1/2} \leq \left(\frac{1}{n}\sum_{t \in T}(1 - a_t)\right)^{1/2p}\left(\frac{1}{n}\sum_{t \in T}\xi_t^{2q}\right)^{1/2q} \lesssim \eta^{1/2 - 1/2q}\sigma\sqrt{q}$$

where $1/p + 1/q = 1$ and we used Lemma 6.5.14 below. Plugging in the value of q gives an upper bound of

$$\sigma\eta^{1/2}\sqrt{\log(1/\eta)}\cdot\exp\left(\max\left(1,\frac{\log\log(1/\delta)\cdot\log(1/\eta)}{4\log(n)}\right)\right).$$

The other change is that in Lemma 6.5.11, we can use the subgaussian property to establish Part 2 without needing boundedness of the noise: we use the generalization of the vector Azuma-Hoeffding inequality to the subgaussian step size setting, Theorem 1.3.23. □

The following Lemma 6.5.14 gives a fairly sharp upper deviation bound for power sums of subgaussian random variables. This result is not so easy to prove directly, but follows from the main result of [Lat97].

Lemma 6.5.14. *Suppose that $Z_1,\dots,Z_n$ are independent σ^2-subgaussian random variables. Then*

$$\left(\frac{1}{n}\sum_i |Z_i|^p\right)^{1/p}\lesssim\sigma\sqrt{p}$$

with probability at least $1-\delta$, provided $n\geq\log(2/\delta)^{p/2}$.

Proof. We rescale so that $\sigma=1$. In this proof we use the notation $\|X\|_q=\mathbb{E}[|X|^q]^{1/q}$ for the function space L_p norm.

Define $S=\sum_i|Z_i|^p$. By Markov's inequality, $\Pr[S\geq t]=\Pr[S^q\geq t^q]\leq\frac{\|S\|_q^q}{t^q}$ for any $q\geq 1$. By Theorem 1 and Corollary 1 of [Lat97], for $q\leq n$ we have

$$\|S\|_q\lesssim\sup\left\{(q/s)(n/q)^{1/s}\max_i\|Z_i^p\|_s : 1\leq s\leq q\right\}.$$

We observe from standard subgaussian moment bounds [RH17, Ver18] that

$$\|Z_i^p\|_s=\|Z_i\|_{sp}^p\lesssim(esp)^{p/2}$$

so

$$\begin{aligned}\|S\|_q&\lesssim(ep)^{p/2}q\sup\left\{(n/q)^{1/s}s^{p/2-1} : 1\leq s\leq q\right\}\\&=(ep)^{p/2}q\sup\left\{\exp((1/s)\log(n/q)+(p/2-1)\log(s) : 1\leq s\leq q\right\}.\end{aligned}$$

We consider the optimization over s inside the exponential. The unique critical point is when $-s^{-2}\log(n/q) + (p/2-1)/s = 0$, i.e. $s = \log(n/q)/(p/2-1)$. Since the function goes to infinity as $s \to 0$ and $s \to \infty$, that critical point must be a minimum. It suffices therefore to consider the boundary points. This shows

$$\|S\|_q \lesssim (ep)^{p/2}\left(n + n^{1/q}q^{p/2-1/q}\right) \lesssim (ep)^{p/2}\left(n + n^{1/q}q^{p/2}\right)$$

using $\max_{q \ge 1} q^{-1/q} = 1$. Now taking $t = e\|S\|_q$ and $q = \log(1/\delta)$ shows

$$S \le e\|S\|_q \lesssim (ep)^{p/2}n(1 + n^{1/\log(1/\delta)-1}\log(1/\delta)^{p/2})$$

with probability at least $1-\delta$. In particular, if $n \ge \log(1/\delta)^{p/2}$ then

$$S \lesssim (ep)^{p/2}n(1 + e^{(p/2)\log\log(1/\delta)/\log(1/\delta)}) \le e^p p^{p/2} n$$

as claimed. □

6.5.5 Stochastic Setting and Generalization Bounds

Finally, we note that while the guarantees in this section so far have been in the usual *fixed design* setting, from these guarantees we also obtain strong results in the stochastic (or *random design*) setting often considered in statistical learning. We first review the setup. We assume there exists a joint distribution $\mathcal{D}_{x,y^*}$ over clean examples (x, y^*) and clean training data $(x_1, y_1^*), \dots, (x_n, y_n^*)$ are sampled identically from this distribution. We define the *population loss* to be the error of w on a fresh clean example (x, y^*) in squared loss,

$$L(w) = \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_{x,y^*}}[(y^* - \langle w, x\rangle)^2],$$

and our goal is to find a near minimizer of the population loss, i.e. compute $\widehat{w}$ from training data such that $\|\widehat{w}\| \le R$ and the gap in population loss $L(\widehat{w}) - L(u^*)$ is as small as possible, where we define

$$u^* \triangleq \arg\min_{\|u\| \le R} L(u)$$

to be the optimal predictor of norm at most R. Concretely, the gap in loss can be rewritten in a more convenient form in the following way

$$\begin{aligned} L(\widehat{w}) - L(u^*) &= \mathbb{E}[(y^* - \langle u^*, x\rangle + \langle u^* - w, x\rangle)^2] - \mathbb{E}[(\langle u^* - w, x\rangle)^2] \\ &= \mathbb{E}[\langle \widehat{w} - u^*, x\rangle^2] + 2\,\mathbb{E}[(y^* - \langle u^*, x\rangle)\langle u^* - w, x\rangle)] \\ &= \|\widehat{w} - u^*\|^2_{\mathbf{\Sigma}^*} + 2\,\mathbb{E}[(y^* - \langle u^*, x\rangle)\langle u^* - w, x\rangle)] \end{aligned}$$

where $\mathbf{\Sigma}^* = \mathbb{E}_{\mathcal{D}_x}[xx^T]$ is the second moment matrix, i.e. covariance matrix if x is mean zero, and the second term on the rhs is $O(\varepsilon\|u^* - w\|_{\mathbf{\Sigma}^*})$ under (6.25), showing that as $\varepsilon \to 0$, the slightly different goals of minimizing $\|u^* - w\|_{\mathbf{\Sigma}^*}$ and minimizing the suboptimality in population loss become exactly equivalent. As before, we assume that the conditional law of y^* given x is

$$y^* = \langle w^*, x\rangle + \varepsilon_x + \xi \tag{6.25}$$

where $\|w^*\| \le R$, $|\varepsilon_x| \le \varepsilon$ is misspecification, and ξ is noise independent of x, ε_x. If $\varepsilon = 0$ then we can take $u^* = w^*$, otherwise we always have $\|u^* - w^*\|_{\Sigma} \le 2\varepsilon$ since $|\langle w^*, x\rangle - \mathbb{E}[y^*|x]| \le \varepsilon$ and u^* is only closer in average squared loss.

We will use the following Lemma to relate the error when measured according to the population second moment matrix $\mathbf{\Sigma}^*$ and the random matrix $\mathbf{\Sigma}_n$: this "localized" generalization bound follows from the main result of [SST10], which builds upon the local Rademacher complexity framework of [BBM+05]; it gives tighter results than e.g. naively applying matrix concentration because it focuses in on the behavior of the bottom singular value. We note that the general connection between generalization theory and the bottom singular value of the empirical covariance matrix is well known and has been used in other contexts, see e.g. [KM15].

Lemma 6.5.15 (Consequence of Theorem 1 of [SST10])**.** *Suppose w^* is any fixed vector with $\|w^*\| \le R$. Suppose that $x_1, \ldots, x_n$ are iid copies of a random variable x with $\mathbf{\Sigma}^* = \mathbb{E}[xx^T]$ and $\|x\| \le 1$ almost surely. Uniformly over all w with $\|w\| \le R$ and with probability at least*

$1-\delta$, *where* $\mathbf{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ *is the empirical second moment matrix, the following holds:*

$$\|w-w^*\|_{\mathbf{\Sigma}^*}^2 - \|w-w^*\|_{\mathbf{\Sigma}_n}^2 \lesssim R\|w-w^*\|_{\mathbf{\Sigma}_n}\sqrt{\frac{\log^3(n)+\log(1/\delta)}{n}} + \frac{R^2(\log^3(n)+\log(1/\delta))}{n}$$

and as a consequence

$$\|w-w^*\|_{\mathbf{\Sigma}^*} \lesssim \|w-w^*\|_{\mathbf{\Sigma}_n} + R\sqrt{\frac{\log^3(n)+\log(1/\delta)}{n}}.$$

Proof. We explain how this follows from Theorem 1 of [SST10], which requires us to interpret the gap $\|w-w^*\|_{\mathbf{\Sigma}^*}^2 - \|w-w^*\|_{\mathbf{\Sigma}_n}^2$ as the generalization gap in a statistical learning problem; we refer the reader there for a detailed explanation of the setup. We now describe the new learning problem, which is not the same as the one considered outside the proof of this Lemma, as it has no noise, contamination, or misspecification. In this problem, x is defined as in the theorem statement, and the label $y = \langle w^*, x\rangle$. The population loss is $\mathbb{E}[\ell(y-\langle w,x\rangle)] = \langle w^*-w, \mathbf{\Sigma}(w^*-w)\rangle$ where $\ell(e) = e^2$ is the squared loss which is 1-smooth, and the empirical loss is $\frac{1}{n}\sum_{i=1}^n \ell(y_i - \langle w,x\rangle) = \langle w^*-w, \mathbf{\Sigma}_n(w^*-w)\rangle$. We observe that the loss $\ell(y_i - \langle w,x\rangle)$ is upper bounded by $4R^2$ almost surely, and finally we use (see [SST10]) that the Rademacher complexity R_n of the function class $\{x \mapsto \langle w,x\rangle : \|w\| \le R\}$ is $O(R\sqrt{1/n})$ where n is the number of samples. Plugging all of this information into Theorem 1 of [SST10] gives

$$\|w-w^*\|_{\mathbf{\Sigma}^*}^2 - \|w-w^*\|_{\mathbf{\Sigma}_n}^2 \lesssim \|w-w^*\|_{\mathbf{\Sigma}_n}\left(R\log^{1.5}(n)\sqrt{\frac{1}{n}} + R\sqrt{\frac{\log(1/\delta)}{n}}\right) + \log^3(n)\frac{R^2}{n} + \frac{R^2\log(1/\delta)}{n}$$

and up to constants this is equivalent to the first stated bound. The second (weaker) bound follows by adding $\|w-w^*\|_{\mathbf{\Sigma}_n}^2$ to the right hand side and taking a square root. □

Given this result, we can immediately obtain versions of all of the previous results for the stochastic setting (e.g. Theorem 6.5.1, Theorem 6.5.13, Theorem 6.5.8). We describe a more involved application below in Section 6.5.6, where we obtain improved results for learning in the stochastic setting by using this generalization bound combined with the generalized median of [M$^+$15].

We note that in the case where the contexts are chosen stochastically, [SLX20] recently showed that a modified version of the reduction from [FR20] can reduce from stochastic contextual bandits to offline regression with stochastic contexts. It should be possible to combine this reduction with our results; however, we omit the details since we will give an algorithm for the more general online setting anyway.

6.5.6 Heavy-Tailed Setting Using Geometric Median

In this section, we focus on the setting where the noise $\{\xi_t\}$ is in L_q with $q \geq 2$ and obtain improved sample complexity guarantees. In this context, there is a fairly general way to boost the success probability of algorithms by using the geometric median [M+15] or a related high-dimensional median of [HS16]; in the context of (uncontaminated) ridge regression itself, this kind of estimator was considered in [HS16], see Theorem 21 there. To take advantage of the geometric median, we start by establishing improved guarantees for our algorithm, but which hold with only a fixed probability of success.

Lemma 6.5.16. *Suppose that $\eta < 1/3$ is an upper bound on the contamination level, define β as in (6.11), and suppose for some $q \in [2, \infty], \sigma_q \geq 0$ and all t that*

$$\|\xi_t\|_q \triangleq \mathbb{E}[|\xi|^q]^{1/q} \leq \sigma_q.$$

Then provided $\eta = 0$ or

$$n \gtrsim \log(\min(n, d))/\eta,$$

we can take $\alpha = \Theta\left(\sqrt{\frac{\eta \log(d)}{n}}\right)$ and $\overline{\eta} = \eta + \Theta(\eta\sqrt{\beta})$ such that the output w of SCRAM *with poly$(R/\sigma, d, n)$ many steps satisfies for oblivious covariates the bound*

$$\beta^{1+1/q}\|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim \eta^{1-1/q}\sigma_q + \eta^{1/2}\varepsilon + \eta^{1/8}R^{1/2}(\varepsilon + \eta^{1/2-1/q}\sigma_q)^{1/2}\sqrt[8]{\frac{\log(\min(n, d))}{n}} + \eta^{1/4}R\sqrt[4]{\frac{\log(\min(n, d))}{n}} + \min\left\{\sigma\sqrt{\frac{d}{n}}, (R\sigma)^{1/2}\sqrt[4]{\frac{1}{n}}\right\} \tag{6.26}$$

with probability at least 0.99*. In the more general case of adaptive covariates, it satisfies the* *bound*

$$\beta^{1+1/q}\|u^* - w\|_{\mathbf{\Sigma}_n} \lesssim \eta^{1-1/q}\sigma_q + \eta^{1/2}\varepsilon + \eta^{1/8}R^{1/2}(\varepsilon + \eta^{1/2-1/q}\sigma_q)^{1/2}\sqrt[8]{\frac{\log(\min(n,d))}{n}}$$
$$+ \eta^{1/4}R\sqrt[4]{\frac{\log(\min(n,d))}{n}} + (R\sigma)^{1/2}\sqrt[4]{\frac{1}{n}}$$

i.e. the same bound except the last term was changed.

Proof. The proof is the same as Theorem 6.5.1 except that we change the analysis of Part 2 in Lemma 6.5.11 to improve the final term in our bound. First, we observe that truncating the noise ξ_i and recentering (the first part of the proof of Theorem 6.5.1) can only make the L_q norm of $|\xi_i|$ larger by a factor of 2 (see the proof of Lemma 2.6.8 in [Ver18]); in what follows, we let ξ_i denote the possibly truncated and recentered noise and use this fact. Now we consider the application of Theorem 6.5.8 in the proof of Theorem 6.5.1 and show how in Part 2 of Lema 6.5.11 we can replace the infinity norm of the noise by the smaller quantity σ_q. Specifically this occurs in Part 2 of Lemma 6.5.11.

For Part 2 (a), we replace Lemma 6.5.10 by the following argument based on Chebyshev's inequality. Let $\xi = (\xi_1, \dots, \xi_n)$ be the vector of (truncated) noise and observe that $\sqrt{\mathbb{E}[\xi_i^2]} \le \mathbb{E}[|\xi_i|^q]^{1/q} = O(\sigma_q)$ by Jensen to see

$$\Pr[\|P_V\xi\| \ge s] \le \frac{\mathbb{E}[\|P_V\xi\|^2]}{s^2} \le \frac{\langle P_V P_V^T, \sigma_q^2 I\rangle}{s^2} \le \frac{2d\sigma_q^2}{s^2}.$$

Similarly for Part 2 (b), we use Chebyshev's inequality and the fact that

$$\mathbb{E}[\|\sum_i \xi_i x_i\|^2] = \sum_i \mathbb{E}[\xi_i^2]\|x_i\|^2 \le 2n\sigma_q^2.$$

to get that $\|\frac{1}{n}\sum_i \xi_i x_i\| = O(\sigma_q/\sqrt{\delta n})$ with probability at least $1 - \delta$.

Taking the union bound and using these estimates in the analysis, otherwise unchanged from the proof of Theorem 6.5.1, gives the result. □

Given this result, we run the algorithm multiple times, and take the geometric median, as described in SCRAM-GM. We recall the key guarantee for geometric median from [M+15] in its contrapositive form, which informally says that if a $1-\alpha > 1/2$ proportion of points cluster near each other, then the geometric median will successfully return a point close to this cluster.

Lemma 6.5.17 (Lemma 2.1 (a) of [M+15])**.** *Suppose $x_1, \dots, x_n$ are points in a d-dimensional Euclidean space with norm $\|\cdot\|$. Suppose $z \in \mathbb{R}^d, r > 0, \alpha \in (0, 1/2)$, let $C_\alpha \triangleq (1-\alpha)\sqrt{\frac{1}{1-2\alpha}}$, and let*

$$y = \arg\min_y \sum_{i=1}^n \|y - x_i\|,$$

be the geometric median. If

$$\#\{i : \|x_j - z\| > r\} \leq \alpha n$$

then $\|y - z\| \leq C_\alpha r$.

Algorithm 16: SCRAM-GM$(x_t, y_t, \delta, \overline{\eta}, \alpha)$

```
Input: Input data $(x_t, y_t)_{t=1}^n$.
Output: Predictor $\widehat{w}$.
Shuffle the data and split into two equal sized groups $C_1, C_2$ and split $C_1$ into
  $k \triangleq \Theta(\log(1/\delta))$ equal-size buckets $B_1, \dots, B_k$.
Run SCRAM with parameters $\overline{\eta}, \alpha$ on each bucket to get predictors $w_1, \dots, w_k$.
Form the geometric median

  $$\widehat{w} = \arg\min_y \sum_{i=1}^k \|y - w_i\|_{\mathbf{\Sigma}_{C_2}}$$

  where $\mathbf{\Sigma}_{C_2} \triangleq \frac{1}{|C_2|} \sum_{t \in C_2} x_t x_t^T$.
return $\widehat{w}$.
```

Theorem 6.5.18. *Suppose that $\eta < 1/3$ is an upper bound on the contamination level, define β as in (6.11), and suppose for some $q \in [2, \infty], \sigma_q \geq 0$ and all t that*

$$\|\xi_t\|_q \triangleq \mathbb{E}[|\xi|^q]^{1/q} \leq \sigma_q.$$

Then provided $\eta = 0$ or

$$\eta \cdot n \gtrsim \log(\min(n, d)),$$

we can take $\alpha = \Theta\left(\sqrt{\frac{\eta \log(d)\log(1/\delta)}{n}}\right)$ *and* $\overline{\eta} = \eta + \Theta(\eta\sqrt{\beta})$ *such that the output* w *of* SCRAM-GM *satisfies*

$$\begin{aligned}\beta^{1+1/q}\|w^* - w\|_{\mathbf{\Sigma}^*} \lesssim \eta^{1-1/q}\sigma_q + \varepsilon + \eta^{1/8}R^{1/2}(\varepsilon + \eta^{1/2-1/q}\sigma_q)^{1/2}\sqrt[8]{\frac{\log(\min(n,d))\log(1/\delta)}{n}} \\ + \eta^{1/4}R\sqrt[4]{\frac{\log(\min(n,d))\log(1/\delta)}{n}} + \min\left\{\sigma\sqrt{\frac{d\log(1/\delta)}{n}}, (R\sigma)^{1/2}\sqrt[4]{\frac{\log(1/\delta)}{n}}\right\} \\ + R\sqrt{\frac{\log^3(n)\log(1/\delta)}{n}}\end{aligned}$$

with probability at least $1-\delta$*.*

Proof. Combining Lemma 6.5.16 and Lemma 6.5.15 gives

$$\|w_i - w^*\|_{\mathbf{\Sigma}^*} \le r \triangleq C(r_0 + R\sqrt{\frac{\log^3(n)\log(1/\delta)}{n}})$$

with probability at least 0.98, where r_0 is the right hand side of (6.26) plus ε (to replace u^* by w^*) and C is an absolute constant. Hence by applying Lemma 6.5.16, independence, and Hoeffding's inequality we see that

$$\#\{i : \beta^{1+1/q}\|w_i - w^*\|_{\mathbf{\Sigma}^*} \le r\} \ge 0.97k$$

with probability at least $1-\delta$ where r is the right hand side of (6.26), including the constant factor. We condition on this event in what follows.

Note that by Bernstein's inequality, for any particular w

$$\left|\|w - w^*\|^2_{\mathbf{\Sigma}_{C_2}} - \|w - w^*\|^2_{\mathbf{\Sigma}^*}\right| \lesssim \|w - w^*\|_{\mathbf{\Sigma}^*} R\sqrt{\frac{\log(1/\delta)}{n}} + \frac{R^2\log(1/\delta)}{n}$$

with probability $1-\delta$, where we used that $\mathbb{E}[\langle w - w^*, X\rangle^4] \le 4R^2\,\mathbb{E}[\langle w - w^*, X\rangle^2]$ to upper bound the variance term. Note that $R\sqrt{\log(1/\delta)/n} = O(r)$. Hence union bounding over $w_1, \ldots, w_k$ we find with probability at least $1-\delta$

$$\#\{i : \beta^{1+1/q}\|w_i - w^*\|_{\Sigma_{C_2}} = O(r)\} \ge 0.97k$$

which by Lemma 6.5.17 gives the result in the norm $\|\cdot\|_{\Sigma_{C_2}}$ and combined with Lemma 6.5.15 gives the desired result in $\|\cdot\|_{\mathbf{\Sigma}^*}$. □

6.6 Optimal Breakdown Point via Sum of Squares Programming

As previously explained, the breakdown point for the estimator SCRAM is at $\eta = 1/3$, because when $\eta \geq 1/3$ the landscape of its objective exhibits bad local minima. Remarkably, if we instead use the natural degree-4 Sum of Squares relaxation of our original combinatorial optimization problem, it maintains the same statistical guarantees as SCRAM (including the optimal η dependence) while also managing to escape the bad local minima of the nonconvex problem and achieve optimal breakdown point $\eta = 1/2$.

As we will see in the analysis, the fundamental fact we use which is true for the SoS relaxation (Program 3) and not true for an arbitrary stationary point or local minima of Program 1 is that the SoS relaxation always computes a lower bound on the original (unrelaxed) problem (6.2), allowing us to compare objective values with the ground truth pair (a^*, u^*).

6.6.1 SoS Algorithm and Analysis

In this section we state the main guarantee for our algorithm when η is large, as well as the result of combining this guarantee with the ones in Section 6.5 to obtain a guarantee for the full range of possible η.

As in Section 6.5, the constants in our result must deteriorate slightly as we approach the (optimal) breakdown point $\eta = 1/2$, so we introduce a parameter ρ which tracks the distance to $1/2$; as long as we are strictly bounded away from this point, ρ is upper bounded by a constant and can be ignored.

We first state a result for *bounded* noise.

Theorem 6.6.1. *Suppose that the contamination rate is* $\eta \in (0.3, 1/2)$*, define* $0 < \rho < 1$ *by*

$\eta = \frac{1}{2+2\rho^2}$, and suppose

$$n \gtrsim \log(\min(n,d)/\delta). \tag{6.27}$$

If the noise $\{\xi_t\}$ satisfies $|\xi_t| \leq \sigma$ for all t with probability 1, then there is a poly(n,d) *algorithm which takes as input $(x_1,y_1),...,(x_n,y_n)$ and, with probability at least $1-\delta$, outputs a vector $\widetilde{w}$ which satisfies*

$$\rho^2\|\widetilde{w}-w^*\|_{\mathbf{\Sigma}_n} \lesssim \sigma+\varepsilon+\rho R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}}+\min\left\{\sigma\sqrt{\frac{d+\log(1/\delta)}{n}},(R\sigma)^{1/2}\rho\cdot\sqrt[4]{\frac{\log(1/\delta)}{n}}\right\}$$

for oblivious covariates and

$$\rho^2\|\widetilde{w}-w^*\|_{\mathbf{\Sigma}_n} \lesssim \sigma+\varepsilon+\rho R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}}+(R\sigma)^{1/2}\rho\cdot\sqrt[4]{\frac{\log(1/\delta)}{n}}$$

for the more general case of adaptive covariates.

By a simple truncation argument, we can also obtain versions of this result for weakly L_q and subgaussian noise. For brevity, we only state the latter:

Theorem 6.6.2. *Let η, ρ, n satisfy the hypotheses of Theorem 6.6.1. If the noise $\{\xi_t\}$ is σ^2-subgaussian, then there is a* poly(n,d) *algorithm which takes as input $(x_1,y_1),...,(x_n,y_n)$ and outputs a vector $\widetilde{w}$ which satisfies*

$$\rho^2\|\widetilde{w}-w^*\|_{\mathbf{\Sigma}_n} \lesssim \sigma\sqrt{\log(1/\rho)}+\varepsilon+\rho R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}}+\sigma\sqrt{\frac{d+\log(1/\delta)}{n}}$$

for oblivious covariates and

$$\rho^2\|\widetilde{w}-w^*\|_{\mathbf{\Sigma}_n} \lesssim \sigma\sqrt{\log(1/\rho)}+\varepsilon+\rho R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}}+(R\sigma)^{1/2}\rho\cdot\sqrt[4]{\frac{\log(1/\delta)}{n}}$$

for the more general case of adaptive covariates.

Proof. For any δ', we know that each of the ξ_t satisfy $|\xi_t| \lesssim \sigma\sqrt{\log(1/\delta')}$ individually with probability $1-\delta'$. If we treat indices t for which this does not hold as corruptions and take δ' to be $1/4-\eta/2$, then we can take the the corruption level in Theorem 6.6.1 to be

$1/4+\eta/2$ and the bound on the noise to be $\sigma\sqrt{\log\left(\frac{4}{1-2\eta}\right)}$. Note that for $\eta \in (1/4, 1/4)$, the ρ corresponding to the new corruption level $1/4+\eta/2$ is within a constant factor of ρ. Also note that $\sqrt{\log(1/\delta')} = O(\log(1/\rho))$. The result then follows by Theorem 6.6.1. □

We now state the full guarantee obtained by combining the above with the results of Section 6.5. For brevity, we will only state the subgaussian case:

Theorem 6.6.3. *Let $0 \le \eta < 1/2$, and define $\rho > 0$ by $\eta = \frac{1}{2+2\rho^2}$, and suppose n satisfies $n \gtrsim \log(\min(n,d)/\delta)$. If the noise $\{\xi_t\}$ is σ^2-subgaussian, then there is a* poly(n,d) *algorithm which takes as input $(x_1,y_1),\dots,(x_n,y_n)$ and, with probability at least $1-\delta$, outputs a vector w which satisfies*

$$\min(1,\rho^2)\|u^*-w\|_{\mathbf{\Sigma}_n} \lesssim c_{\delta,\eta,n}\eta\sigma + \eta^{1/2}\rho^2\varepsilon + \eta^{1/8}R^{1/2}(\sqrt{c_{\delta,\eta,n}\eta}\sigma+\varepsilon)^{1/2}\sqrt[8]{\frac{\log(\min(n,d)/\delta)}{n}}$$
$$+\eta^{1/4}R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}} + \min\left\{\sigma\sqrt{\frac{d+\log(2/\delta)}{n}}, (R\sigma)^{1/2}\sqrt[4]{\frac{\log(2/\delta)}{n}}\right\}$$

for oblivious covariates, where $c_{\delta,\eta,n}$ is defined in (6.24). In the more general case of adaptive covariates, w satisfies

$$\min(1,\rho^2)\cdot\|u^*-w\|_{\mathbf{\Sigma}_n} \lesssim c_{\delta,\eta,n}\eta\sigma + \eta^{1/2}\rho^2\varepsilon + \eta^{1/8}R^{1/2}(\sqrt{c_{\delta,\eta,n}\eta}\sigma+\varepsilon)^{1/2}\sqrt[8]{\frac{\log(\min(n,d)/\delta)}{n}}$$
$$+\eta^{1/4}R\sqrt[4]{\frac{\log(\min(n,d)/\delta)}{n}} + (R\sigma)^{1/2}\sqrt[4]{\frac{\log(2/\delta)}{n}}, \tag{6.28}$$

i.e. the same bound except the last term was changed. Recall that u^* here is the best norm-R linear predictor of the uncorrupted and unnoised data, that is,

$$u^* \triangleq \arg\min_{u:\|u\|\le R}\frac{1}{n}\sum_t (y_t^* - \langle u, x_t\rangle)^2.$$

Proof. If $0 \le \eta < 0.3$, apply Theorem 6.5.13, noting that the parameter β in that theorem is an absolute constant for this range of η. Otherwise, apply Theorem 6.6.2, noting that

$\eta = \Theta(1)$ in this case, and that $\|u^* - w\|_{\mathbf{\Sigma}_n}$ and $\|w^* - w\|_{\mathbf{\Sigma}_n}$ differ by $O(\varepsilon)$. □

Sum-of-Squares Program and Feasibility

```
Algorithm 17: SoSRegression(D)
  Input: Dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
  Output: Vector $\widetilde{w}$ for which $\|\widetilde{w} - w^*\|_{\mathbf{\Sigma}_n}$ is small (see Theorem 6.6.1)
  Let $\widetilde{\mathbb{E}}[\cdot]$ be the pseudoexpectation optimizing Program 6.2.
  return $\widetilde{\mathbb{E}}[w]$.
```

We will condition on the events of Lemma 6.5.11. Now consider the following set of polynomial constraints.

Program 3. *Let $\alpha > 0$ be a parameters to be tuned later. The program variables are $\{a_t\}_{t\in[n]}$ and w, and the constraints are*

1. *(Norm bound) $\sum_{i=1}^{d} w_i^2 \le R^2$.*

2. *(Booleanity) $a_t^2 = a_t$ for all $t \in [n]$.*

3. *(Large fraction of inliers) $\frac{1}{n}\sum_{t=1}^{n} a_t \ge 1 - \eta - \alpha$.*

4. *(Outliers sub-sample the empirical covariance[2])*

$$\frac{1}{n}\sum_{t=1}^{n}(1 - a_t)x_t x_t^\top \preceq \eta\mathbf{\Sigma}_n + \alpha \cdot Id\,.$$

The program objective is to minimize

$$\min \widetilde{\mathbb{E}}\left[\sum_{t=1}^{n} a_t(y_t - \langle w, x_t\rangle)^2\right]$$

over degree-4 SoS-pseudoexpectations satisfying the above constraints.

We first show that conditioned on the events of Lemma 6.5.11 holding, there always exists a feasible solution to the above polynomial system.

[2] One can use matrix inequalities in SoS: see e.g. Section 7.1 in [HL18].

Lemma 6.6.4 (Satisfiability)**.** *For any $\delta > 0$, if n satisfies the bound in (6.27), then for any sequence of $x_1, ..., x_n$ chosen during the process in Definition 6.4.1, we have that with probability at least $1 - \delta$ over the randomness of the Ber(η) coins generating $a_1^*, ..., a_n^*$ and over the randomness of $\xi_1, ..., \xi_n$, the choice of $a_t = a_t^*$ and*

$$v = \arg \min_{\|v\| \leq R} \sum_{t=1}^{n} a_t (y_t - \langle v, x_t \rangle)^2 \tag{6.29}$$

is a feasible solution to Program 3. As a consequence, for any $\|v\| \leq R$ the objective value of Program 3 is at most

$$\frac{1}{n} \sum_{t=1}^{n} a_t^* (y_t - \langle v, x_t \rangle)^2.$$

Proof. Clearly Constraints 1 and 2 are satisfied. Part 1 of Lemma 6.5.11 implies that Constraint 3 is satisfied with probability $1-\delta/3$. Part 3 of Lemma 6.5.11 implies that Constraint 4 is satisfied with probability at least $1 - \delta/3$. Finally, the first-order stationarity condition is satisfied because v is the optimizer of (6.29).

To get the consequence, we use that such an upper bound holds with v the minimizer of (6.29) by feasibility of (a^*, v), and then use the fact that it is the minimizer to extend to conclusion to all (not necessarily first-order stationary) v. □

Bounding Clean Square Loss

We now proceed to the sum-of-squares proof that the constraints of Program 3 imply a bound on the clean square loss achieved by w, under the degree-4 SoS proof system.

Let v^* be defined as

$$v^* \triangleq \arg \min_{v : \|v\| \leq R} \frac{1}{n} \sum_{t=1}^{n} a_t^* (y_t^* - \langle v, x_t \rangle)^2. \tag{6.30}$$

The following Lemma is needed only for the misspecified setting: if $\varepsilon = 0$ we will trivially have $v^* = w^*$. In the misspecified setting v^* will naturally appear in the analysis, instead of w^*, because it gives the optimal bounded norm linear function approximating the true regression function $x_t \mapsto \langle w^*, x_t \rangle + \varepsilon_t$. We define $\mathbf{\Sigma}_n' \triangleq \frac{1}{n} \sum a_t^* \cdot x_t x_t^\top$.

Lemma 6.6.5. *For v^* as defined above, we have $\|v^* - w^*\|^2_{\mathbf{\Sigma}'_n} = O(\varepsilon^2)$ and also, if we define*

$$\varepsilon'_t \triangleq y^*_t - \langle v^*, x_t\rangle,$$

then for all w with $\|w\| \le R$ we have:

$$\sum_{t=1}^n a^*_t \varepsilon'_t \langle w - v^*, x_t\rangle \le 0 \tag{6.31}$$

Proof. Since $\nabla_v (y^*_t - \langle v^*, x_t\rangle)^2 = -2(y^*_t - \langle v^*, x_t\rangle)x_t$, we see that the first order optimality condition for (6.30) implies for any w with $\|w\| \le R$ we have

$$\frac{-2}{n}\sum_{t=1}^n a^*_t \varepsilon'_t \langle w - v^*, x_t\rangle \ge 0$$

which gives (6.31).

It remains to upper bound $\|v^* - w^*\|^2_{\Sigma'}$. By writing it out, we see

$$\|v^* - w^*\|^2_{\Sigma^*_t} = \frac{1}{n}\sum_{t=1}^n a^*_t \langle v^* - w^*, x_t\rangle^2 = \frac{1}{n}\sum_{t=1}^n a^*_t (y^*_t - \varepsilon_t - \langle v^*, x_t\rangle)^2 \le \frac{2}{n}\sum_{t=1}^n a^*_t (\varepsilon^2_t + (\varepsilon'_t)^2) \le 2\varepsilon^2$$

where in the second-to-last step we used $(a+b)^2 \le 2a^2 + 2b^2$ and in the last step we used that v^* minimizes (6.30). □

We can now prove Theorem 6.6.1.

Proof of Theorem 6.6.1. Let $\widetilde{\mathbb{E}}[\cdot]$ be the pseudo-expectation optimizing the objective in Program 3, and define $\widetilde{w} \triangleq \widetilde{\mathbb{E}}[w]$. By part 3 of Lemma 6.5.11 and Constraint 1, we have that

$$(1-\eta)\|\widetilde{w} - w^*\|^2_{\mathbf{\Sigma}_n} \le \|\widetilde{w} - w^*\|^2_{\mathbf{\Sigma}'_n} + \alpha\|\widetilde{w} - w^*\|^2 \le \widetilde{\mathbb{E}}[\|w - v^*\|^2_{\mathbf{\Sigma}'_n}] + \alpha R^2 + 2\varepsilon^2,$$

where $\mathbf{\Sigma}'_n \triangleq \frac{1}{n}\sum a^*_t \cdot x_t x_t^\top$ and $\|\cdot\|_{\mathbf{\Sigma}'_n}$ is the induced norm, and in the last step we used the first part of Lemma 6.6.5, Lemma 1.3.43, and Constraint 1.

We can further bound

$$\begin{aligned}
&\widetilde{\mathbb{E}}[\|w - v^*\|^2_{\boldsymbol{\Sigma}'_n}] \\
&\quad= \frac{1}{n}\sum_{t=1}^{n} a_t^* \widetilde{\mathbb{E}}[\langle w - v^*, x_t\rangle^2] \\
&\quad= \frac{1}{n}\sum_{t=1}^{n} a_t^* \widetilde{\mathbb{E}}\big[(y_t - \langle w, x_t\rangle) - (y_t - \langle v^*, x_t\rangle)^2\big] \\
&\quad= \frac{1}{n}\sum_{t=1}^{n} a_t^* \left[\widetilde{\mathbb{E}}\big[(y_t - \langle w, x_t\rangle)^2\big] - (y_t - \langle v^*, x_t\rangle)^2\right] + \frac{2}{n}\sum_{t=1}^{n} a_t^*(y_t - \langle v^*, x_t\rangle) \cdot \left\langle \widetilde{\mathbb{E}}[w] - v^*, x_t\right\rangle \\
&\quad= \underbrace{\frac{1}{n}\sum_{t=1}^{n} a_t^* \left[\widetilde{\mathbb{E}}\big[(y_t - \langle w, x_t\rangle)^2\big] - (y_t - \langle v^*, x_t\rangle)^2\right]}_{\textcircled{1}} + \underbrace{\left\langle \widetilde{\mathbb{E}}[w] - v^*, \frac{2}{n}\sum_{t=1}^{n} a_t^*(\xi_t + \varepsilon'_t)x_t \right\rangle}_{\textcircled{2}}
\end{aligned}$$

where in the fourth step we used the identity $(a-b)^2 = a^2 - b^2 - 2b(a-b)$ and $\varepsilon'_t := y_t - \xi_t - \langle v^*, x_t\rangle$ as defined in Lemma 6.6.5.

Because of Lemma 6.6.5 and $\|\widetilde{\mathbb{E}}[w]\|^2 \le R^2$ from Constraint 1 we know that

$$\langle \widetilde{\mathbb{E}}[w] - v^*, \frac{2}{n}\sum_{t=1}^{n} a_t^* \varepsilon'_t x_t\rangle \le 0$$

so we can drop this term from $\textcircled{2}$. Then by part 2 of Lemma 6.5.11, together with Cauchy-Schwarz,

$$\textcircled{2} \le 2\sigma\left(\lambda\big\|\widetilde{\mathbb{E}}[w] - v^*\big\|_{\boldsymbol{\Sigma}'_n} + \lambda'\big\|\widetilde{\mathbb{E}}[w] - v^*\big\|\right) \le O\left(\big\|\widetilde{\mathbb{E}}[w] - v^*\big\|_{\boldsymbol{\Sigma}'_n}\sigma\lambda + R\sigma\lambda'\right).$$

It remains to upper bound $\textcircled{1}$, and this is the bulk of the analysis. Concretely, we need to show that the constraints of the program SoS-imply an upper bound on the quantity $\frac{1}{n}\sum_{t=1}^{n} a_t^*(y_t - \langle w, x_t\rangle)^2 - \frac{1}{n}\sum_{t=1}^{n} a_t^*(y_t - \langle w^*, x_t\rangle^2)$ of $c\|w - v^*\|^2_{\boldsymbol{\Sigma}'_n} + O(\cdot)$ with $c \in [0, 1)$, so that we can solve for an upper bound on $\|w - v^*\|^2_{\boldsymbol{\Sigma}'_n}$. We do so in Lemma 6.6.6 below and get $c = \frac{(1+\rho^2)\overline{\eta}}{1-\overline{\eta}}$. Choosing ρ to be the solution to $\overline{\eta} = \frac{1}{2+2\rho^2}$ and observing that

$$\frac{1}{1-c} = \frac{1-\overline{\eta}}{1-(2+\rho^2)\overline{\eta}} = \frac{1+2\rho^2}{\rho^2} = 2 + 1/\rho^2,$$

we get that

$$\left\|\widetilde{\mathbb{E}}[w] - v^*\right\|_{\mathbf{\Sigma}'_n}^2 \le O(1/\rho^2) \cdot \left(\left\|\widetilde{\mathbb{E}}[w] - v^*\right\|_{\mathbf{\Sigma}'_n} \sigma\lambda + \mathcal{E}\right)$$

for $\mathcal{E} \triangleq R\sigma\lambda' + \frac{\sigma^2+\varepsilon^2}{\rho^2} + \alpha R^2$. We do case analysis based on which of the two terms on the right-hand side dominates:

1. If the former dominates, then the bound simplifies to

$$\left\|\widetilde{\mathbb{E}}[w] - v^*\right\|_{\mathbf{\Sigma}'_n} \lesssim \sigma\lambda/\rho^2$$

2. Otherwise, if $\mathcal{E}$ dominates, then after taking a square root, the bound can be rewritten as

$$\left\|\widetilde{\mathbb{E}}[w] - v^*\right\|_{\mathbf{\Sigma}'_n} \lesssim \rho^{-1} \cdot \left((R\sigma\lambda')^{1/2} + \frac{\sigma+\varepsilon}{\rho} + \alpha^{1/2}R\right)$$

In either case, we conclude that

$$\left\|\widetilde{\mathbb{E}}[w] - v^*\right\|_{\mathbf{\Sigma}'_n} \lesssim \sigma\lambda/\rho^2 + \rho^{-1} \cdot \left((R\sigma\lambda')^{1/2} + \frac{\sigma+\varepsilon}{\rho} + \alpha^{1/2}R\right).$$

If the covariates are adaptively chosen, we get

$$\left\|\widetilde{\mathbb{E}}[w] - v^*\right\|_{\mathbf{\Sigma}'_n} \lesssim \frac{R^{1/2}\sigma^{1/2}}{\rho} \cdot \sqrt[4]{\frac{\log(1/\delta)}{n}} + \frac{\sigma+\varepsilon}{\rho^2} + \frac{\alpha^{1/2}R}{\rho}.$$

If the covariates are obliviously chosen, then we could also obtain

$$\left\|\widetilde{\mathbb{E}}[w] - v^*\right\|_{\mathbf{\Sigma}'_n} \lesssim \frac{\sigma}{\rho^2} \cdot \sqrt{\frac{d+\log(1/\delta)}{n}} + \frac{\sigma+\varepsilon}{\rho^2} + \frac{\alpha^{1/2}R}{\rho}.$$

Plugging in $\alpha = \Theta\left(\sqrt{\eta \log(\min(n,d)/\delta) n}\right)$ as in Section 6.5 completes the proof. □

Lemma 6.6.6. *Conditioned on the four parts of Lemma 6.5.11 holding, we have for any $\rho \in (0,1]$ that*

$$\widetilde{\mathbb{E}}\left[\frac{1}{n}\sum_{t=1}^{n} a_t^*(y_t - \langle w, x_t\rangle)^2\right] \le \frac{1}{n}\sum_{t=1}^{n} a_t^*(y_t - \langle v^*, x_t\rangle)^2 + \frac{(1+2\rho^2)\eta}{1-\eta}\|v^* - w\|^2_{\boldsymbol{\Sigma}'_n} + \\ O\left(\frac{\sigma^2+\varepsilon^2}{\rho^2} + \alpha R^2\right) \quad (6.32)$$

as long as $\widetilde{\mathbb{E}}[\cdot]$ is a SoS degree-4 pseudoexpectation satisfying the constraints of the program.

Proof. Let $\circledast$ denote the quantity inside the pseudoexpectation on the left-hand side of (6.32). Then in the SoS degree-4 proof system we can show the following bound

$$\begin{aligned}
\circledast &= \frac{1}{n}\sum_{t=1}^{n} a_t^* a_t(y_t - \langle w, x_t\rangle)^2 + \frac{1}{n}\sum_{t=1}^{n} a_t^*(1-a_t)(y_t - \langle w, x_t\rangle)^2 \\
&\le \frac{1}{n}\sum_{t=1}^{n} a_t(y_t - \langle w, x_t\rangle)^2 + \frac{1}{n}\sum_{t=1}^{n} a_t^*(1-a_t)(y_t - \langle w, x_t\rangle)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n} a_t(y_t - \langle w, x_t\rangle)^2 + \frac{1}{n}\sum_{t=1}^{n} a_t^*(1-a_t)(y_t - \varepsilon'_t - \langle v^*, x_t\rangle + \langle v^* - w, x_t\rangle + \varepsilon'_t)^2 \\
&\le \frac{1}{n}\sum_{t=1}^{n} a_t(y_t - \langle w, x_t\rangle)^2 + \frac{2+1/\rho^2}{n}\sum_{t=1}^{n} a_t^*(1-a_t)(y_t - \varepsilon'_t - \langle v^*, x_t\rangle)^2 \\
&\quad + \frac{1+2\rho^2}{n}\sum_{t=1}^{n} a_t^*(1-a_t)\langle v^* - w, x_t\rangle^2 + \frac{2+1/\rho^2}{n}\sum_{t=1}^{n} a_t^*(1-a_t)(\varepsilon'_t)^2 \\
&\le \frac{1}{n}\sum_{t=1}^{n} a_t(y_t - \langle w, x_t\rangle)^2 + \frac{2+1/\rho^2}{n}\sum_{t=1}^{n} a_t^*(1-a_t)\xi_t^2 + \\
&\qquad \frac{1+2\rho^2}{n}\sum_{t=1}^{n} a_t^*(1-a_t)\langle v^* - w, x_t\rangle^2 + (2+1/\rho^2)\varepsilon^2
\end{aligned}$$

where in the second step we use Constraint 2 to get $a_t^* a_t \le a_t$, in the fourth step we use the SOS Cauchy-Schwartz inequality to show $(a+b+c)^2 = (\rho a/\rho + b + \rho c/\rho)^2 \le (1+2\rho^2)(a^2/\rho^2 + b^2 + c^2/\rho^2)$, and in the fifth step we used that $\sum_{t=1}^n a_t^*(\varepsilon'_t)^2 \le \sum_{t=1}^n a_t^*\varepsilon_t^2 \le \varepsilon^2$ by construction (see (6.30)).

Therefore, we can upper bound $\widetilde{\mathbb{E}}[\textcircled{*}]$ by

$$\underbrace{\widetilde{\mathbb{E}}\left[\frac{1}{n}\sum_{t=1}^{n} a_t(y_t - \langle w, x_t\rangle)^2\right]}_{\textcircled{\scriptsize I}} + \underbrace{\frac{2+1/\rho^2}{n}\sum_{t=1}^{n} a_t^*(1-\widetilde{\mathbb{E}}[a_t])\xi_t^2}_{\textcircled{\scriptsize II}} + \underbrace{\widetilde{\mathbb{E}}\left[\frac{1+2\rho^2}{n}\sum_{t=1}^{n} a_t^*(1-a_t)\langle v^* - w, x_t\rangle^2\right]}_{\textcircled{\scriptsize III}} + (2+1/\rho^2)\varepsilon^2.$$

From the last part of Lemma 6.6.4, we know $\textcircled{\scriptsize I} \leq \frac{1}{n}\sum_{t=1}^{n} a_t^*(y_t - \langle v^*, x_t\rangle)^2$. And as we are in the bounded noise setting, we can upper bound $\textcircled{\scriptsize II}$ by $(2+1/\rho^2)\cdot\sigma^2(\eta+\alpha) \leq O(\sigma^2/\rho^2)$, where in the last inequality we used that η is upper and lower bounded by absolute constants by assumption.

Finally, to bound $\textcircled{\scriptsize III}$, we can finally apply Constraint 4. We get that

$$\begin{aligned}
\frac{1+2\rho^2}{n}\sum_{t=1}^{n} a_t^*(1-a_t)\langle v^* - w, x_t\rangle^2 &\leq \frac{1+2\rho^2}{n}\sum_{t=1}^{n}(1-a_t)\langle v^* - w, x_t\rangle^2 \\
&\leq \frac{(1+2\rho^2)\eta}{n}\sum_{t=1}^{n}\langle v^* - w, x_t\rangle^2 + 3\alpha\|v^* - w\|_2^2 \\
&\leq \frac{(1+2\rho^2)\eta}{(1-\eta)n}\sum_{t=1}^{n} a_t^*\langle v^* - w, x_t\rangle^2 + \frac{3\eta\alpha R^2}{1-\eta} + 3\alpha R^2 \\
&= \frac{(1+2\rho^2)\eta}{(1-\eta)}\|v^* - w\|_{\mathbf{\Sigma}_n'}^2 + O(\alpha R^2),
\end{aligned}$$

where the second step follows by Constraint 4 and $\rho \leq 1$, the third step follows by part 3 of Lemma 6.5.11 which we are conditioning on in this section, and the fourth step uses the definition of $\mathbf{\Sigma}_n'$ together with the assumption that η is at least some absolute constant. □

6.7 Online Regression

6.7.1 Cutting Plane Algorithm

In this section we leverage the guarantees of Section 6.5 to design an efficient algorithm for Huber-contaminated online regression. For brevity, in this section we restrict our attention to the case of sub-Gaussian noise, though our techniques extend easily to handle k-hypercontractive noise.

The basic trick we use is to combine the offline regression oracle with a cutting plane method, so that we can keep efficiently cutting down the space of linear predictors until we find one near w^*. Essentially, the algorithm collects a large batch of samples, compares it's current performance on this batch to the optimal robust regression result in hindsight (estimated by SCRAM), and if it finds its performance is poor it cuts out a large set of possible predictors and updates to use a new predictor.

The algorithm, which we will refer to as AMCutter, can be based upon any central cutting-plane optimization method like ellipsoid or Vaidya's algorithm; here we use Vaidya's algorithm since it is oracle-efficient. More specifically, we recall the following guarantee for Vaidya's algorithm:

Theorem 6.7.1 ([Vai89], see e.g. Section 2.3 of [Bub14]). *Suppose that $\mathcal{K}$ is an (unknown) convex body in $\mathbb{R}^d$ which contains a Euclidean ball of radius $r > 0$ and is contained in a Euclidean ball centered at the origin of radius $R > 0$. There exists an algorithm which, given access to a separation oracle for $\mathcal{K}$, finds a point $x \in \mathcal{K}$, runs in time poly$(\log(R/r), d)$, and makes $O(d \log(Rd/r))$ calls to the separation oracle.*

Now we describe the algorithm. C_0 and N_0 are constants to be determined later. SeparationOracle (see Algorithm 18) implements the separation oracle (which is also where most of the interaction with Nature occurs). Here the input w lies in $\mathcal{W} = \{w : \|w\| \le R\}$ and Nature's inputs are x_t with $\|x_t\| \le 1$. Finally, we note that if SeparationOracle gets to the final round T of the online regression problem, then it may not return to Vaidya's algorithm (so step 2 of AMCutter is never reached), but as we will see, even if this happens the algorithm still achieves the correct regret bound.

```
Algorithm 18: SeparationOracle($w, x_t, C_0, D$)
Input: Vector $w \in \mathcal{W}$
Output: Separating hyperplane between $w$ and the target region
        $\{w' : \|w' - w^*\| \leq r\}$, if $w$ lies outside
$D \leftarrow \emptyset$.
for each new point $x_t$ input by Nature do
    Predict $\widehat{y}_t = \langle w, x_t \rangle$ and observe $y_t$.
    Append $(x_t, y_t)$ to $D$.
    $v_t \leftarrow \mathrm{SCRAM}(D)$.
    $\mathbf{\Sigma}_t \leftarrow \frac{1}{|D|} \sum_{(x_t,y_t)\in D} x_t x_t^\top$. Define $\varphi_t(u) \triangleq \|u - v_t\|^2_{\mathbf{\Sigma}_t}$.
    if $|D| \geq N_0$ and $\varphi_t(w) \geq C_0$ then
        // intersect current feasible region with $\{u : \langle u - w, \nabla\varphi_t(w)\rangle < 0\}$
        return separating hyperplane given by $\nabla\varphi_t(w)$.
```

```
Algorithm 19: AMCutter($r, R, N_0, C_0, T$)
Input: Radius $r$ of target ball around $w^*$, parameter $R$ from Assumption 5,
       parameters $N_0, C_0$ to be tuned, number of rounds $T$
Output: Sequence of predictions $\widehat{y}_1, \ldots, \widehat{y}_T$
Let $w$ be the output of running Vaidya's algorithm [Vai89] with
  SeparationOracle defined above and parameters $r, R$, and let $\widehat{y}_1, ..., \widehat{y}_{t_1}$ be the
  predictions made in the course of running SeparationOracle.
for $t_1 + 1 \leq t \leq T$ do
    Given new point $x_t$ input by Nature, predict $\widehat{y}_t = \langle w, x_t \rangle$.
return $\widehat{y}_1, \ldots, \widehat{y}_T$.
```

As far as the choice of constants, based on (6.28) and Theorem 6.6.3 we will leave N_0 to be optimized later and take

$$C_0 \triangleq 4Rr + \max(1, 1/\rho^4) \cdot O\left(c^2_{\delta/T,\eta,N_0}\eta^2\sigma^2 + \rho^4\varepsilon^2 + \eta^{1/4}R(\sqrt{c_{\delta/T,\eta,N_0}}\eta\sigma + \varepsilon)\sqrt[4]{\frac{\log(T/\delta)}{N_0}} \right.$$
$$\left. + \eta^{1/2}R^2\sqrt{\frac{\log(T/\delta)}{N_0}} + R\sigma\sqrt{\frac{\log(T/\delta)}{N_0}} \right),$$

where $\delta > 0$ is the desired overall probability of success. With this choice of parameters we can guarantee with probability at least $1 - \delta$:

1. At every step where $|D| \geq N_0$ in SeparationOracle, the guarantee (6.28) is satisfied by the vector v_t output by SCRAM, by applying Theorem 6.6.3 and the union bound over all rounds. In particular, by triangle inequality, we have $\|w^* - v_t\|^2_{\mathbf{\Sigma}_n} \leq C_0 - 4Rr$

2. If w lies outside the ball of radius r around w^*, the result of SeparationOracle is a valid separating hyperplane between w and the ball. By convexity of φ, to see that the ball of radius r around w^* is never cut, we just need to show that all w' with $\|w' - w^*\| \leq r$ satisfy $\varphi_t(w') \leq C_0$. For w^* we have the stronger guarantee $\varphi_t(w^*) \lesssim C_0 - 4Rr$, just from the guarantee of step 1. For other w' in the ball of radius r, we deduce the claim by triangle inequality from the guarantee for w^*, using that

   $$\varphi_t(w') - \varphi_t(w^*) \leq \langle \nabla\varphi_t(w'), w' - w^* \rangle = 2\langle \mathbf{\Sigma}_t(w' - v_t), w' - w^* \rangle \leq 4R\|w' - w^*\| \leq 4Rr$$

   where the first inequality is by convexity, and the second inequality uses that $\|\hat{\Sigma}_t\| \leq 1$ and that the diameter of $\mathcal{W}$ is at most $2R$.

Recall that the separation oracle can only be called $I = O(d\log(R/r))$ many times, since this is the oracle complexity guarantee from Theorem 6.7.1: after this many rounds the algorithm is guaranteed to return or query a point in the ball of radius r around w^*. Let D_i be the collected dataset D built during the i-th invocation of the oracle. Since we know by the triangle inequality and AM-GM that

$$\|w - w^*\|_{\mathbf{\Sigma}_t}^2 \leq 2\|w - v_t\|_{\mathbf{\Sigma}_t}^2 + 2\|v_t - w^*\|_{\mathbf{\Sigma}_t}^2$$

it follows that after $|D_i|$ gets to size N_0 and up to the step before returning a hyperplane, we are guaranteed that $\|w - w^*\|_{\mathbf{\Sigma}_t}^2 \leq 4C_0$. For all of the steps before $|D_i|$ gets to size N_0, the error incurred per step is trivially upper bounded by $4R^2$. It follows that the regret incurred per call of the separation is upper bounded by $\max\{4N_0R^2, 4|D_i|C_0 + 4R^2\}$. Hence, the total regret incurred in step 1 of AMCutter is upper bounded by

$$\sum_{i=1}^{I}(4N_0R^2 + 4|D_i|C_0) \leq 4N_0IR^2 + 4C_0T = O\left(N_0dR^2\log(R/r) + C_0T\right) \tag{6.33}$$

using that the total number of oracle calls is $I = O(d\log(R/r))$, and $\sum_i |D_i| \leq T$. If t_1 is the time step at which the algorithm enters step 2, then the total regret in step 2 of AMCutter is upper bounded by

$$\sum_{t=t_1}^{T}(\langle w^*, x_t\rangle + \varepsilon_t - \langle w, x_t\rangle)^2 \leq \sum_{t=t_1}^{T}(r + |\varepsilon_t|)^2 \leq 2T(r^2 + \varepsilon^2) \tag{6.34}$$

where in the last step we used the basic inequality $(a+b)^2 \leq 2a^2 + 2b^2$. In particular, the leading term in the regret is $O(k\sigma^2\eta^{2-2/k}T)$ as expected. We formalize this in the following Theorem.

Theorem 6.7.2. *For the Huber-Contaminated Online Regression problem with $\eta \leq \overline{\eta} < 1/2$ and $\overline{\eta} = \frac{1}{2+2\rho^2}$, Algorithm* AMCutter *with parameters R and $r \triangleq 1/T$ satisfies the following regret guarantee:*

$$\sum_{t=1}^{T}(y_t^* - \hat{y}_t)^2 \lesssim \left(\eta^2\log(1/\eta)\sigma^2\rho^{-4} + \varepsilon^2\right)T + \eta^{1/4}R\rho^{-4}\left(\eta^{1/2}\sqrt[4]{\log(1/\eta)}\cdot\sigma + \varepsilon\right)\sqrt[4]{\log T}\cdot d^{1/6}T^{5/6}$$
$$+ \left(\eta^{1/2}R^2 + R\sigma\right)\cdot\rho^{-4}d^{1/3}T^{2/3}\sqrt{\log T} + d^{1/3}R^2\log(RT)T^{2/3} \tag{6.35}$$

with probability $1 - 1/\mathrm{poly}(T)$ over the randomness of the coin flips. In particular, for *sufficiently large T, this quantity is dominated by $(\eta^2 \log(1/\eta)\sigma^2\rho^{-4} + \varepsilon^2)T$.*

Proof. From the above (6.33) and (6.34), we see that the total regret is upper bounded by

$$O\left(N_0 d R^2 \log(R/r) + C_0 T\right) + 2T(r^2 + \varepsilon^2).$$

so by taking $N_0 = d^{-2/3}T^{2/3}$ and $r = 1/T$, we get the claimed regret bound upon noting that $c_{1/10T,\eta,d^{-2/3}T^{2/3}} = O(\sqrt{\log(1/\eta)})$. □

6.7.2 Gradient Descent Algorithm

For the high-dimensional setting, cutting planes don't work because their guarantees are dimension-dependent. Fortunately, we can fix this by using gradient descent instead, see Algorithm 20. We recall the following guarantee for online gradient descent from [Zin03].

```
Algorithm 20: AM-GD($R, N_0, C_1, \gamma, T$)
  Input: Parameter $R$ from Assumption 5, number of rounds $T$, parameters
         $r, N_0, C_1, \gamma$ to be tuned
  Output: Sequence of predictions $\widehat{y}_1, \ldots, \widehat{y}_T$ (via interaction with Nature)
  Let $w_1 = 0$.
  while there are more inputs do
      Let $g_s$ be the output of SeparationOracle run with parameters $r \triangleq 0, R, C_1$
        and input $w_s$
      Let $w_{s+1} = w_s - \frac{\gamma}{\sqrt{T}} g_s$.
      Set $s \leftarrow s + 1$.
```

Theorem 6.7.3 ([Zin03, Haz19])**.** *Suppose that $f_1, \ldots, f_T$ is a sequence of convex functions such that $\|\nabla f_t(w)\| \le G$ for any w with $\|w\| \le R$. Let $w_1 = 0$ and suppose that*

$$w_{t+1} \triangleq \Pi_R\left(w_t - \frac{2R}{G\sqrt{T}}\nabla f_t(w_t)\right)$$

where $\Pi_R(x) \triangleq \frac{x}{\max(R,\|x\|)}$ is the projection onto the Euclidean ball of norm R. Then for any w^* with $\|w^*\| \le R$,

$$\sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w^*) \le \sum_{t=1}^{T} \langle \nabla f_t(w_t), w_t - w^* \rangle \le 3RG\sqrt{T}.$$

We now discuss parameter selection: we define

$$C_0 \triangleq \max(1, 1/\rho^4) \cdot O\left(c^2_{\delta/T,\eta,N_0}\eta^2\sigma^2 + \eta\rho^4\varepsilon^2 + \eta^{1/4}R(\sqrt{c_{\delta/T,\eta,N_0}\eta}\sigma + \varepsilon)\sqrt[4]{\frac{\log(N_0T/\delta)}{N_0}} \right.$$
$$\left. + \eta^{1/2}R^2\sqrt{\frac{\log(N_0T/\delta)}{N_0}} + R\sigma\sqrt{\frac{\log(2T/\delta)}{N_0}} \right)$$

where $\delta > 0$ is the overall acceptable probability of failure, based upon the right-hand side of (6.28) and take $C_1 \triangleq 2C_0$.

Theorem 6.7.4. *For the Huber-Contaminated Online Regression problem with $\eta \le \overline{\eta} < 1/2$ and $\overline{\eta} = \frac{1}{2+2\rho^2}$, Algorithm* AM-GD *with parameters R and $\gamma = \Theta(1)$ satisfies the following regret guarantee:*

$$\sum_{t=1}^{T}(y^*_t - \hat{y}_t)^2 \lesssim \left(\eta^2\log(1/\eta)\sigma^2\rho^{-4} + \varepsilon^2\right)T + \eta^{1/4}R\rho^{-4}\left(\eta^{1/2}\sqrt[4]{\log(1/\eta)}\cdot\sigma + \varepsilon\right)\sqrt[4]{\log T}\cdot T^{9/10}$$
$$+ \left(\eta^{1/2}R^2\rho^{-4}\sqrt{\log T} + R\rho^{-4}\sigma\sqrt{\log T} + R^2/\eta\right)\cdot T^{4/5} \quad (6.36)$$

with probability $1 - 1/\mathrm{poly}(T)$ over the randomness of the coin flips. In particular, for sufficiently large T, this quantity is dominated by $(\eta\varepsilon^2 + k\sigma^2\eta^{2-2/k})T$.

Note that in (6.36) there is a term $R^2T^{4/5}/\eta$ which increases as $\eta \to 0$. As discussed previously, for very small contamination rate η one can simply apply the above Theorem with slightly larger η to get meaningful bounds.

Proof. As in the proof of Theorem 6.7.2, we first bound the regret incurred in a single call of SeparationOracle by $4N_0R^2 + 8|D_i|C_0$ where D_i is the dataset D collected in call i. It follows then that if V is the total number of calls made to SeparationOracle then the total clean regret is upper bounded by $O(N_0R^2V + TC_0)$ where we used that $\sum_i |D_i| \le T$. On the other hand, we know from Theorem 6.7.3 that if we define φ_i to be the function whose gradient is returned at the end of Algorithm SeparationOracle, then

$$C_0V = (C_1 - C_0)V \le \sum_{s=1}^{V}(\varphi_i(w_s) - \varphi_i(w^*)) \le 6R^2\sqrt{V}$$

since $\|\nabla\varphi_i(w')\| \le \|\mathbf{\Sigma}_t(w' - v_t)\| \le 2R$ and using the corresponding choice of γ. Therefore $V = O(R^4/C_0^2)$. Hence the clean regret is upper bounded by $O(N_0 R^6/C_0^2 + TC_0)$.

Finally, it remains to choose N_0. At this point the optimal choice for N_0 is given by equalizing N_0 and the terms involving N_0 but not η in $C_0^3 T/R^6$. Since the leading order term in C_0 of this kind is of order $N_0^{-1/2}$ we can roughly minimize by taking $N_0 = T^{2/5}$. In this case,

$$\frac{N_0 R^6}{C_0^2} \lesssim \frac{N_0 R^6}{\max(1, 1/\rho^4)\cdot \eta R^4 \log(T)/N_0} \le \max(1, 1/\rho^4)\cdot (R^2/\eta)\cdot T^{4/5},$$

so the claimed bound follows. □

6.8 Putting Everything Together

In this section we record consequences of applying our results on Huber-contaminated online regression to the reduction of [FR20] (see Appendix 6.10).

The first consequence is the following pseudo-regret/regret bound for Huber-contaminated contextual bandits in the finite-dimensional case.

Theorem 6.8.1 (Main, formal version of Theorem 6.1.6)**.** *For the Huber-Contaminated Contextual Bandits problem with contamination rate $0 \le \eta < 1/2$ and corresponding parameter ρ given by $\eta = \frac{1}{2+2\rho^2}$, σ^2-subgaussian noise $\{\xi_t\}$, misspecification rate ε, range parameter R, noise parameter σ, action space of size K, and d-dimensional contexts, then there is a* $\mathrm{poly}(n,d)$*-time algorithm which achieves clean pseudo-regret $\widetilde{\mathrm{Reg}}_{\mathsf{HCB}}(T)$ at most*

$$O(\sqrt{K})\left((\eta\sqrt{\log(1/\eta)}\sigma\rho^{-2}+\varepsilon)T + \eta^{1/8}R^{1/2}\rho^{-2}\left(\eta^{1/4}\sqrt[8]{\log(1/\eta)}\cdot\sigma^{1/2}+\varepsilon^{1/2}\right)\sqrt[8]{\log T}\cdot d^{1/12}T^{11/12} + \left(\eta^{1/4}R + R^{1/2}\sigma^{1/2}\right)\cdot\rho^{-2}d^{1/6}T^{5/6}\sqrt{\log T} + d^{1/6}R\sqrt{\log(RT)}T^{5/6}\right).$$

In particular, for sufficiently large T, this quantity is dominated by $\left(\eta\sqrt{\log(1/\eta)}\sigma\rho^{-2}+\varepsilon\right)\sqrt{K}T$.

In the special case where $\varepsilon = 0$, there is a $\mathrm{poly}(n,d)$*-time algorithm which achieves clean regret* $\mathrm{Reg}_{\mathsf{HCB}}(T)$ *at most*

$$O(\sqrt{K})\left(\eta\sqrt{\log(1/\eta)}\sigma\rho^{-2}T+\eta^{1/8}R^{1/2}\rho^{-2}\left(\eta^{1/4}\sqrt[8]{\log(1/\eta)}\cdot\sigma^{1/2}\right)\sqrt[8]{\log T}\cdot d^{1/12}T^{11/12}\right.$$
$$\left.+\left(\eta^{1/4}R+R^{1/2}\sigma^{1/2}\right)\cdot\rho^{-2}d^{1/6}T^{5/6}\sqrt{\log T}+d^{1/6}R\sqrt{\log(RT)}T^{5/6}\right).$$

with probability $1-1/\mathrm{poly}(T)$*. For sufficiently large* T*, this is dominated by* $\eta\sqrt{\log(1/\eta)}\sigma\rho^{-2}\sqrt{K}T$*.*

Proof. For the first part of the theorem, we can apply Theorem 6.7.2 with failure probability $T^{-1/3}$ to get that the clean square loss regret incurred by AMCutter is given by (6.35) with probability at least $1-T$ and is otherwise upper bounded by R^2T. So the expectation of this quantity is at most the quantity in (6.35) plus $R^2T^{1/3}$, which is dominated by the $d^{1/3}R^2\log(RT)T^{2/3}$ term in (6.35). The result then follows from applying the clean pseudo-regret bound of Theorem 6.10.1 and using the elementary fact that for positive numbers $\{a_i\}_{i\in[s]}$, $\left(\sum_{i=1}^{s}a_i\right)^{1/2}\leq\sum_{i=1}^{s}\sqrt{a_i}$.

For the second part of the theorem, we can directly apply the high-probability guarantee Theorem 6.7.2 together with the high-probability guarantee of Theorem 6.10.3 and a union bound. □

Theorem 6.8.2 (High-dimensional variant of Theorem 6.8.1)**.** *Let* $\eta,\rho,\varepsilon,R,\sigma,K$ *be the same as in Theorem 6.8.1, but now we make no assumptions on the dimension of the context space* $\mathcal{X}$*. There exists an algorithm which runs in polynomial time and achieves clean pseudo-regret* $\widetilde{\mathrm{Reg}}_{\mathsf{HCB}}(T)$ *at most*

$$O(\sqrt{K})\cdot\left(\left(\eta\sqrt{\log(1/\eta)}\sigma\rho^{-2}+\varepsilon\right)T+\eta^{1/8}R^{1/2}\rho^{-2}\left(\eta^{1/4}\sqrt[8]{\log(1/\eta)}\cdot\sigma^{1/2}+\varepsilon^{1/2}\right)\sqrt[8]{\log T}\cdot T^{19/20}\right.$$
$$\left.+\left(\eta^{1/4}R\rho^{-2}\sqrt[4]{\log T}+R^{1/2}\rho^{-2}\sigma^{1/2}\sqrt[4]{\log T}+R/\sqrt{\eta}\right)\cdot T^{9/10}\right).$$

In particular, for sufficiently large T*, this quantity is dominated by* $\left(\eta\sqrt{\log(1/\eta)}\sigma\rho^{-2}+\varepsilon\right)\sqrt{K}T$*. When* $\varepsilon=0$*, we can similarly achieve a bound on the clean regret* $\mathrm{Reg}_{\mathsf{HCB}}(T)$ *with high probability.*

Proof. The proof is identical to Theorem 6.8.1, except that we replaced the use of Theorem 6.7.2 by Theorem 6.7.4 and AMCutter by AM-GD. □

6.9 Lower Bound Against Convex Surrogates

We exhibit an $\Omega(\eta^3 \sigma R)$ lower bound against regression using convex losses. This lower bound captures natural approaches like Huber regression, L_1/LAD regression, and OLS. By rescaling, we can assume $\sigma = 1$ without loss of generality, which we do in the statement of the result below; also, just for this example we scale (without loss of generality) so that $\|w^*\| \leq 1$ and $\|x_t\| \leq R$, because this makes the equations slightly cleaner.

Theorem 6.9.1. *For any convex loss $h(\cdot)$, there exists a distribution over covariates $x \sim \mathcal{D}_x$ with support in $[-R, R]$ and true regressor $\ell \in [-1, 1]$ such that the following is true. Let $y \sim \ell \cdot x + \zeta$ with noise $\zeta \sim \mathcal{N}(0, 1)$, and let $\mathcal{C}$ denote the joint distribution over (x, y). Furthermore, let $\widehat{y}$ denote the Huber contaminated labels drawn $y \sim (1 - \eta)(\ell \cdot x + \zeta) + \eta \mathcal{Q}$ where $\mathcal{Q}$ is an arbitrary distribution with support in $[-R, R]$ for $R \geq \frac{1}{\eta}$ and $\eta \in [0, \frac{1}{2})$. Let $\mathcal{H}$ be the joint distribution of the contaminated data $(x, \widehat{y})$. For any $b \in [0, 1]$, let $w :=$ argmin$_{\ell \in [-b,b]} \mathbb{E}_{(x,\widehat{y}) \sim \mathcal{H}}[h(y - \ell \cdot x)]$ be the minimizer of the loss on contaminated data. Then the clean square loss of w is lower bounded as $\mathbb{E}_{(x,y) \sim \mathcal{C}}[(y - w \cdot x)^2] \geq \min\left(\frac{\eta^3 R}{40}, \frac{(1-b)^2 R^2}{2}\right)$.*

Proof. First, we consider the case where the constraint parameter b is less than 1. In this case, we can just consider a simple clean example, e.g. the covariate distribution $x = 0$ with probability 1/2 and $x = R$ with probability 1/2, and take $\ell = 1$. If $b < 1$ then the best predictor in $[-b, b]$ makes squared loss at least $(1 - b)^2 R^2 / 2$, which proves the second lower bound.

We now consider the more interesting case where $b = 1$. Our hard instance is constructed as follows. Let $\mathcal{D}_x \triangleq m_1 \delta(1) + (1 - m_1)\delta(-R)$ where $\delta(\cdot)$ is the dirac delta and $m_1 = 1 - \frac{\eta}{10R}$. Let the true regressor $\ell = 0$ so that the uncorrupted $y \sim \mathcal{N}(0, 1)$ for all $x \in [-R, R]$. Let the corrupted labels be $\widehat{y}$ defined as follows

$$\widehat{y} = \begin{cases} (1 - \eta)\mathcal{N}(0, 1) + \eta\delta(R + 1) & x = 1 \\ \mathcal{N}(0, 1) & x = -R \end{cases}$$

Let $h'(\cdot)$ be the right derivative of $h(\cdot)$, which is well defined because every convex function on an open convex domain is semi-differentiable. Let $g(v) \triangleq -\mathbb{E}_{y \sim \mathcal{N}(0,1)}[h'(y - v)]$. By convexity of $h(\cdot)$ we have the right derivative evaluated at w is greater than or equal to zero.

$$\lim_{\varepsilon\to 0}\frac{\mathbb{E}_{(x,y)\sim\mathcal{H}}[h(y-(v+\varepsilon)\cdot x)]-\mathbb{E}_{(x,y)\sim\mathcal{H}}[h(y-v\cdot x)]}{\varepsilon}\Big|_{v=w}$$
$$=(1-\eta)m_1\cdot g(w)-h'(R+1-w)\eta\cdot m_1+(1-m_1)Rg(-Rw)\geq 0$$

Rearranging we obtain

$$g(w)\geq\frac{h'(R+1-w)\eta\cdot m_1-(1-m_1)Rg(-Rw)}{(1-\eta)m_1}\tag{6.37}$$

Let $g^{-1}(\cdot)$ denote the left inverse of $g(\cdot)$. Note that $h(\cdot)$ is convex implies $-h'(\cdot)$ is monotonically decreasing implies $g(\cdot)$ is monotonically increasing implies $g^{-1}(\cdot)$ is monotonically increasing. Thus, applying $g^{-1}(\cdot)$ to both sides of (6.37) we obtain

$$w\geq g^{-1}\big(\frac{h'(R+1-w)\eta\cdot m_1-(1-m_1)Rg(-Rw)}{(1-\eta)m_1}\big)\tag{6.38}$$

To lower bound w it suffices to lower bound the argument of $g^{-1}(\cdot)$. We obtain,

$$\frac{h'(R+1-w)\eta\cdot m_1-(1-m_1)R\cdot g(-Rw)}{(1-\eta)m_1}\geq\frac{h'(R)\eta\cdot m_1+h'(R)R(1-m_1)}{(1-\eta)m_1}$$

Where we lower bounded the first term in the numerator using the fact that $h'(\cdot)$ is monotonically increasing and $w\in[-1,1]$ to conclude $h'(R+1-w)\geq h'(R)$. We lower bounded the second term in the numerator using the fact that $g(\cdot)$ is monotonically increasing and that $h'(R)\geq\max_{[-R,R]}|h'(x)|$ (monotonicity of $h'(\cdot)$) to conclude $g(-Rw)\geq g(-R)\geq -h'(R)$. Further lower bounding, we obtain

$$=\frac{h'(R)(\eta m_1-(1-m_1)R)}{(1-\eta)m_1}=\frac{h'(R)(\eta(1-\frac{\eta}{10R})-\frac{\eta}{10})}{(1-\eta)m_1}\geq\frac{h'(R)\eta}{2(1-\eta)m_1}\geq\frac{h'(R)\eta}{2}$$

Where in the first inequality we use that $R\geq\frac{1}{\eta}$. Substituting this lower bound into (6.38) we obtain $w\geq g^{-1}\big(\frac{h'(R)\eta}{2}\big)$. Once again using the fact that $h'(R)\geq\max_{[-R,R]}|h'(x)|$ we observe that

$$g(\rho) - g(g^{-1}(0)) \leq \frac{(\rho - g^{-1}(0))h'(R)}{\sqrt{2\pi}}$$

for any $\rho \geq g^{-1}(0)$. This follows by the definition of $g(\cdot)$ and the fact that the mode of the standard gaussian is $\frac{1}{\sqrt{2\pi}}$. Setting $\rho = g^{-1}(\frac{h'(R)\eta}{2})$ we obtain

$$\frac{h'(R)\eta}{2} = g(g^{-1}(\frac{h'(R)\eta}{2})) - g(g^{-1}(0)) \leq \frac{(g^{-1}(\frac{h'(R)\eta}{2}) - g^{-1}(0))h'(R)}{\sqrt{2\pi}}$$

which implies

$$w \geq g^{-1}(\frac{h'(R)\eta}{2}) \geq \eta + g^{-1}(0) \tag{6.39}$$

We then have two possibilities.

Case 1: Either $g^{-1}(0) \geq \frac{-\eta}{2}$ in which case the loss is lower bounded by

$$\mathbb{E}_{(x,y)\sim\mathcal{C}}[(y - w \cdot x)^2] \geq \mathbb{E}_{(x,y)\sim\mathcal{C}}[(y - w \cdot x)^2 | x = -R]\mathbb{P}_{\mathcal{D}_x}(x = -R) = (1 - m_1)R^2(w)^2$$
$$\geq (1 - m_1)R^2(\eta + g^{-1}(0))^2 \geq \frac{\eta^3 R}{40}$$

Where in the first inequality we use the law of total expectation, and in the second inequality we used (6.39) and $g^{-1}(0) \geq \frac{-\eta}{2}$. This is the desired lower bound.

Case 2: In the other case we have $g^{-1}(0) \leq \frac{-\eta}{2}$. Then we flip the sign of the corruptions placed by the adversary. Let the corrupted distribution be

$$\widehat{y} = \begin{cases} (1-\eta)\mathcal{N}(0,1) + \eta\delta(-R-1) & x = 1 \\ \mathcal{N}(0,1) & x = -R \end{cases}$$

Then working through the same calculations flipping signs at the right places we obtain $w \leq g^{-1}\big(-\frac{h'(R)\eta}{2}\big)$. Once again, using that

$$g(\rho) - g(g^{-1}(0)) \geq \frac{(\rho - g^{-1}(0))h'(R)}{\sqrt{2\pi}}$$

for any $\rho \leq g^{-1}(0)$, and setting $\rho = g^{-1}\big(-\frac{h'(R)\eta}{2}\big)$ we obtain

$$-\frac{h'(R)\eta}{2} = g(g^{-1}(-\frac{h'(R)\eta}{2})) - g(g^{-1}(0)) \geq \frac{(g^{-1}\big(-\frac{h'(R)\eta}{2}\big) - g^{-1}(0))h'(R)}{\sqrt{2\pi}}$$

Rearranging we obtain

$$w \leq g^{-1}\big(-\frac{h'(R)\eta}{2}\big) \leq g^{-1}(0) - \eta \leq \frac{-3\eta}{2}$$

Where the last inequality follows by $g^{-1}(0) \leq \frac{-\eta}{2}$. The loss is then lower bounded by

$$\mathbb{E}_{(x,y)\sim\mathcal{C}}[(y - w\cdot x)^2] \geq \mathbb{E}_{(x,y)\sim\mathcal{C}}[(y-w\cdot x)^2 | x = -R]\mathbb{P}_{\mathcal{D}_x}(x = -R) \geq (1-m_1)R^2(w)^2 \geq \frac{9\eta^3 R}{40}$$

where in the last inequality we use $w \leq \frac{-3\eta}{2}$. This is our desired lower bound.

□

6.10 Appendix: Reduction from Contextual Bandits to Online Regression

In this section we verify that the reduction given in [FR20], specifically the proof of Theorem 5 in their paper, also applies to our Huber-contaminated setting as well. Formally, we show the following:

Theorem 6.10.1 (Bandits to Regression Reduction)**.** *Given any oracle $\mathcal{O}$ for Huber-contaminated online regression achieving clean square loss regret* $\mathrm{Reg}_{\mathsf{HSq}}(T)$ *in the sense of Definition 6.4.1, we can produce a learner for Huber-contaminated contextual bandits in the sense of Definition 6.4.4 that achieves clean pseudo-regret* $O\left(\sqrt{KT\cdot\mathrm{Reg}_{\mathsf{HSq}}(T)} + \varepsilon\sqrt{K}T\right)$.

We will use the SquareCB algorithm from [FR20], which draws upon ideas from [AL99], and which we repeat here for completeness (see Algorithm 21).

Proof of Theorem 6.10.1. Fix any policy $\pi : \mathcal{X} \to \mathcal{A}$ and consider the learner given by SquareCB (Algorithm 21) above for a regression oracle $\mathcal{O}$ achieving square loss $\mathrm{Reg}_{\mathsf{HSq}}(T)$, which is some random variable depending on the interactions with Nature. Recall that for

Algorithm 21: SquareCB(A, γ, μ)

```
1 .
  Input: Online regression oracle $\mathcal{O}$, learning rate $\gamma > 0$, exploration parameter $\mu > 0$
  Output: Sequence of actions, in the setting of Definition 6.4.4
2 for $t \in [T]$ do
3     Get context $z_t$ from Nature.
4     For every $a \in \mathcal{A}$, use regression oracle $\mathcal{O}$ to compute prediction $\widehat{y}_{t,a} \triangleq \widehat{y}_t(z_t, a)$.
5     Define $b_t \triangleq \arg\min_{a \in \mathcal{A}} \widehat{y}_{t,a}$.
6     For $a \neq b_t$, define $p_{t,a} = \frac{1}{\mu + \gamma(\widehat{y}_{t,a} - \widehat{y}_{t,b_t})}$ and let $p_{t,b_t} = 1 - \sum_{a \neq b_t} p_{t,a}$. The numbers
      $\{p_{t,a}\}_a$ define a distribution $p_t$ over actions.
7     Sample $a_t$ from $p_t$ and observe loss $\ell$, and update $\mathcal{O}$ with example $((x_t, a_t), \ell)$.
```

this choice of learner, $\mathrm{Reg}_{\mathsf{HCB}}(T)$ is the supremum of

$$\mathbb{E}\left[\sum_{t=1}^{T}(\ell_t^*(a_t) - \ell_t^*(\pi(z_t)))\right]$$

over all such π. Define the filtration

$$\mathfrak{F}_{t-1} \triangleq \sigma((z_1, a_1, \ell_1^*(a_1), \ell_1(a_1), \gamma_1), \ldots, (z_{t-1}, a_{t-1}, \ell_{t-1}^*(a_{t-1}), \ell_{t-1}(a_{t-1}), \gamma_{t-1}), (z_t, \gamma_t)).$$

We can write the sum of conditional expectations of immediate regrets incurred by π as

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}[(\ell_t^*(a_t) - \ell_t^*(\pi(z_t))) \mid \mathfrak{F}_{t-1}] &\leq \sum_{t=1}^{T} \mathbb{E}[(f(z_t, a_t) - f(z_t, \pi(z_t))) \mid \mathfrak{F}_{t-1}] + 2\varepsilon T \\ &\leq \sum_{t=1}^{T} \mathbb{E}[(f(z_t, a_t) - f(z_t, \pi_f(z_t))) \mid \mathfrak{F}_{t-1}] + 2\varepsilon T \\ &= \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} p_{t,a}(f(z_t, a) - f(z_t, \pi_f(z_t))) + 2\varepsilon T. \qquad (6.40)\end{aligned}$$

where recall from Definition 6.4.4 that $\pi_f(z) \triangleq \arg\max_a f(z, a)$, and $p_{t,a}$ is defined in Step 6 of SquareCB

The following lemma is a key ingredient in the reduction of [FR20]:

Lemma 6.10.2 (Lemma 3, [FR20]). *For any collection of numbers $\{\widehat{y}_a\}_{a \in \mathcal{A}} \in [-R, R]^K$, let p be the corresponding probability distribution computed in Step 6. For any collection of* *numbers* $\{f_a\}_{a\in\mathcal{A}} \in \{-R, R\}^K$*, if we define* $a^* \triangleq \arg\max_a f_a$*, we have that*

$$\sum_{a\in\mathcal{A}} p_a \Big[(f_a - f_{a^*}) - \frac{\gamma}{4}(\widehat{y}_a - f_a)^2\Big] \leq \frac{2K}{\gamma}$$

Applying Lemma 6.10.2, we can upper bound (6.40) by

$$\frac{\gamma}{4}\sum_{t=1}^{T} \mathbb{E}[(\widehat{y}_t(z_t, a_t) - f(z_t, a_t))^2 \mid \mathfrak{F}_{t-1}] + \frac{2KT}{\gamma} + 2\varepsilon T.$$

By this and law of total expectation, the pseudo-regret incurred by policy π can be upper bounded by

$$\frac{\gamma}{4}\,\mathbb{E}\big[(\widehat{y}_t(z_t, a_t) - f(z_t, a_t))^2\big] + \frac{2KT}{\gamma} + 2\varepsilon T. \tag{6.41}$$

To bound the prediction error in (6.41), using the identity $b^2 \leq (a+b)^2 - 2ab$, we can upper bound $(\widehat{y}_t(z_t, a_t) - f(z_t, a_t))^2$ by

$$(\widehat{y}_t(z_t, a_t) - \ell_t^*(a_t))^2 - 2(f(z_t, a_t) - \ell_t^*(a_t))(\widehat{y}_t(z_t, a_t) - f(z_t, a_t)). \tag{6.42}$$

Recall from (6.7) that the misspecification adversary is oblivious, that is, conditioned on $\mathfrak{F}_{t-1}$, $f(z_t, a_t) - \ell_t^*(a_t)$ is equal to $-\varepsilon_t(z_t, a_t)$. Putting this and (6.42) together and applying law of total expectation, we can bound the expectation of the prediction error in (6.41) by

$$\begin{aligned}
&\mathbb{E}\big[(\widehat{y}_t(z_t, a_t) - f(z_t, a_t))^2\big] \\
&\quad\leq \mathbb{E}[\mathrm{Reg}_{\mathsf{HSq}}(T)] + 2\,\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[\varepsilon_t(z_t, a_t)(\widehat{y}_t(z_t, a_t) - f(z_t, a_t)) \mid \mathfrak{F}_{t-1}]\right] \\
&\quad\leq \mathbb{E}[\mathrm{Reg}_{\mathsf{HSq}}(T)] + 2\,\mathbb{E}\left[\sum_{t=1}^{T} \varepsilon_t^2(z_t, a_t) + \frac{1}{4}\sum_{t=1}^{T} \mathbb{E}[(\widehat{y}_t(z_t, a_t) - f(z_t, a_t))^2 \mid \mathfrak{F}_{t-1}]\right] \\
&\quad\leq \mathbb{E}[\mathrm{Reg}_{\mathsf{HSq}}(T)] + 2\varepsilon^2 T + \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}[(\widehat{y}_t(z_t, a_t) - f(z_t, a_t))^2],
\end{aligned}$$

which upon rearranging gives

$$\mathbb{E}\big[(\widehat{y}_t(z_t, a_t) - f(z_t, a_t))^2\big] \leq 2\,\mathbb{E}[\mathrm{Reg}_{\mathsf{HSq}}(T)] + 4\varepsilon^2 T.$$

Substituting this into (6.41), and taking $\gamma = 2\sqrt{KT/(\mathbb{E}[\mathrm{Reg}_{\mathsf{HSq}}(T)] + 2\varepsilon^2 T)}$ and $\mu = K$, we conclude that the pseudo-regret incurred by π is upper bounded by

$$\frac{\gamma}{2}(\mathbb{E}[\mathrm{Reg}_{\mathsf{HSq}}(T)] + 2\varepsilon^2 T) + \frac{2KT}{\gamma} + 2\varepsilon T \le 2\sqrt{KT \cdot \mathbb{E}[\mathrm{Reg}_{\mathsf{HSq}}(T)]} + 5\varepsilon\sqrt{K}T$$

as desired. □

In the special case where $\varepsilon = 0$, [FR20] also gives a *high-probability* bound on the *regret* (see their Theorem 1). By adapting their argument, we can show an analogous statement in this setting:

Theorem 6.10.3 (Bandits to Regression Reduction)**.** *Fix any $\delta > 0$. Given any oracle $\mathcal{O}$ for Huber-contaminated online regression achieving clean square loss regret $\mathrm{Reg}_{\mathsf{HSq}}(T)$ in the sense of Definition 6.4.1 with $\varepsilon = 0$, we can produce a learner for Huber-contaminated contextual bandits in the sense of Definition 6.4.4 that with probability at least $1-\delta$ achieves achieves clean regret at most $4\sqrt{KT \cdot \mathrm{Reg}_{\mathsf{HSq}}(T)} + 8\sqrt{KT\log(2/\delta)}$.*

6.11 Appendix: Proof of Theorem 1.3.23

In this section we give a self-contained proof of Theorem 1.3.23, largely following the proof of Equation 5.18 in [KS91].

First, we recall the statement. Suppose that $X_1, \ldots, X_n$ are random vectors in $\mathbb{R}^d$ with $\|X_t\| \le 1$ for all t, and $\xi_1, \ldots, \xi_n$ are random variables such that almost surely, the law of ξ_t conditional on $X_1, \ldots, X_t, \xi_1, \ldots, \xi_{t-1}$ is mean-zero and σ^2-subgaussian. Then

$$\Pr\left[\left\|\frac{1}{n}\sum_{i=1}^n \xi_i X_i\right\| \ge s\right] \le 2\exp\left(\frac{-ns^2}{2\pi\sigma^2}\right).$$

Proof of Theorem 1.3.23. Without loss of generality, we rescale so that $\sigma = 1$. The key observation is that for any $a \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$,

$$F_a \triangleq \mathbb{E}[e^{\lambda \sum_i \xi_i \langle X_i, a\rangle - \lambda^2 \sum_i \langle X_i, a\rangle^2/2}] \le 1. \tag{6.43}$$

The proof of (6.43) follows by an inductive argument. Let $\mathcal{F}_t$ be the filtration generated by $X_1, \ldots, X_t, \xi_1, \ldots, \xi_{t-1}$. Then the first step of the induction is to observe

$$\begin{aligned}\mathbb{E}[e^{\lambda \sum_{i=1}^n \xi_i \langle X_i, a\rangle - \lambda^2 \sum_{i=1}^n \langle X_i, a\rangle^2/2} \mid \mathcal{F}_n] &= e^{\lambda \sum_{i=1}^{n-1} \xi_i \langle X_i, a\rangle - \lambda^2 \sum_{i=1}^{n-1} \langle X_i, a\rangle^2/2}\, \mathbb{E}[e^{\lambda \xi_n \langle X_n, a\rangle - \lambda^2 \langle X_n, a\rangle^2/2} \mid \mathcal{F}_n] \\ &\le e^{\lambda \sum_{i=1}^{n-1} \xi_i \langle X_i, a\rangle - \lambda^2 \sum_{i=1}^{n-1} \langle X_i, a\rangle^2/2}\end{aligned}$$

by the conditional subgaussian assumption on ξ_n. Iterating this argument shows (6.43).

From here the argument follows [KS91]. We let $Z \sim N(0, I_{d\times d})$ be a Gaussian vector independent of everything else, and letting $\gamma = \lambda\sqrt{\pi/2}$ we have

$$\begin{aligned}\mathbb{E}[e^{\lambda \left\| \sum_{i=1}^n \xi_i X_i \right\|}] &\le \mathbb{E}[e^{\gamma\, \mathbb{E}_Z[\|\langle Z, \sum_{i=1}^n \xi_i X_i\rangle|] + [\gamma^2/2](n - \mathbb{E}_Z[\sum_i \langle X_i, Z\rangle^2]))}] \\ &\le e^{n\gamma^2/2}\, \mathbb{E}[e^{\gamma \|\langle Z, \sum_{i=1}^n \xi_i X_i\rangle| - \sum_i \langle X_i, Z\rangle^2})]\end{aligned}$$

where in the first inequality we used $\mathbb{E}[|\langle Z, u\rangle|] = \sqrt{2/\pi}\|u\|$ and $\mathbb{E}_Z[\sum_i \langle X_i, Z\rangle^2] = \sum_i \|X_i\|^2 \le n$ almost surely, and the second step is Jensen's inequality. Using the inequality $e^{|x|} \le e^x + e^{-x}$ gives

$$\mathbb{E}[e^{\gamma \|\langle Z, \sum_{i=1}^n \xi_i X_i\rangle| - (\gamma^2/2) \sum_i \langle X_i, Z\rangle^2})] \le \mathop{\mathbb{E}}_{Z}[F_Z + F_{-Z}] \le 2$$

by (6.43). This shows $e^{\lambda \| \sum_{i=1}^n \xi_i X_i \|} \le 2e^{n\lambda^2\pi/2}$ hence

$$\Pr[e^{\lambda \| \sum_{i=1}^n \xi_i X_i \|} \ge e^{\lambda s}] \le 2e^{n\lambda^2\pi/2 - \lambda s}$$

and taking $\lambda = s/n\pi$ makes the rhs $e^{-s^2/2n\pi}$ which is equivalent to the result. □

Part III

Learning from Heterogeneous Data

Chapter 7

Mixtures of Product Distributions

7.1 Introduction

In this chapter, we turn to the first of three mixture models that we study in this thesis. We begin with the following natural problem, originally introduced in Definition 1.2.19 in slightly different terminology. Recall that a *mixture of subcubes* is a distribution on the Boolean hypercube where each sample is drawn as follows:

(1) There are k mixing weights $\pi^1, \pi^2, \cdots, \pi^k$ and centers $\mu^1, \mu^2, \cdots, \mu^k \in \{0, 1/2, 1\}^n$.

(2) We choose a center proportional to its mixing weight and then sample a point uniformly at random from its corresponding subcube. More precisely, if we choose the i^{th} center, each coordinate is independent and the j^{th} coordinate has expectation μ_i^j.

Our goal here is to give efficient algorithms for estimating the distribution in the PAC-style model of Kearns et al. [KMR+94]. It is not always possible to learn the parameters because two mixtures of subcubes[1] can give rise to identical distributions. Instead, the goal is to output a distribution that is close to the true distribution in total variation distance.

As discussed immediately proceeding Definition 1.2.19, the problem of learning mixtures of subcubes contains various classic problems in computational learning theory as a special case, and is itself a special case of others. For example, for any k-leaf decision tree, the

[1] Even with different numbers of components.

uniform distribution on assignments that satisfy it is a mixture of k subcubes. Likewise, for any function that depends on just j variables (a j-junta), the uniform distribution on assignments that satisfy it is a mixture of 2^j-subcubes. And when we allow the centers μ^i to instead be in the set $[0,1]^n$ it becomes the problem of learning mixtures of binary product distributions.

Each of these problems has a long history of study. Ehrenfeucht and Haussler [EH89] gave an $n^{O(\log k)}$ time algorithm for learning k-leaf decision trees. Blum [Blu92] showed that k-leaf decision trees can be represented as a $\log k$-width decision list and Rivest [Riv87] gave an algorithm for learning ℓ-width decision lists in time $n^{O(\ell)}$. Mossel, O'Donnell and Servedio [MOS03] gave an $n^{j\frac{\omega}{\omega+1}}$ time algorithm for learning j-juntas where ω is the exponent for fast matrix multiplication. Valiant [Val12] gave an improved algorithm that runs in $n^{j\frac{\omega}{4}}$ time. Freund and Mansour [FM99] gave the first algorithm for learning mixtures of two product distributions. Feldman, O'Donnell and Servedio [FOS05] gave an $n^{O(k^3)}$ time algorithm for learning mixtures of k product distributions.

What makes the problem of learning mixtures of subcubes an interesting compromise between expressive power and structure is that it admits surprisingly efficient learning algorithms. The main result of this chapter is an $n^{O(\log k)}$ time algorithm for learning mixtures of subcubes. We also give applications of our algorithm to learning k-leaf decision trees with at most s stochastic transitions on any root-to-leaf path (which also capture interesting scenarios where the transitions are deterministic but there are latent variables). Using our algorithm for learning mixtures of subcubes, we can approximate the error of the Bayes optimal classifier within an additive ε in $n^{O(s+\log k)} \cdot \text{poly}(1/\varepsilon)$ time with an inverse polynomial dependence on the accuracy parameter ε. The classic algorithms of [Riv87, Blu92, EH89] for learning decision trees with zero stochastic transitions achieve this runtime, but because they are Occam algorithms, they break down in the presence of stochastic transitions. Alternatively, the low-degree algorithm [LMN93] is able to get a constant factor approximation to the optimal error (again within an additive ε), while running in time $n^{O(s+\log(k/\varepsilon))}$. The quasipolynomial dependence on $1/\varepsilon$ is inherent to the low-degree approach because the degree needs to grow as the target accuracy decreases, which is undesirable when ε is small as a function of k.

In contrast, we show that mixtures of k subcubes are uniquely identified by their $2 \log k$ order moments. Ultimately our algorithm for learning mixtures of subcubes will allow us to simultaneously match the polynomial dependence on $1/\varepsilon$ of Occam algorithms and achieve the flexibility of the low-degree algorithm in being able to accommodate stochastic transitions. We emphasize that proving identifiability from $2 \log k$ order moments is only a first step in a much more technical argument: There are many subtleties about how we can algorithmically exploit the structure of these moments to solve our learning problem.

7.1.1 Our Results and Techniques

Our main result is an $n^{O(\log k)}$ time algorithm for learning mixtures of subcubes.

Theorem 7.1.1. *Let $\varepsilon, \delta > 0$ be given and let $\mathcal{D}$ be a mixture of k subcubes. There is an algorithm that given samples from $\mathcal{D}$ runs in time $O_k(n^{O(\log k)}(1/\varepsilon)^{O(1)} \log 1/\delta)$ and outputs a mixture $\mathcal{D}'$ of $f(k)$ subcubes that satisfies $d_{TV}(\mathcal{D}, \mathcal{D}') \leq \varepsilon$ with probability at least $1 - \delta$. Moreover the sample complexity is $O_k((\log n/\varepsilon)^{O(1)} \log 1/\delta)$.*[2]

The starting point for our algorithm is the following simple but powerful identifiability result:

Lemma 7.1.2 (Informal)**.** *A mixture of k subcubes is uniquely determined by its $2 \log k$ order moments.*

In contrast, for many sorts of mixture models with k components, typically one needs $\Theta(k)$ moments to establish identifiability [MV10] and this translates to algorithms with running time at least $n^{\Omega(k)}$ and sometimes even much larger than that. In part, this is because the notion of identifiability we are aiming for needs to be weaker and as a result is more subtle. We cannot hope to learn the subcubes and their mixing weights because there are mixtures of subcubes that can be represented in many different ways, sometimes with the same number of subcubes. But as distributions, two mixtures of subcubes are the same if they match on their first $2 \log k$ moments. It turns out that proving this is equivalent to the following basic problem in linear algebra:

[2] Throughout, the hidden constant depending on k will be $O(k^{k^3})$, which we have made no attempt to optimize.

Q2. *Given a matrix $M \in \{0, 1/2, 1\}^{n \times k}$, what is the minimum d for which the set of all entrywise products of at most d rows of M spans the set of all entrywise products of rows of M?*

We show that d can be at most $2 \log k$, which is easily shown to be tight up to constant factors. We will return to a variant of this question later when we discuss why learning mixtures of product distributions requires much higher-order moments.

Unsurprisingly, our algorithm for learning mixtures of subcubes is based on the method of moments. But there is an essential subtlety. For any distribution on the hypercube, $x_i^2 = x_i$. From a technical standpoint, this means that when we compute moments, there is never any reason to take a power of x_i larger than one. We call these *multilinear moments*, and characterizing the way that the multilinear moments determine the distribution (but cannot determine its parameters) is the central challenge. Note that multilinearity makes our problem quite different from typical settings where tensor decompositions can be applied.

Now collect the centers $\mu^1, \mu^2, \cdots, \mu^k$ into a $n \times k$ size matrix that we call the *marginals matrix* and denote by $\mathbf{m}$. The key step in our algorithm is constructing a basis for the entrywise products of rows from this matrix. However we cannot afford to simply brute-force search for this basis among all sets of at most k entrywise products of up to $2 \log k$ rows of $\mathbf{m}$ because the resulting algorithm would run in time $n^{O(k \log k)}$. Instead we construct a basis incrementally.

The first challenge that we need to overcome is that we cannot directly observe the entrywise product of a set of rows of the marginals matrix. But we can observe its weighted inner-product with various other vectors. More precisely, if u, v are respectively the entrywise products of subsets S and T of rows of some marginals matrix $\mathbf{m}$ that realizes the distribution and π is the associated vector of mixing weights, then the relation

$$\sum_{i=1}^{k} \pi^i u_i v_i = \mathbb{E}\left[\prod_{i \in S \cup T} x_i\right]$$

holds if S and T are disjoint. When S and T intersect, this relation is no longer true because in order to express the left hand side in terms of the x_i's we would need to take some powers to be larger than one, which no longer correspond to multilinear moments that can be estimated from samples.

Now suppose we are given a collection $\mathcal{B} = \{T_1, T_2, \cdots, T_k\}$ of subsets of rows of $\mathbf{m}$ and we want to check if the vectors $\{v_1, v_2, \cdots, v_k\}$ (where v_i is the entrywise product of the rows in T_i) are linearly independent. Set $J = \cup_i T_i$. We can define a helper matrix whose columns are indexed by the T_i's and whose rows are indexed by subsets of $[n]\backslash J$. The entry in column i, row S is $\mathbb{E}[\prod_{j \in S \cup T_i} x_j]$ and it is easy to show that if this helper matrix has full row rank then the vectors $\{v_1, v_2, \cdots, v_k\}$ are indeed linearly independent.

The second challenge is that this is an imperfect test. Even if the helper matrix is not full rank, $\{v_1, v_2, \cdots, v_k\}$ might still be linearly independent. Even worse, we can encounter situations where our current collection $\mathcal{B}$ is not yet a basis, and yet for any set we try to add, we cannot certify that the associated entrywise product of rows is outside the span of the vectors we have so far. Our algorithm is based on a win-win analysis. We show that when we get stuck in this way, it is because there is some $S \subseteq [n]$ with $|S| \leq 2 \log k$ where the order $2 \log k$ entrywise products of subets of rows from $[n]\backslash(J \cup S)$ do not span the full k-dimensional space. We show how to identify such an S by repeatedly solving systems of linear equations. Once we identify such an S it turns out that for any string $s \in \{0,1\}^{|J \cup S|}$ we can condition on $x_{J \cup S} = s$ and the resulting conditional distribution will be a mixture of strictly fewer subcubes, which we can then recurse on.

7.1.2 Applications

We demonstrate the power of our $n^{O(\log k)}$ time algorithm for learning mixtures of subcubes by applying it to learning decision trees with stochastic transitions. Specifically suppose we are given a sample x that is uniform on the hypercube, but instead of computing its label based on a k-leaf decision tree with deterministic transitions, some of the transitions are stochastic — they read a bit and based on its value proceed down either the left or right subtree with some unknown probabilities. Such models are popular in medicine [HPS98] and finance [HS65] when features of the system are partially or completely unobserved and the transitions that depend on these features appear to an outside observer to be stochastic. Thus we can also think about decision trees with deterministic transitions but with latent variables as having stochastic transitions when we marginalize on the observed variables.

With stochastic transitions, it is no longer possible to perfectly predict the label even if you know the stochastic decision tree. This rules out many forms of learning like Occam algorithms such as [EH89, Blu92, Riv87] that are based on succinctly explaining a large portion of the observed samples. It turns out that by accurately estimating the distribution on positive examples — via our algorithm for learning mixtures of subcubes — it is possible to approach the Bayes optimal classifier in $n^{O(\log k)}$ time and with only a polylogarithmic number of samples:

Theorem 7.1.3. *Let $\varepsilon, \delta > 0$ be given and let $\mathcal{D}$ be a distribution on labelled examples from a stochastic decision tree under the uniform distribution. Suppose further that the stochastic decision tree has k leaves and along any root-to-leaf path there are at most s stochastic transitions. There is an algorithm that given samples from $\mathcal{D}$ runs in time $O_{k,s}(n^{O(s+\log k)}(1/\varepsilon)^{O(1)} \log 1/\delta)$ and with probability at least $1-\delta$ outputs a classifier whose probability of error is at most $opt + \varepsilon$ where opt is the error of the Bayes optimal classifier. Moreover the sample complexity is $O_{k,s}((\log n/\varepsilon)^{O(1)} \log 1/\delta)$.*

Recall that the low-degree algorithm [LMN93] is able to learn k-leaf decision trees in time $n^{O(\log(k/\varepsilon))}$ by approximating them by $O(\log(k/\varepsilon))$ degree polynomials. These results also generalize to stochastic settings [AM91]. Recently, Hazan, Klivans and Yuan [HKY17] were able to improve the sample complexity even in the presence of adversarial noise using the low-degree Fourier approximation approach together with ideas from compressed sensing for learning low-degree, sparse Boolean functions [SK12]. Although our algorithm is tailored to handle stochastic rather than adversarial noise, our algorithm has a much tamer dependence on ε which yields much faster algorithms when ε is small as a function of k. Moreover we achieve a considerably stronger (and nearly optimal) error guarantee of $\mathrm{opt} + \varepsilon$ rather than $c \cdot \mathrm{opt} + \varepsilon$ for some constant c. Our algorithm even works in the natural variations of the problem [Den98, LDG00, DDS14] where it is only given positive examples.

Lastly, we remark that [DDS14] studied a similar setting where the learner is given samples from the uniform distribution $\mathcal{D}$ over satisfying assignments of some Boolean function f and the goal is to output a distribution close to $\mathcal{D}$. Their techniques seem quite different from ours and also the low-degree algorithm. Among their results, the one most relevant to ours is the incomparable result that there is an $n^{O(\log(k/\varepsilon))}$-time learning algorithm for when

f is a k-term DNF formula.

7.1.3 More Results

As we discussed earlier, mixtures of subcubes are a special case of mixtures of binary product distributions. The best known algorithm for learning mixtures of k product distributions is due to Feldman, O'Donnell and Servedio [FOS05] and runs in time $n^{O(k^3)}$. A natural question which a number of researchers have thought about is whether the dependence on k can be improved, perhaps to $n^{O(\log k)}$. This would match the best known statistical query (SQ) lower bound for learning mixtures of product distributions, which follows from the fact that the uniform distribution over inputs accepted by a decision tree is a mixture of product distributions and therefore from Blum et al.'s $n^{O(\log k)}$ SQ lower bound [BFJ+94].

As we will show, it turns out that mixtures of product distributions require much higher-order moments even to distinguish a mixture of k product distributions from the uniform distribution on $\{0,1\}^n$. As before, this turns out to be related to a basic problem in linear algebra:

Q3. *For a given k, what is the largest possible collection of vectors $v_1, v_2, \cdots, v_m \in \mathbb{R}^k$ for which* (1) *the entries in the entrywise product of any $t < m$ vectors sum to zero and* (2) *the entries in the entrywise product of all m vectors do not sum to zero?*[3]

We show a rather surprising construction that achieves $m = c\sqrt{k}$. An obvious upper bound for m is k. It is not clear what the correct answer ought to be. In any case, we show that this translates to the following negative result:

Lemma 7.1.4 (Informal)**.** *There is a family of mixtures of product distributions that are all different as distributions but which match on all $c\sqrt{k}$ order moments.*

Given a construction for Question 3, the idea for building this family is the same idea that goes into the $n^{\Omega(s)}$ SQ lower bound for s-sparse parity [Kea98] and the $n^{\Omega(k)}$ SQ lower bound for density estimation of mixtures of k Gaussians [DKS17], namely that of hiding a low-dimensional moment-matching example inside a high-dimensional product measure. We leverage Lemma 7.1.4 to show an SQ lower bound for learning mixtures of product

[3] In Section 7.2.5 we discuss the relationship between Questions 2 and 3.

distributions that holds for small values of ε, which is exactly the scenario we are interested in, particularly in applications to learning stochastic decision trees.

Theorem 7.1.5 (Informal)**.** *Any algorithm given $\Omega(n^{-\sqrt{k}/3})$-accurate statistical query access to a mixture $\mathcal{D}$ of k binary product distributions that outputs a distribution $\mathcal{D}'$ satisfying $d_{TV}(\mathcal{D},\mathcal{D}') \leq \varepsilon$ for $\varepsilon \leq k^{-c\sqrt{k}}$ must make at least $n^{c'\sqrt{k}}$ queries.*

This improves upon the previously best known SQ lower bound of $n^{\Omega(\log k)}$, although for larger values of ε our construction breaks down. In any case, in a natural dimension-independent range of parameters, mixtures of product distributions are substantially harder to learn using SQ algorithms than the special case of mixtures of subcubes.

Finally, we leverage the insights we developed for reasoning about higher-order multilinear moments to give improved algorithms for learning mixtures of binary product distributions:

Theorem 7.1.6. *Let $\varepsilon, \delta > 0$ be given and let $\mathcal{D}$ be a mixture of k binary product distributions. There is an algorithm that given samples from $\mathcal{D}$ runs in time $O_k((n/\varepsilon)^{O(k^2)} \log 1/\delta)$ and outputs a mixture $\mathcal{D}'$ of $f(k)$ binary product distributions that satisfies $d_{TV}(\mathcal{D},\mathcal{D}') \leq \varepsilon$ with probability at least $1-\delta$.*

Here we can afford to brute-force search for a basis. However a different issue arises. In the case of mixtures of subcubes, when a collection of vectors that come from entrywise products of rows are linearly independent we can also upper bound their condition number, which allows us to get a handle on the fact that we only have access to the moments of the distribution up to some sampling noise. But when the centers are allowed to take on arbitrary values in $[0,1]^n$ there is no a priori upper bound on the condition number. To handle sampling noise, instead of finding just any basis, we find a barycentric spanner.[4] We proceed via a similar win-win analysis as for mixtures of subcubes: in the case that condition number poses an issue for learning the distribution, we argue that after conditioning on the coordinates of the barycentric spanner, the distribution is *close* to a mixture of fewer product distributions. A key step in showing this is to prove the following *robust* identifiability result that may be of independent interest:

Lemma 7.1.7 (Informal)**.** *Two mixtures of k product distributions are ε-far in statistical*

[4] Specifically, we find a barycentric spanner for just the rows of the *marginals matrix*, rather than for the set of entrywise products of rows of the marginals matrix.

distance if and only if they differ by $\mathrm{poly}(n, 1/\varepsilon, 2^k)^{-O(k)}$ *on a* $2k$*-order moment.*

In fact this is tight in the sense that $o(k)$-order moments are insufficient to distinguish between some mixtures of k product distributions (see the discussion in Section 7.2.5). Another important point is that in the case of mixtures of subcubes, exact identifiability by $O(\log k)$-order moments (Lemma 7.1.2) is non-obvious but, once proven, can be bootstrapped in a black-box fashion to robust identifiability using the abovementioned condition number bound. On the other hand, for mixtures of product distributions, exact identifiability by $O(k)$-order moments is straightforward, but without a condition number bound, it is much more challenging to turn this into a result about robust identifiability.

7.1.4 Organization

The rest of this chapter is organized as follows:

- Section 7.2 — we set up basic definitions, notation, and facts about mixtures of product distributions and provide an overview of our techniques.

- Section 7.3 — we describe our algorithm for learning mixtures of subcubes and give the main ingredients in the proof of Theorem 7.1.1.

- Section 7.4 — we prove the statistical query lower bound of Theorem 7.1.5.

- Section 7.5 — we describe our algorithm for learning general mixtures of product distributions, prove a robust low-degree identifiability lemma in Section 7.5.4, give the main ingredients in the proof of Theorem 7.1.6, and conclude in Section 7.5.6 with a comparison of our techniques to those of [FOS05].

- Appendix 7.6 — we make precise the sampling tree-based framework that our algorithms follow.

- Appendix 7.7 — we complete the proof of Theorem 7.1.1

- Appendix 7.8 — we complete the proof of Theorem 7.1.6

- Appendix 7.9 — we make precise the connection between mixtures of subcubes and various classical learning theory problems, including stochastic decision trees, juntas, and sparse parity with noise, and prove Theorem 7.1.3.

7.2 Preliminaries

7.2.1 Notation and Definitions

Given a matrix A and a set S, we denote $A|_S$ as the restriction of A to rows in S. And similarly $A|^T$ is the restriction of A to columns in T. In this chapter, we will let $\|A\|_\infty$ denote the induced L_∞ operator norm of A, that is, the maximum absolute row sum. We will also make frequent use of entrywise products of vectors and their relation to the multilinear moments of the mixture model.

Definition 7.2.1. *The* entrywise product $\bigodot_{j\in S} v^j$ *of a collection of vectors* $\{v^j\}_{j\in S}$ *is the vector whose* i^{th} *coordinate is* $\prod_{j\in S} v_i^j$. *When* $S=\emptyset$, $\bigodot_{j\in S} v_i$ *is the all ones vector.*

Given a set J, we use 2^J to denote the powerset of J. Let U_n be the uniform distribution over $\{0,1\}^n$. Also let $\mathcal{R}(J) = 2^{[n]\setminus J}$ for convenience. Let $\mathcal{D}(x)$ denote the density of $\mathcal{D}$ at x. Let 1^n be the all ones string of length n.

Definition 7.2.2. *For* $S \subseteq [n]$, *the* S-moment *of* $\mathcal{D}$ *is* $Pr_{\mathcal{D}}[x_S = 1^{|S|}]$. *We will sometimes use the shorthand* $\mathbb{E}_{\mathcal{D}}[x_S]$.

There can be many choices of mixing weights and centers that yield the same mixture of product distributions $\mathcal{D}$. We will refer to any valid choice of parameters as a realization of $\mathcal{D}$.

Definition 7.2.3. *A mixture of* k *product distributions* $\mathcal{D}$ *is a* mixture of k subcubes *if there is a realization of* $\mathcal{D}$ *with mixing weights* $\pi^1,\pi^2,\cdots,\pi^k$ *and centers* $\mu^1,\mu^2,\cdots,\mu^k$ *for which each center has only* $\{0,1/2,1\}$ *values.*

In this chapter, when referring to mixing weights, our superscript notation is only for indexing and never for powering.

There are three main matrices we will be concerned with.

Definition 7.2.4. *The* marginals matrix $\mathbf{m}$ *is a $n \times k$ matrix obtained by concatenating the centers $\mu^1, \mu^2, \cdots, \mu^k$, for some realization. The* moment matrix $\mathbf{M}$ *is a $2^n \times k$ matrix whose rows are indexed by sets $S \subseteq [n]$ and*

$$\mathbf{M}_S = \bigodot_{i \in S} \mathbf{m}_i$$

Finally the cross-check matrix $\mathbf{C}$ *is a $2^n \times 2^n$ matrix whose rows and columns are indexed by sets $S, T \subseteq [n]$ and whose entries are in $[0,1] \cup \{?\}$ where*

$$\mathbf{C}_S^T = \begin{cases} \mathbb{E}_{\mathcal{D}}[x_{S \cup T}] & \text{if } S \cap T = \emptyset \\ ? & \text{otherwise} \end{cases}$$

We say that an entry of $\mathbf{C}$ is accessible *if it is not equal to ?.*

It is important to note that $\mathbf{m}$ and $\mathbf{M}$ depend on the choice of a particular realization of $\mathcal{D}$, but that $\mathbf{C}$ does not because its entries are defined through the moments of $\mathcal{D}$. The starting point for our algorithms is the following observation about the relationship between $\mathbf{M}$ and $\mathbf{C}$:

Observation 7.2.5. *For any realization of $\mathcal{D}$ with mixing weights π and centers $\mu^1, \mu^2, \cdots, \mu^k$. Then*

(1) For any set $S \subseteq [n]$ we have $\mathbf{M}_S \cdot \pi = \mathbb{E}_{\mathcal{D}}[x_S]$

(2) For any pair of sets $S, T \subseteq [n]$ with $S \cap T = \emptyset$ we have

$$\mathbf{C}_S^T = \Big(\mathbf{M} \cdot diag(\pi) \cdot \mathbf{M}^\top\Big)_S^T$$

The idea behind our algorithms are to find a basis for the rows of $\mathbf{M}$ or failing that to find some coordinates to condition on which result in a mixture of fewer product distributions. The major complications come from the fact that we can only estimate the accessible entries of $\mathbf{C}$ from samples from our distribution. If we had access to all of them, it would be straightforward to use the above relationship between $\mathbf{M}$ and $\mathbf{C}$ to find a set of rows of $\mathbf{M}$ that span the row space.

7.2.2 Rank of the Moment Matrix and Conditioning

First we will show that without loss of generality we can assume that the moment matrix $\mathbf{M}$ has full column rank. If it does not, we will be able to find a new realization of $\mathcal{D}$ as a mixture of strictly fewer product distributions.

Definition 7.2.6. *A realization of $\mathcal{D}$ is a* full rank realization *if $\mathbf{M}$ has full column rank and all the mixing weights are nonzero. Furthermore if* $\text{rank}(\mathbf{M}) = k$ *we will say $\mathcal{D}$ has rank k.*

Lemma 7.2.7. *Fix a realization of $\mathcal{D}$ with mixing weights π and centers $\mu^1, \mu^2, \cdots, \mu^k$ and let $\mathbf{M}$ be the moment matrix. If* $\text{rank}(\mathbf{M}) = r < k$ *then there are new mixing weights π' such that:*

(1) π' has r nonzeros

(2) π' and $\mu^1, \mu^2, \cdots, \mu^k$ also realize $\mathcal{D}$.

Moreover the submatrix $\mathbf{M}'$ consisting of the columns of $\mathbf{M}$ with nonzero mixing weight in π' has rank r.

Proof. We will proceed by induction on r. When $r = k - 1$ there is a vector $v \in \ker(\mathbf{M})$. The sum of the entries in v must be zero because the first row of $\mathbf{M}$ is the all ones vector. Now if we take the line $\pi + tv$ as we increase t, there is a first time t_0 when a coordinate becomes zero. Let $\pi' = \pi + t_0 v$. By construction, π' is nonnegative and its entries sum to one and it has at most $k - 1$ nonzeros. We can continue in this fashion until the columns corresponding to the support of π' in $\mathbf{M}$ are linearly independent. Note that as we change the mixing weights, the moment matrix $\mathbf{M}$ stays the same. Also the resulting matrix $\mathbf{M}'$ that we get must have rank r because each time we update π we are adding a multiple of a vector in the kernel of $\mathbf{M}$ so the columns whose mixing weight is changing are linearly dependent. □

Thus when we fix an (unknown) realization of $\mathcal{D}$ in our analysis, we may as well assume that it is a full rank realization. This is true even if we restrict our attention to mixtures of subcubes where the above lemma shows that if $\mathbf{M}$ does not have full column rank, there is a mixture of $r < k$ subcubes that realizes $\mathcal{D}$. Next we show that mixtures of product distributions behave nicely under conditioning:

Lemma 7.2.8. *Fix a realization of $\mathcal{D}$ with mixing weights π and centers $\mu^1, \mu^2, \cdots, \mu^k$. Let $S \subseteq [n]$ and $s \in \{0,1\}^{|S|}$. The the conditional distribution $\mathcal{D}|_{x_S=s}$ can be realized as a mixture of k product distributions with mixing weights π' and centers*

$$\mu^1|_{[n]\setminus S}, \mu^2|_{[n]\setminus S}, \cdots, \mu^k|_{[n]\setminus S}$$

Proof. Using Bayes' rule we can write out the mixing weights π' explicitly as

$$\pi' = \frac{\pi \bigodot \left(\bigodot_{i \in S} \gamma^i \right)}{\Pr_{\mathcal{D}}[x_S = s]}$$

where we have abused notation and used $\bigodot$ as an infix operator and where $\gamma^i = \mu^i + (1 - s_i) \cdot (\mathbf{1} - 2\mu^i)$. This follows because the map $x \mapsto x + (1 - s) \cdot (1 - 2x)$ is the identity when $s = 1$ and $x \mapsto 1 - x$ when $s = 0$ □

We can straightforwardly combine Lemma 7.2.7 and Lemma 7.2.8 to conclude that if $\operatorname{rank}(\mathbf{M}|_{2^{[n]\setminus S}}) = r$ then for any $s \in \{0,1\}^{|S|}$ there is a realization of $\mathcal{D}|_{x_S=s}$ as a mixture of r product distributions. Moreover if $\mathcal{D}$ was a mixture of subcubes then so too would the realization of $\mathcal{D}|_{x_S=s}$ be.

7.2.3 Linear Algebraic Relations between M and C

Even though not all of the entries of $\mathbf{C}$ are accessible (i.e. can be estimated from samples from $\mathcal{D}$) we can still use it to deduce linear algebraic properties among the rows of $\mathbf{M}$. All of the results in this subsection are elementary consequences of Observation 7.2.5.

Lemma 7.2.9. *Let $T_1, T_2, \cdots, T_r \subseteq [n]$ and set $J = \cup_i T_i$. If the columns*

$$\mathbf{C}^{T_1}|_{\mathcal{R}(J)}, \mathbf{C}^{T_2}|_{\mathcal{R}(J)}, \cdots, \mathbf{C}^{T_r}|_{\mathcal{R}(J)}$$

are linearly independent then for any realization of $\mathcal{D}$ the rows $\mathbf{M}_{T_1}, \mathbf{M}_{T_2}, \cdots, \mathbf{M}_{T_r}$ are also linearly independent.

Proof. Fix any realization of $\mathcal{D}$. Using Observation 7.2.5, we can write:

$$\mathbf{C}|_{\mathcal{R}(J)}^{T_1,\dots,T_r} = \mathbf{M}|_{\mathcal{R}(J)} \cdot \text{diag}(\pi) \cdot \Big(\mathbf{M}^\top\Big)|^{T_1,\dots,T_r}$$

Now suppose for the sake of contradiction that the rows of $\mathbf{M}|_{T_1,\dots,T_r}$ are not linearly independent. Then there is a nonzero vector u so that $(\mathbf{M}^\top)|^{T_1,\dots,T_r} u = 0$ which by the above equation immediately implies that the columns of $\mathbf{C}|_{\mathcal{R}(J)}^{T_1,\dots,T_r}$ are not linearly independent, which yields our contradiction. □

Next we prove a partial converse to the above lemma:

Lemma 7.2.10. *Fix a realization of $\mathcal{D}$ and let $\mathcal{D}$ have rank k. Let $T_1, T_2, \cdots, T_r \subseteq [n]$ and set $J = \cup_i T_i$. If $\text{rank}(\mathbf{M}|_{\mathcal{R}(J)}) = k$ and there are coefficients $\alpha_1, \alpha_2, \cdots, \alpha_r$ so that*

$$\sum_{i=1}^{r} \alpha_i \mathbf{C}^{T_i}|_{\mathcal{R}(J)} = 0$$

then the corresponding rows of $\mathbf{M}$ are linearly dependent too — i.e. $\sum_{i=1}^{r} \alpha_i \mathbf{M}_{T_i} = 0$.

Proof. By the assumptions of the lemma, we have that

$$\mathbf{M}|_{\mathcal{R}(J)} \cdot \text{diag}(\pi) \cdot \Big(\mathbf{M}^\top\Big)|^{T_1,\dots,T_r} \alpha = 0$$

Now $\text{rank}(\mathbf{M}|_{\mathcal{R}(J)}) = k$ and the fact that the mixing weights are nonzero implies that $\mathbf{M}|_{\mathcal{R}(J)} \cdot \text{diag}(\pi)$ is invertible. Hence we conclude that $\Big(\mathbf{M}^\top\Big)|^{T_1,\dots,T_r} \alpha = 0$ as desired. □

Of course, we don't actually have exact estimates of the moments of $\mathcal{D}$, so in Appendix 7.7 we prove the sampling noise-robust analogues of Lemma 7.2.9 and Lemma 7.2.10 (see Lemma 7.7.1) needed to get an actual learning algorithm.

7.2.4 Technical Overview for Learning Mixtures of Subcubes

With these basic linear algebraic relations in hand, we can explain the intuition behind our algorithms. Our starting point is the observation that if we know a collection of sets $T_1, ..., T_k \subset [n]$ indexing a row basis of $\mathbf{M}$, then we can guess one of the $3^{k\cdot|T_1\cup\cdots\cup T_k|}$ possibilities for the entries of $\mathbf{m}|_{T_1\cup\cdots\cup T_k}$. Using a correct guess, we can solve for the mixing weights using (1) from Observation 7.2.5. The point is that because $T_1, ..., T_k$ index a row basis of $\mathbf{M}$, the system of equations

$$\mathbf{M}_{T_j} \cdot \pi = \mathbb{E}_{\mathcal{D}}[x_{T_j}], \quad j = 1, ..., k \tag{7.1}$$

has a unique solution which thus must be the true mixing weights in the realization $(\pi, \mathbf{m})$. We can then solve for the remaining rows of $\mathbf{m}$ using part 2 of Observation 7.2.5, i.e. for every $i \notin T_1 \cup \cdots \cup T_k$ we can solve

$$\mathbf{M}_{T_j} \cdot \mathrm{diag}(\pi) \cdot \mathbf{m}_i^\top = \mathbb{E}_{\mathcal{D}}[x_{T_j\cup\{i\}}] \quad \forall j = 1, ..., k. \tag{7.2}$$

Again, because the rows $\mathbf{M}_{T_i}$ are linearly independent and π has no zero entries, we conclude that the true value of $\mathbf{m}_i$ is the unique solution.

There are three main challenges to implementing this strategy:

A **Identifiability**. How do we know whether a given guess for $\mathbf{m}|_{T_1\cup\cdots\cup T_k}$ is correct? More generally, how do we efficiently test whether a given distribution is close to the underlying mixture of subcubes?

B **Building a Basis**. How do we produce a row basis for $\mathbf{M}$ without knowing $\mathbf{M}$, let alone one for which $T_1\cup\cdots\cup T_k$ is small enough that we can actually try all $3^{k\cdot|T_1\cup\cdots\cup T_k|}$ possibilities for $\mathbf{m}|_{T_1\cup\cdots\cup T_k}$?

C **Sampling Noise**. Technically we only have approximate access to the moments of $\mathcal{D}$, so even from a correct guess for $\mathbf{m}|_{T_1\cup\cdots\cup T_k}$ we only obtain approximations to π and the remaining rows of $\mathbf{m}$. How does sampling noise affect the quality of these approximations?

Identifiability

As our algorithms will be based on the method of moments, an essential first question to answer is that of identifiability: what is the minimum d for which mixtures of k subcubes are uniquely identified by their moments of degree at most d? As alluded to in Section 7.1.1, it is enough to answer Question 2, which we can restate in our current notation as:

Q4. *Given a matrix* $\mathbf{m} \in \{0, 1/2, 1\}^{n\times k}$ *with associated* $2^n \times k$ *moment matrix* $\mathbf{M}$*, what is the minimum* d *for which the rows* $\{\mathbf{M}_S\}_{|S|\le d}$ *span all rows of* $\mathbf{M}$*?*

Let $d(k)$ be the largest d for Question 4 among all $\mathbf{m} \in \{0, 1/2, 1\}^{n\times k}$. Note that $d(k) = \Omega(\log k)$ just from considering a $O(\log k)$-sparse parity with noise instance as a mixture of k subcubes. The reason getting upper bounds on $d(k)$ is directly related to identifiability is that k subcubes are uniquely identified by their moments of degree at most $d(2k)$. Indeed, if $(\pi_1, \mathbf{m}_1)$ and $(\pi_2, \mathbf{m}_2)$ realize different distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, then there must exist $S \subseteq [n]$ for which

$$(\mathbf{M}_1)_S \cdot \pi_1 = \mathbb{E}_{\mathcal{D}_1}[x_S] \neq \mathbb{E}_{\mathcal{D}_2}[x_S] = (\mathbf{M}_2)_S \cdot \pi_2.$$

In other words, the vector $(\pi_1| - \pi_2) \in \mathbb{R}^{2k}$ does not lie in the right kernel of the matrix $2^n \times 2k$ matrix $(\mathbf{M}_1|\mathbf{M}_2)$. But because $\mathbb{N} \triangleq (\mathbf{M}_1|\mathbf{M}_2)$ is the moment matrix of the matrix $(\mathbf{m}_1|\mathbf{m}_2) \in \{0, 1/2, 1\}^{n\times 2k}$, its rows are spanned by the rows $(\mathbb{N}_S)_{|S|\le d(2k)}$, so there in fact exists S' of size at most $d(2k)$ for which $\mathbb{E}_{\mathcal{D}_1}[x_{S'}] \neq \mathbb{E}_{\mathcal{D}_2}[x_{S'}]$. Finally, note also that the reverse direction of this argument holds, that is, if mixtures of k subcubes $\mathcal{D}_1$ and $\mathcal{D}_2$ agree on all moments of degree at most $d(2k)$, then they are identical as distributions.

In Section 7.3.1, we show that $d(k) = \Theta(\log k)$. The idea is that there is a natural correspondence between 1) linear relations among the rows of $\mathbf{M}_S$ for $|S| \le d$ and 2) multilinear polynomials of degree at most d which vanish on the rows of $\mathbf{m}$. The bound on $d(k)$ then follows from cleverly constructing an appropriate low-degree multilinear polynomial.

Note that the above discussion only pertains to *exact identifiability*. For the purposes of our learning algorithm, we want *robust identifiability*, i.e. there is some $d'(k)$ such that $\mathcal{D}_1$ and $\mathcal{D}_2$ are far in statistical distance if and only if they differ noticeably on some moment of degree at most $d'(k)$. It turns out that it suffices to take $d'(k)$ to be the same $\Theta(\log k)$, and in Section 7.2.4 below, we sketch how we achieve this.

Once we have robust identifiability in hand, we have a way to resolve Challenge A above: to check whether a given guess for $\mathbf{m}|_{T_1\cup\cdots\cup T_k}$ is correct, compute the moments of degree at most $\Theta(\log k)$ of the corresponding candidate mixture of subcubes and compare them to empirical estimates of the moments of the underlying mixture. If they are close, then the mixture of subcubes we have learned is close to the true distribution.

As we will see below though, while the bound of $d(k) = \Theta(\log k)$ is a necessary first step to achieving a quasipolynomial running time for our learning algorithm, there will be many more steps and subtleties along the way to getting an actual algorithm.

Building a Basis

We now describe how we address Challenge B. The key issue is that we do not have access to the entries of $\mathbf{M}$ (and $\mathbf{M}$ itself depends on the choice of a particular realization). Given the preceding discussion about Question 4, a naive way to circumvent this is simply to guess a basis from among all combinations of at most k rows from $\{\mathbf{M}_S\}_{|S|\le d(k)}$, but this would take time $n^{\Theta(k\log k)}$.

As we hinted at in Section 7.1.1, we will overcome the issue of not having access to $\mathbf{M}$ by using the accessible entries of $\mathbf{C}$, which we can easily estimate by drawing samples from $\mathcal{D}$, as a surrogate for $\mathbf{M}$ (see Lemmas 7.2.9 and 7.2.10). To this end, one might first try to use $\mathbf{C}$ to find a row basis for $\mathbf{M}$ by looking at the submatrix of $\mathbf{C}$ consisting of entries $\{\mathbf{C}^T_S\}_{S,T:|S|,|T|\le d(k)}$ and simply picking out a column basis $\{T_1, ..., T_r\}$ for this submatrix. Of course, the crucial issue is that we can only use the accessible entries of $\mathbf{C}$.

Instead, we will incrementally build up a row basis. Suppose at some point we have found a list of subsets $T_1, ..., T_m$ indexing linearly independent rows of $\mathbf{M}$ for some realization of $\mathcal{D}$ and are deciding whether to add some set T to this list. By Lemmas 7.2.9 and 7.2.10, if $\text{rank}(\mathbf{M}|_{\mathcal{R}(J)}) = k$, where $J = T \cup (T_1 \cup \cdots \cup T_m)$, then $\mathbf{M}_T$ is linearly independent from $\mathbf{M}_{T_1}, ..., \mathbf{M}_{T_m}$ if and only if the column vector $\mathbf{C}^T|_{\mathcal{R}(J)}$ is linearly independent from column vectors $\mathbf{C}^{T_1}|_{\mathcal{R}(J)}, ..., \mathbf{C}^{T_m}|_{\mathcal{R}(J)}$.[5]

If we make the strong assumption that we always have that $\text{rank}(\mathbf{M}|_{\mathcal{R}(J)}) = k$ in the

[5] Note that while the dimension of these column vectors is exponential in n, the discussion in Section 7.2.4 implies that it suffices to look only at the coordinates of these columns that are indexed by S with $|S| \le d(k) = \Theta(\log k)$.

course of running this procedure, the problem of finding a row basis for $\mathbf{M}$ reduces to the following basic question:

Q5. *Given $T_1, ..., T_m$ indexing linearly independent rows of a moment matrix $\mathbf{M}$, as well as access to an oracle which on input T decides whether $\mathbf{M}_T$ lies in the span of $\mathbf{M}_{T_1}, ..., \mathbf{M}_{T_m}$, how many oracle calls does it take to either find T for which $\mathbf{M}_T$ lies outside the span of $\mathbf{M}_{T_1}, ..., \mathbf{M}_{T_m}$ or successfully conclude that $\mathbf{M}_{T_1}, ..., \mathbf{M}_{T_m}$ are a row basis for $\mathbf{M}$?*

Section 7.2.4 tells us it suffices to look at all remaining subsets of size at most $d(k)$ which have not yet been considered, which requires checking at most $n^{O(\log k)}$ subsets before we decide whether to add a new subset to our basis.

Later, in Section 7.3.4, we will show the following alternative approach which we call GrowByOne suffices: simply consider all subsets of the form $T_j \cup \{i\}$ for $1 \leq i \leq m$ and $i \notin T_1 \cup \cdots \cup T_m$. If $T_1, ..., T_m$ have up to this point been constructed in this incremental fashion, we prove that if no such $T_j \cup \{i\}$ can be added to our list and moreover we have that $\text{rank}(\mathbf{M}|_{\mathcal{R}(J)}) = \text{rank}(\mathbf{M}|_{\mathcal{R}(T_1 \cup \cdots \cup T_m \cup \{i\})}) = k$ for every i, then $T_1, ..., T_m$ indexes a row basis for $\mathbf{M}$.

The advantages of GrowByOne are that it 1) only requires checking at most nk subsets before we decide whether to add a new subset to our basis, 2) it works even when we assume $\mathbf{M}$ is the moment matrix of a mixture of *arbitrary product distributions*, and 3) it will simplify our analysis regarding issues of sampling noise.

Making Progress When Basis-Building Fails

The main subtlety is that the correctness of GrowByOne as outlined in Section 7.2.4 hinges on the fact that $\text{rank}(\mathbf{M}|_{\mathcal{R}(J)}) = k$ at every point in the algorithm. But if this is not the case and yet $\mathbf{C}^T_{\mathcal{R}(J)}$ lies in the span of $\mathbf{C}^{T_1}_{\mathcal{R}(J)}, ..., \mathbf{C}^{T_m}_{\mathcal{R}(J)}$, we cannot conclude whether $\mathbf{M}_T$ lies in the span of $\mathbf{M}_{T_1}, ..., \mathbf{M}_{T_m}$. In particular, suppose we found that $\mathbf{C}^T_{\mathcal{R}(J)}$ lies in the span of $\mathbf{C}^{T_1}_{\mathcal{R}(J)}, ..., \mathbf{C}^{T_m}_{\mathcal{R}(J)}$ for every candidate subset $T = T_j \cup \{i\}$ and therefore decided to add nothing more to the list $T_1, ..., T_m$. Then while Lemma 7.2.9 guarantees that the rows of $\mathbf{M}$ corresponding to $T_1, ..., T_m$ are linearly independent, we can no longer ascertain that they span all the rows of $\mathbf{M}$.

The key idea is that if this is the case, then there must have been some candidate $T = T_j \cup$

$\{i\}$ such that $\text{rank}(\mathbf{M}|_{\mathcal{R}(T_1\cup\cdots\cup T_m\cup\{i\})}) < k$. We call the set of all such i the set of *impostors*. By Lemma 7.2.7, if i is an impostor, the conditional distribution $(\mathcal{D}|x_{T_1\cup\cdots\cup T_m\cup\{i\}} = s)$ can be realized as a mixture of strictly fewer than k subcubes for any bitstring s. The upshot is that even if the list $T_1, ..., T_m$ output by GrowByOne does not correspond to a row basis of $\mathbf{M}$, we can make progress by conditioning on the coordinates $T_1 \cup \cdots \cup T_m \cup \{i\}$ for an impostor i and recursively learning mixtures of fewer subcubes.

On the other hand, the issue of actually identifying an impostor $i \notin T_1 \cup \cdots \cup T_m$ is quite delicate. Because there may be up to k levels of recursion, we cannot afford to simply brute force over all $n - |T_1 \cup \cdots \cup T_n|$ possible coordinates. Instead, the idea will be to pretend that $T_1, ..., T_m$ actually corresponds to a row basis of $\mathbf{M}$ and use this to attempt to learn the parameters of the mixture. It turns out that either the resulting mixture will be close to $\mathcal{D}$ on all low-degree moments and robust identifiability will imply we have successfully learned $\mathcal{D}$, or it will disagree on some low-degree moment, and we show in Section 7.3.3 that this low-degree moment must contain an impostor i.

Sampling Noise

Obviously we only have access to empirical estimates of the entries of $\mathbf{C}$, so for instance, instead of checking whether a column of $\mathbf{C}$ lies in the span of other columns of $\mathbf{C}$, we look at the corresponding L_∞ regression problem. In this setting, the above arguments still carry over provided that the submatrices of $\mathbf{M}$ and $\mathbf{C}$ used are well-conditioned. We show in Section 7.3.5 that the former are well-conditioned by Cramer's, as they are matrices whose entries are low-degree powers of $1/2$, and this on its own can already be used to show robust identifiability. By Observation 7.2.5, the submatrices of $\mathbf{C}$ used in the above arguments are also well-conditioned provided that π has no small entries. But if π has small entries, intuitively we might as well ignore these entries and only attempt to learn the subcubes of the mixture which have non-negligible mixing weight.

In Section 7.3.5, we explain in greater detail the subtleties that go into dealing with these issues of sampling noise.

7.2.5 Technical Overview for SQ Lower Bound

To understand the limitations of the method of moments for more general mixtures of product distributions, we can first ask Question 4 more generally for arbitrary matrices $\mathbf{m} \in \mathbb{R}^{n\times k}$, but in this case it is not hard to see that the minimum d for which the rows $\{\mathbf{M}_S\}_{|S|\le d}$ span all rows of $\mathbf{M}$ can be as high as $k-1$. Simply take $\mathbf{m}$ to have identical rows, each of which consists of k distinct entries $z_1, ..., z_k \in [0,1]$. Then $\mathbf{M}_S = (z_1^{|S|}, ..., z_k^{|S|})$, so by usual properties of Vandermonde matrices, the rows $\{\mathbf{M}_S\}_{|S|\le d}$ will not span the rows of $\mathbf{M}$ until $d \ge k-1$.[6]

From such an $\mathbf{m}$, we immediately get a pair of mixtures $(\mu_1, \mathbf{m}_1)$ and $(\mu_2, \mathbf{m}_2)$ that agree on all moments of degree at most $k-2$ but differ on moments of degree $k-1$: let μ_1 and $-\mu_2$ up to scaling be the positive and negative parts of an element in the kernel of $\{M_S\}_{|S|<k-1}$, and let $\mathbf{m}_1$ and $\mathbf{m}_2$ be the corresponding disjoint submatrices of $\mathbf{m}$. But this is not yet sufficient to establish an SQ lower bound of $n^{\Omega(k)}$.

Instead, we will exhibit a large collection $\mathcal{C}$ of mixtures of k product distributions that all agree with the uniform distribution over $\{0,1\}^n$ on moments up to some degree $d^*(k)-1$ but differ on some moment of degree $d^*(k)$. This will be enough to give an SQ lower bound of $n^{\Omega(d^*(k))}$.

The general approach is to construct a mixture $\mathcal{A}$ of product distributions over $\{0,1\}^{d^*(k)}$ whose top-degree moment differs noticeably from $2^{-d^*(k)}$ but whose other moments agree with that of the uniform distribution over $\{0,1\}^{d^*(k)}$. The collection $\mathcal{C}$ of mixtures will then consist of all product measures given by $\mathcal{A}$ in some $d^*(k)$ coordinates S and the uniform distribution over $\{0,1\}^{n-d^*(k)}$ in the remaining coordinates $[n]\backslash S$. This general strategy of embedding a low-dimensional moment-matching distribution $\mathcal{A}$ in some hidden set of coordinates is the same principle behind SQ lower bounds for learning sparse parity [Kea98], robust estimation and density estimation of mixtures of Gaussians [DKS17], etc.

The main challenge is to actually construct the mixture $\mathcal{A}$. We reduce this problem to Question 3 and give an explicit construction in Section 7.4 with $d^*(k) = \Theta(\sqrt{k})$.

[6] Note that by the connection between linear relations among rows of $\mathbf{M}_S$ and multilinear polynomials vanishing on the rows of $\mathbf{m}$, this example is also tight, i.e. $\{\mathbf{M}_S\}_{|S|\le k-1}$ will span the rows of $\mathbf{M}$ for any $m \in \mathbb{R}^{n\times k}$.

7.2.6 Technical Overview for Learning Mixtures of Product Distributions

The main difficulty with learning mixtures of general product distributions is that moment matrices can be arbitrarily ill-conditioned, which makes it far more difficult to handle sampling noise. Indeed, with exact access to the accessible entries of $\mathbf{C}$, one can in fact show there exists a $n^{O(d^*(k))}$ algorithm for learning mixtures of general product distributions, where $d^*(k)$ is the answer to Question 3, though we omit the proof of this in this work. In the presence of sampling noise, it is not immediately clear how to adapt the approach from Section 7.2.4. The three main challenges are:

A **Robust Identifiability**. For mixtures of subcubes, robust identifiability essentially followed from exact identifiability and a condition number bound on $\mathbf{M}$. Now that $\mathbf{M}$ can be arbitrarily ill-conditioned, how do we still show that two mixtures of product distributions that are far in statistical distance must differ noticeably on some low-degree moment?

B **Using $\mathbf{C}$ as a Proxy for $\mathbf{M}$**. Without a condition number bound, can approximate access to $\mathbf{C}$ still be useful for deducing (approximate) linear algebraic relations among the rows of $\mathbf{M}$?

C **Guessing Entries of $\mathbf{m}$**. Entries of $\mathbf{m}$ are arbitrary scalars now, rather than numbers from $\{0, 1/2, 1\}$. We can still try discretizing by guessing integer multiples $0, \eta, 2\eta, ..., 1$ of some small scalar η, but how small must η be for this to work?

For Challenge A, we will show that if two mixtures of k product distributions are far in statistical distance, they must differ noticeably on some moment of degree at most $2k$. Roughly, the proof is by induction on the total number of product distributions in the two mixtures, though the inductive step is rather involved and we defer the details to Section 7.5.4, which can be read independently of the other parts of the proof of Theorem 7.1.6.

Next, we make Challenges B and C more manageable by shifting our goal: instead of a row basis for $\mathbf{M}$, we would like a row basis for $\mathbf{m}$ that is well-conditioned in an appropriate sense. Specifically, we want a row basis $J \subset [n]$ for $\mathbf{m}$ such that if we express any other row of $\mathbf{m}$ as a linear combination of this basis, the corresponding coefficients are small. This is precisely the notion of *barycentric spanner* introduced in [AK08], where it was shown that any collection of vectors has a barycentric spanner. We can find a barycentric spanner for the rows of $\mathbf{m}$ by simply guessing all $\binom{n}{k}$ possibilities. We then show that if $J = \{i_1, ..., i_r\}$ is a barycentric spanner and $\mathbf{M}|_{\mathcal{R}(J \cup i_j)}$ is well-conditioned in an L_∞ sense for all $1 \leq j \leq r$, then in analogy with Lemma 7.2.10, one can learn good approximations to the true coefficients expressing the remaining rows of $\mathbf{m}$ in terms of $\mathbf{m}_{i_1}, ..., \mathbf{m}_{i_r}$. Furthermore, these approximations are good enough that it suffices to pick the discretization parameter in Challenge C to be $\eta = \text{poly}(\varepsilon/n)$, in which case the k^2 entries of $\mathbf{m}|_J$ can be guessed in time $(n/\varepsilon)^{O(k^2)}$.

If instead $\mathbf{M}|_{\mathcal{R}(J \cup \{i_j\})}$ is ill-conditioned for some "impostor" $1 \leq j \leq r$, we can afford now to simply brute-force search for the impostor, but we cannot appeal to Lemma 7.2.7 to argue as before that each of the conditional distributions $(\mathcal{D}|x_{J \cup \{i_j\}} = s)$ is a mixture of fewer than k product distributions, because $\mathbf{M}|_{\mathcal{R}(J \cup \{i_j\})}$ might still have rank k. Instead, we show in Section 7.5.5 that robust identifiability implies that these conditional distributions are *close* to mixtures of at most $k - 1$ product distributions, and this is enough for us to make progress and recursively learn.

7.3 Learning Mixtures of Subcubes in Quasipolynomial Time

7.3.1 Logarithmic Moments Suffice

Recall that a mixture of k subcubes can represent the distribution on positive examples from an s-sparse parity with noise when $k = 2^{s-1} + 1$. It is well known that every $s - 1$ moments of such a distribution are indistinguishable from the uniform distribution. Here we prove a converse and show that for mixtures of k subcubes all of the relevant information is contained within the $O(\log k)$ moments. More precisely we show:

Lemma 7.3.1. *Let $\mathcal{D}$ be a mixture of k subcubes and fix a realization where the centers are*

$\{0, 1/2, 1\}$*-valued. Let* $\mathbf{M}$ *be the corresponding moment matrix. Then*

$$\Big\{\mathbf{M}_T \Big| |T| < 2\log k\Big\}$$

span the rows of $\mathbf{M}$.

Proof. Fix any set $S \subseteq [n]$ of size $m = 2\log k$. Without loss of generality suppose that $S = \{1, 2, \cdots, m\}$. We want to show that $\mathbf{M}_S$ lies in the span of $\mathbf{M}_T$ for all $T \subsetneq S$. Our goal is to show that there are coefficients α_T so that

$$\sum_{T \subseteq S} \alpha_T \mathbf{M}_T = 0$$

and that α_S is nonzero. If we can do this, then we will be done. First we construct a multilinear polynomial

$$p(x) = \prod_{i=1}^{m} \Big(x_i - \lambda_i\Big)$$

where each $\lambda_i \in \{0, 1/2, 1\}$ and with the property that for any j, $p(\mathbf{m}^j|_S) = 0$. If we had such a polynomial, we could expand

$$p(x) = \sum_{T \subseteq S} \alpha_T \prod_{i \in T} x_i$$

By construction $\alpha_S = 1$. And now for any j we can see that the j^{th} coordinate of $\sum_{T \subseteq S} \alpha_T \mathbf{M}_T$ is exactly $p(\mathbf{m}^j|_S)$, which yields the desired linear dependence.

All that remains is to construct the polynomial p. We will do this by induction. Suppose we have constructed a polynomial $p_t(x) = \prod_{i=1}^{t}(x_i - \lambda_i)$ and let

$$R_t = \Big\{j \Big| p_t(\mathbf{m}^j|_S) \neq 0\Big\}$$

In particular $R_t \subseteq [k]$ is the set of surviving columns. By the pigeonhole principle we can choose $\lambda_{t+1} \in \{0, 1/2, 1\}$ so that $|R_{t+1}| \leq \lfloor (2/3)|R_t| \rfloor$. For some $\ell \leq m$ we have that $R_\ell = \emptyset$

at which point we can choose

$$p(x) = \Big(\prod_{i=1}^{\ell}(x_i - \lambda_i)\Big) \cdot \prod_{i=\ell+1}^{m} x_i$$

which completes the proof. □

Recall that $\mathcal{R}(J) = 2^{[n]\setminus J}$. Now Lemma 7.3.1 implies that

$$\operatorname{rank}(\mathbf{M}|_{\mathcal{R}(J)}) = \operatorname{rank}(\mathbf{M}|_{\mathcal{R}'(J)})$$

where $\mathcal{R}'(J)$ is the set of all subsets $T \subseteq [n]\setminus J$ with $|T| < 2\log k$. Thus we can certify whether a basis $\mathbf{M}_{T_1}, \mathbf{M}_{T_2}, \cdots, \mathbf{M}_{T_k}$ is a basis by, instead of computing the entire vector $\mathbf{C}^{T_i}|_{\mathcal{R}(J)}$, working with the much smaller vector $\mathbf{C}^{T_i}|_{\mathcal{R}'(J)}$, where as usual $J = \cup_i T_i$.

We remark that if $\mathcal{D}$ were not a mixture of subcubes, but a general mixture of product distributions, then we would need to look at $\mathbf{M}_T$ for $|T| \leq k-1$ in order to span the rows of $\mathbf{M}$. First this is necessary because we could set v to be a length k vector with k distinct entries in the range $[0,1]$. Now set each row of $\mathbf{m}$ to be v. In this example, the entrywise product of v with itself $k-1$ times is linearly independent of the vectors we get from taking the entrywise product between zero and $k-2$ times. On the other hand, this is tight:

Lemma 7.3.2. *Let $\mathcal{D}$ be a mixture of k product distributions and fix a realization. Let $\mathbf{M}$ be the corresponding moment matrix. Then*

$$\Big\{\mathbf{M}_T \Big| |T| < k\Big\}$$

span the rows of $\mathbf{M}$.

Proof. The proof is almost identical to the proof of Lemma 7.3.1. The only difference is that we allow $\lambda_i \in [0,1]$ and instead of reducing the size of R_t geometrically each time, we could reduce it by one. □

7.3.2 Local Maximality

In the following three subsections, we explain in greater detail how to produce a row basis for $\mathbf{M}$, as outlined in Sections 7.2.4 and 7.2.4. Recall that Lemma 7.2.9 and Lemma 7.2.10 give us a way to certify that the sets we are adding to $\mathcal{B}$ correspond to rows of $\mathbf{M}$ that are linearly independent of the ones we have selected so far. Motivated by these lemmas, we introduce the following key definitions:

Definition 7.3.3. *Given a collection $\mathcal{B} = \{T_1, T_2, \cdots, T_r\}$ of subsets we say that $\mathcal{B}$ is* certified full rank *if $\mathbf{C}|_{\mathcal{R}'(J)}^{T_1,T_2,\cdots,T_r}$ has full column rank, where $J = \cup_i T_i$.*

Note here we have used $\mathcal{R}'(J) = \{T \subseteq [n] \backslash J : |T| < 2 \log k\}$ with Lemma 7.3.1 in mind.

Definition 7.3.4. *Let $\mathcal{B} = \{T_1, T_2, \cdots, T_r\}$ be certified full column rank. Let $J = \cup_i T_i$. Suppose there is no*

(1) $T' \subseteq J$ or

(2) $T' = T_i \cup \{j\}$ for $j \notin J$

for which $\mathbf{C}|_{\mathcal{R}'(J')}^{T_1,T_2,\cdots,T_r,T'}$ has full column rank, where $J' = J \cup T'$. Then we say that $\mathcal{B}$ is locally maximal.

We are working towards showing that any certified full rank and locally maximal $\mathcal{B}$ spans a particular subset of the rows of $\mathbf{M}$. First we will show the following helper lemma:

Lemma 7.3.5. *Let $\mathcal{B} = \{T_1, T_2, \cdots, T_r\}$ and $J = \cup_i T_i$ as usual. Suppose that*

(1) the rows of $\mathbf{M}|_{\mathcal{B}}$ are a basis for the rows of $\mathbf{M}|_{2^J}$ and

(2) for any T_i and any $j \notin J$, the row $\mathbf{M}_{T_i \cup \{j\}}$ is in the row span of $\mathbf{M}|_{\mathcal{B}}$

Then the rows of $\mathbf{M}|_{\mathcal{B}}$ are a basis for the rows of $\mathbf{M}$.

Proof. We will proceed by induction. Suppose that the rows of $\mathbf{M}|_{\mathcal{B}}$ are a basis for the rows of $\mathbf{M}|_{2^{J'}}$ for some $J' \supseteq J$. Consider any $j \notin J'$. Then the rows

$$\mathbf{M}_{T_1}, \mathbf{M}_{T_2}, \cdots, \mathbf{M}_{T_r} \text{ and } \mathbf{M}_{T_1 \cup \{j\}}, \mathbf{M}_{T_2 \cup \{j\}}, \cdots, \mathbf{M}_{T_r \cup \{j\}}$$

are a basis for the rows of $\mathbf{M}|_{2^{J' \cup \{j\}}}$. But by assumption each row $\mathbf{M}_{T_i \cup \{j\}}$ is in the row span of $\mathbf{M}|_{\mathcal{B}}$. Thus the rows of $\mathbf{M}|_{\mathcal{B}}$ are also a basis for the rows of $\mathbf{M}|_{2^{J' \cup \{j\}}}$, as desired. □

Now we are ready to prove the main lemma in this subsection:

Lemma 7.3.6. *Let $\mathcal{D}$ have rank k and fix a full rank realization of $\mathcal{D}$. Let $\mathcal{B} = \{T_1, T_2, \cdots, T_r\}$ be certified full rank and locally maximal. Let $J = \cup_i T_i$ and*

$$K = \left\{ i \middle| i \notin J \textit{ and } \mathrm{rank}(\mathbf{M}|_{\mathcal{R}'(J \cup \{i\})}) = k \right\}$$

If $K \neq \emptyset$ then the rows of $\mathbf{M}|_{\mathcal{B}}$ are a basis for the rows of $\mathbf{M}|_{2^{J \cup K}}$.

Proof. Our strategy is to apply Lemma 7.3.5 to the set $J \cup K$ which will give the desired conclusion. To do this we just need to verify that the conditions in Lemma 7.3.5 hold. We will need to pay special attention to the distinction between $\mathcal{R}(J)$ and $\mathcal{R}'(J)$. First take any $i \in K$. Then

$$k = \mathrm{rank}(\mathbf{M}|_{\mathcal{R}'(J \cup \{i\})}) = \mathrm{rank}(\mathbf{M}|_{\mathcal{R}(J \cup \{i\})}) = \mathrm{rank}(\mathbf{M}|_{\mathcal{R}(J)})$$

The first equality follows from how we constructed K. The second equality follows from Lemma 7.3.1 when applied to the set $[n] \backslash J \cup \{i\}$. The third equality follows because the rows of $\mathbf{M}|_{\mathcal{R}(J \cup \{i\})}$ are a subset of the rows of $\mathbf{M}|_{\mathcal{R}(J)}$ and $\mathbf{M}$ has rank k.

Now the first condition of local maximality implies that there is no $T' \subseteq J$ where $\mathbf{C}|_{\mathcal{R}'(J)}^{T_1, T_2, \cdots, T_r, T'}$ has full column rank. Lemma 7.3.1 implies that $\mathbf{C}|_{\mathcal{R}(J)}^{T_1, T_2, \cdots, T_r, T'}$ also does not have full column rank because the additional rows of the latter can be obtained as linear combinations of the rows in the former. Now we can invoke Lemma 7.2.10 which implies that $\mathbf{M}_{T'}$ is in the span of $\mathbf{M}|_{\mathcal{B}}$. Thus the rows of $\mathbf{M}|_{\mathcal{B}}$ are indeed a basis for the rows of $\mathbf{M}|_{2^J}$, which is the first condition we needed to check.

For the second condition, the chain of reasoning is similar. Consider any $i \in K$ and any $T_{i'} \in \mathcal{B}$. Set $T' = T_{i'} \cup \{i\}$ and $J' = J \cup \{i\}$. Then $\mathrm{rank}(\mathbf{M}|_{\mathcal{R}'(J')}) = k$. Now the second condition of local maximality implies that $\mathbf{C}|_{\mathcal{R}'(J')}^{T_1, T_2, \cdots, T_r, T'}$ does not have full column rank. Lemma 7.3.1 implies that $\mathbf{C}|_{\mathcal{R}(J')}^{T_1, T_2, \cdots, T_r, T'}$ does not have full column rank either. We can once again invoke Lemma 7.2.10 which implies that $\mathbf{M}_{T'}$ is in the span of $\mathbf{M}|_{\mathcal{B}}$, which is the second condition we needed to verify. This completes the proof. $\square$

See Lemma 7.7.5 in Section 7.7.1 for the sampling noise-robust analogue of this.

7.3.3 Tracking Down an Impostor

First we give a name to a concept that is implicit in Lemma 7.3.6:

Definition 7.3.7. *Let $\mathcal{D}$ have rank k and fix a full rank realization of $\mathcal{D}$. Let $\mathcal{B} = \{T_1, T_2, \cdots, T_r\}$ be certified full rank and locally maximal. Let $J = \cup_i T_i$ and*

$$I = \left\{ i \middle| i \notin J \textit{ and } \operatorname{rank}(\mathbf{M}|_{\mathcal{R}'(J\cup\{i\})}) < k \right\}$$

We call I the set of impostors *and K the set of* non-impostors*.*

We emphasize that the notion of an impostor depends on a particular realization. If there are no impostors then Lemma 7.3.6 implies that the rows of $\mathbf{M}_{\mathcal{B}}$ are a basis for the rows of $\mathbf{M}$ and so we can directly use the algorithm outlined at the beginning of Section 7.2.4 to learn the parameters. If instead there is an impostor i we can condition on $x_S = s$ for $S = J \cup \{i\}$ and any $s \in \{0,1\}^{|S|}$ and get $\mathcal{D}|_{x_S=s}$ which by Lemma 7.2.7 and Lemma 7.2.8 is a mixture of strictly fewer than k subcubes. In particular, we can condition on $x_S = s$ for every $s \in \{0,1\}^{|S|}$, recursively learn these $2^{|S|}$ mixtures of strictly fewer than k subcubes in $\{0,1\}^{n\backslash S}$, estimate $\Pr_{x\sim\mathcal{D}}[x_S = s]$ for each s, and combine these mixtures into a single mixture over $\{0,1\}^n$ in the natural way (see Appendix 7.6 for details on this combining procedure).

But how do we find an impostor? It turns out that regardless of whether there exist impostors, we can still use the algorithm outlined at the beginning of Section 7.2.4 to learn a mixture of subcubes $\mathcal{D}'$ where either

(a) all the moments of $\mathcal{D}'$ up to size $c \log k$ are close to the true moments or

(b) there is a size at most $c \log k$ moment which is different, which in turn identifies a set S that is guaranteed to contain an impostor

And thus we will be able to make progress one way or the other. With this roadmap in hand, we can prove the main lemma in this subsection.

Lemma 7.3.8. *Let $\mathcal{D}$ have rank k and fix a full rank realization of $\mathcal{D}$. Let $\mathcal{B} = \{T_1, T_2, \cdots, T_r\}$ be certified full rank and locally maximal. Let $J = \cup_i T_i$. Let I be the set of impostors and K be the set of non-impostors.*

There is a guess $\mathbf{m}'|_J \in \{0, 1/2, 1\}^{|J|\times r}$ so that if we solve (7.1) and solve (7.2) for each $i \in K$ we get parameters that generate a mixture of subcubes $\mathcal{D}'$ on $J \cup K$ that satisfy $\mathbb{E}_{\mathcal{D}'}[x_S] = \mathbb{E}_{\mathcal{D}}[x_S]$ for all $S \subseteq J \cup K$.

Proof. For any $i \in K$ we have $\text{rank}(\mathbf{M}|_{\mathcal{R}'(J\cup\{i\})}) = k$. By Lemma 7.3.6 we know that $\mathbf{M}_{\mathcal{B}}$ is a row basis for $\mathbf{M}_{2^{J\cup K}}$. In particular $\text{rank}(\mathbf{M}_{2^{J\cup K}}) = r$. Thus using Lemma 7.2.7 there is a mixture of r subcubes with mixing weights π' and marginals matrix $\mathbf{m}' \in \{0, 1/2, 1\}^{|J\cup K|\times r}$ that realizes the same distribution as projecting $\mathcal{D}$ onto coordinates in $J \cup K$ (i.e. without conditioning on any coordinates outside of this set).

Let $\mathbf{M}'$ be the corresponding moment matrix. Then by construction $\mathbf{M}'$ consists of a subset of the columns of $\mathbf{M}_{2^{J\cup K}}$. Thus the rows of $\mathbf{M}'_{\mathcal{B}}$ still span the rows of $\mathbf{M}'$. Also by construction $\mathbf{M}'$ has rank r and hence the rows of $\mathbf{M}'_{\mathcal{B}}$ are linearly independent. Now if we take our guess to be $\mathbf{m}'|_J$ where $\mathbf{m}'$ is as above, (7.1) has a unique solution, namely π'. Also for each $i \in K$, (7.2) has a unique solution namely $\mathbf{m}'_i$. Now if we take our learned parameters we get a mixture of subcubes $\mathcal{D}'$ on $J \cup K$ that satisfies $\mathbb{E}_{\mathcal{D}'}[x_S] = \mathbb{E}_{\mathcal{D}}[x_S]$ for all $S \subseteq J \cup K$ because $\mathcal{D}'$ and projecting $\mathcal{D}$ onto coordinates in $J \cup K$ realize the same distribution. This completes the proof. □

See Lemma 7.7.6 in Section 7.7.2 for the sampling noise-robust analogue of this.

To connect this lemma to the discussion above, we will guess $\mathbf{m}'|_J \in \{0, 1/2, 1\}^{|J|\times r}$ and solve (7.1) and solve (7.2) for each $i \in [n]\backslash J$ (because we do not know the set of impostors). We can then check whether the parameters we get generate a mixture of subcubes $\mathcal{D}'$ that satisfies

$$\mathbb{E}_{\mathcal{D}'}[x_S] = \mathbb{E}_{\mathcal{D}}[x_S]$$

for all S with $|S| \leq c \log k$. If it does, then $\mathcal{D}' = \mathcal{D}$ and we are done. But if there is an S where the equation above is violated (and our guess was correct) then S cannot be a subset of $J \cup K$ which means that it contains an impostor. Thus the fact that we can check the equation above only up to logarithmic sized moments gives us a way to trace an impostor down to a logarithmic sized set, so that we can condition on $S \cup J$ and make progress without needing to fix too many coordinates.

Algorithm 22: N-List$(\mathcal{D}, k)$

```
Input: Mixture of subcubes $\mathcal{D}$, counter $k$
Output: Mixture of subcubes close to $\mathcal{D}$, or FAIL
if $k \le 0$ then
    return FAIL.
Run GrowByOne, which outputs either a certified full rank and locally maximal
  $\mathcal{B} = \{T_1, T_2, \cdots, T_r\}$, or FAIL and a set $J \subseteq [n]$.
if GrowByOne outputs FAIL and $J$ then
    Condition on $J$ by running N-List$(\mathcal{D}|_{x_J = s}, k-1)$ for all choices of $s \in \{0,1\}^{|J|}$.
    return the resulting distribution.
else
    $J \leftarrow \cup_i T_i$.
Initialize an empty list $L$ of candidate mixtures.
for $\mathbf{m}'|_J \subseteq \{0, 1/2, 1\}^{|J| \times r}$ do
    Solve (7.1) for $\pi' \in \Delta^r$.
    For each $i \notin J$, solve (7.2) for $\mathbf{m}'_i \in \{0, 1/2, 1\}^r$. If no such solution exists, skip
      to the next guess $\mathbf{m}'|_J$.
    If $\mathbf{M}'_S \cdot \pi' \neq \mathbb{E}_{\mathcal{D}}[x_S]$ for some $|S| \le 2\log(2k)$, then condition on $J \cup S$.
      Specifically, run N-List$(\mathcal{D}|_{x_{J \cup S} = s}, k-1)$ for all choices of $s \in \{0,1\}^{|J \cup S|}$,
      estimate $\Pr_{x \sim \mathcal{D}}[x_S = s]$ for all $s$, and combine the resulting mixtures into a
      single mixture over $\{0,1\}^n$. Add this mixture to $L$.
Run hypothesis selection on $L$ to find a distribution close to $\mathcal{D}$.
if distribution close to $\mathcal{D}$ exists then
    return this distribution.
else
    /* Every $i \notin J$ is an impostor */
    Select an arbitrary $i \notin J$ and condition on $J \cup \{i\}$ by running
      N-List$(\mathcal{D}|_{x_{J \cup \{i\}} = s}, k-1)$ for all choices of $s \in \{0,1\}^{|J \cup \{i\}|}$.
```

Again, we stress that while the algorithm as stated assumes access to the exact moments of $\mathcal{D}$, we show in the appendices how to lift this assumption entirely. Our final algorithm for learning mixtures of subcubes is actually Algorithm 24 (see Appendix 7.6) which invokes Algorithm 26 (see Appendix 7.7) as a subroutine.

As a final observation, if all of our guesses are correct, we would need to condition and recurse at most k times (because each time the number of components strictly decreases). So if ever we have too many recursive calls, we can simply terminate because we know that at least some guess along the way was incorrect. Algorithm 22 collects together all of these ideas into pseudocode and frames it as a non-deterministic algorithm for listing not too many candidate hypotheses, at least one of which will be close to a projection of $\mathcal{D}$. What remains is to implement GrowByOne to construct a certified full rank and locally maximal basis. Then we will move on to giving variants of our algorithm that work when we only have estimates of the moments (from random samples) and analyzing how the errors compound to give our full algorithm for learning mixtures of subcubes.

7.3.4 Finding a Certified Full Rank and Locally Maximal Set

It remains to implement Step 3 of N-List. $\mathbf{C}$ has 2^n columns, so it is not immediately clear how to efficiently find a set $\mathcal{B}$ of columns that is locally maximal certified full rank. We prove that it is always possible to greedily pick out an $\mathcal{B}$ such that either $\mathcal{B}$ is locally maximal certified full rank or $\text{rank}(\mathbf{M}|_{\mathcal{R}(J)}) < k$ for some rank-k realization of $\mathcal{D}$. If the latter happens and Step 10 of N-List fails, then Step 18 will succeed. Our greedy procedure GrowByOne is given in Algorithm 23.

When we assume exact access to the accessible entries of $\mathbf{C}$, the subroutine InSpan in GrowByOne is basic linear algebra. In the appendix, we show how to implement InSpan even if we only have estimates of the accessible entries of $\mathbf{C}$ up to some additive sampling error (see Algorithm 25 in Appendix 7.7).

Lemma 7.3.9. *If* GrowByOne *outputs FAIL and some set J^*, then* $\text{rank}(\mathbf{M}|_{\mathcal{R}'(J^*)}) < k$ *for some rank-k realization of $\mathcal{D}$. Otherwise,* GrowByOne *outputs $\mathcal{B}^* = \{T_1, \cdots, T_r\}$, and $\mathcal{B}^*$ is certified full rank and locally maximal.*

Proof. Set J^* either to be the output of GrowByOne if it outputs FAIL, or if it outputs $\mathcal{B}^*$ then set $J^* = \cup_i T_i$. Now fix any rank-k realization of $\mathcal{D}$ and let $\mathbf{M}$ be the corresponding moment matrix. Whenever the algorithm reaches Step 5 for some $i \in J^*$, $\mathcal{B} = \{T_1, \cdots, T_r\}$, there are two possibilities. If $\text{rank}(\mathbf{M}|_{\mathcal{R}'(J\cup\{i\})}) < k$, then $\text{rank}(\mathbf{M}|_{\mathcal{R}'(J^*)}) < k$ because J^*

Algorithm 23: GrowByOne($\mathcal{D}$)

```
Input: Mixture of subcubes $\mathcal{D}$
Output: Either $\mathcal{B} = \{T_1, \cdots, T_r\}$ such that $\mathcal{B}$ is certified full rank and locally
        maximal, or FAIL and some set $J$, in which case there is a rank-$k$
        realization of $\mathcal{D}$ for which $\mathrm{rank}(\mathbf{M}|_{\mathcal{R}(J)}) < k$.
$\mathcal{B} \leftarrow \{\emptyset\}$.
$J \leftarrow \emptyset$.
while True do
    for $i \notin J$ do
        $\mathcal{B}' \leftarrow \mathcal{B}$.
        for $T \in \mathcal{B}$ do
            Run InSpan($\mathcal{D}, \mathcal{B}', T \cup \{i\}$) to check if $\mathbf{C}|^{T\cup\{i\}}_{\mathcal{R}'(J\cup\{i\})}$ lies in the span of
              $\mathbf{C}|^{\mathcal{B}'}_{\mathcal{R}'(J\cup\{i\})}$.
            If so, add $T \cup \{i\}$ to $\mathcal{B}'$.
        $\mathcal{B} \leftarrow \mathcal{B}'$ and update $J$ to be the union of all elements of $\mathcal{B}$.
    If after trying all $i \notin J$, $\mathcal{B}$ remains unchanged, exit the loop.
for $S \subseteq J$ for which $S \notin \mathcal{B}$ do
    Run InSpan($\mathcal{D}, \mathcal{B}, S$) to check if $\mathbf{C}|^{S}_{\mathcal{R}'(J)}$ lies in the span of $\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J)}$.
    if exists S for which this is not the case then
        return FAIL.
    else
        return $\mathcal{B}$.
```

obviously contains $J \cup \{i\}$. Otherwise, inductively we know that $\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J\cup\{i\})}$ is a column basis for $\mathbf{C}^{2^J}_{\mathcal{R}'(J\cup\{i\})}$, so by Lemma 7.2.9 and Lemma 7.2.10, $\mathbf{M}|_{\mathcal{B}}$ is a row basis for $\mathbf{M}|_{2^J}$. So rows

$$T_1, \cdots, T_r, T_1 \cup \{i\}, \cdots, T_r \cup \{i\}$$

of $\mathbf{M}$ span the rows of $\mathbf{M}|_{2^{J\cup\{i\}}}$. By Lemma 7.2.9, columns

$$T_1, \cdots, T_r, T_1 \cup \{i\}, \cdots, T_r \cup \{i\}$$

of $\mathbf{C}|_{\mathcal{R}'(J\cup\{i\})}$ thus span the columns of $\mathbf{C}|^{2^{J\cup\{i\}}}_{\mathcal{R}'(J\cup\{i\})}$. Step 8 of GrowByOne simply finds a basis for these columns.

Thus when we exit the loop, either (a) $\mathcal{B}^*$ indexes a column basis for $\mathbf{C}|^{2^{J^*}}_{\mathcal{R}'(J^*)}$ or (b) at some iteration of Step 3 J satisfies $\text{rank}(\mathbf{M}|_{\mathcal{R}'(J)}) < k$ and thus $\text{rank}(\mathbf{M}|_{\mathcal{R}'(J^*)}) < k$.

If (a) holds GrowByOne will reach Step 16 and output $\mathcal{B}^*$. The fact that $\mathcal{B}^*$ is a column basis implies that $\mathcal{B}^*$ is certified full rank and, together with the exit condition in Step 10, that it is also locally maximal. On the other hand, if GrowByOne terminates at Step 13, we know that (b) holds, so it successfully outputs FAIL together with J^* satisfying $\text{rank}(\mathbf{M}|_{\mathcal{R}(J^*)}) < k$. □

See Lemma 7.7.4 in Section 7.7.1 for the sampling noise-robust analogue of this.

7.3.5 Sampling Noise and Small Mixture Weights

It remains to show that N-List works even when it only has access to the entries of $\mathbf{C}$ up to sampling noise $\varepsilon_{\text{samp}}$. We defer most of the details to the appendix but present here the crucial ingredients that ensure sampling noise-robust analogues of the above lemmas still hold.

We first need to show that $\mathbf{M}$ and $\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J)}$ are well-conditioned. Because the entries of these matrices are $[0,1]$-valued and thus have bounded Frobenius norm, it's enough to bound their minimal singular values. For our purposes, it will be more convenient to bound $\sigma^{\infty}_{\min}(A) \triangleq \min_x \|Ax\|_\infty / \|x\|_\infty$ for $A = \mathbf{M}, \mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J)}$.

Lemma 7.3.10. *Take any realization of $\mathcal{D}$ with moment matrix* $\mathbf{M}$ *such that* $\mathbf{M}$ *is full-rank and* $\operatorname{rank}(\mathbf{M}) = k$. *For* $d \geq 2\log k$, *let* M *be any subset of the rows of* $\mathbf{M}$ *with full column rank and which are all entrywise products of fewer than* d *rows of* $\mathbf{m}$. *Then* $\sigma^{\infty}_{\min}(M) \geq 2^{-O(dk)} \cdot k^{-O(k)}$.

In particular, for $d = 2\log k$, *there exists an absolute constant* $c_{13} > 0$ *for which* $\sigma^{\infty}_{\min}(M) \geq k^{-c_{13}k}$. *For* $d = k$, *there exists an absolute constant* $c_{14} > 0$ *for which* $\sigma^{\infty}_{\min}(M) \geq 2^{-c_{14}k^2}$.

Proof. Because adding rows will simply increase $\sigma^{\infty}_{\min}$, assume without loss of generality that M is $k \times k$. We show that the largest entry of M^{-1} is at most $2^{O(dk)} \cdot k^{O(k)}$.

Note that the entries of M take values among $\{0, 1, 1/2, 1/4, ..., 1/2^{d-1}\}$. The determinant of any $(k-1) \times (k-1)$ minor is at most $(k-1)! \sim k^{O(k)}$, while $\det(M)$ is some nonzero integral multiple of $1/2^{(d-1)k}$, so by Cramer's we obtain the desired bound on the largest entry of M^{-1}. □

Lemma 7.3.10 allows us to prove the following robust low-degree identifiability lemma, which says that mixtures of subcubes which agree on all $O(\log k)$-degree moments are close in total variation distance.

Lemma 7.3.11. *Let* $\mathcal{D}_1, \mathcal{D}_2$ *be mixtures of* k *subcubes in* $\{0,1\}^n$ *with mixing weights* π^1 *and* π^2 *and moment matrices* $\mathbf{M}_1$ *and* $\mathbf{M}_2$ *respectively. If* $d_{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$, *there is some* S *for which* $|S| < 2\log(k_1 + k_2)$ *and* $|\mathbb{E}_{\mathcal{D}_1}[x_S] - \mathbb{E}_{\mathcal{D}_2}[x_S]| > \varepsilon \cdot k^{-c_{15}k}$ *for an absolute constant* $c_{15} > 0$.

For convenience, define $k = k_1 + k_2$ and $d = 2\log k$. First observe that the largest moment discrepancies $\max_{S:|S|<d} |\mathbb{E}_{\mathcal{D}_1}[x_S] - \mathbb{E}_{\mathcal{D}_2}[x_S]|$ can be interpreted as follows. Denote the moment matrices of $\mathcal{D}_1$ and $\mathcal{D}_2$ by $\mathbf{M}_1$ and $\mathbf{M}_2$. Define $\mathbf{N}$ to be the $2^{\binom{n}{d}} \times (k)$ matrix $\Big((\mathbf{M}_1)_{<d} \| (\mathbf{M}_2)_{<d}\Big)$ where $(\mathbf{M}_i)_{<d}$ denotes rows of $\mathbf{M}_i$ each given by entrywise products of fewer than d rows of $\mathbf{m}$. Define $\pi \in \mathbb{R}^k$ to be $(\pi^1 \| - \pi^2)$. Note that because $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) > 0$, Lemma 7.3.1 implies that their degree d-moments cannot all be identical, i.e. $\pi \notin \ker(\mathbf{N})$.

Denote the $2^n \times k$ concatenation of the distribution matrices of $\mathcal{D}_1$ and $\mathcal{D}_2$ by $\mathbf{D}$ and observe that we have chosen d so that the rows of $\mathbf{N}$ span those of $\mathbf{D}$ by the proof of

Lemma 7.3.1. Then it is easy to check that

$$\max_{S:|S|<d} |\mathbb{E}_{\mathcal{D}_1}[x_S] - \mathbb{E}_{\mathcal{D}_2}[x_S]| = \|\mathbf{N}\pi\|_\infty.$$

Lemma 7.3.12. *For any $v \in \ker(\mathbf{N})$, $\|\pi + v\|_\infty > \varepsilon/k$.*

Proof. Suppose to the contrary there existed a $v \in \ker(\mathbf{N})$ for which $\|\pi+v\|_\infty \le \varepsilon/k$. Denote $\pi + v$ by π' and the $2^n \times k$ concatenation of the distribution matrices of $\mathcal{D}_1$ and $\mathcal{D}_2$ by $\mathbf{D}$ again. We have that

$$d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) = \|\mathbf{D}\pi\|_1 = \|\mathbf{D}(v - \pi')\|_1 \le \|\mathbf{D}v\|_1 + \|\mathbf{D}\pi'\|_1.$$

But note that because the row spans of $\mathbf{N}$ and $\mathbf{D}$ agree, $v \in \ker(\mathbf{D})$, so $\|\mathbf{D}v\|_1 = 0$. Moreover,

$$\|\mathbf{D}\pi'\|_1 \le \sum_{j=1}^{k} \|\pi'_j \mathbf{D}^j\|_1 = \|\pi'\|_1 \le \varepsilon,$$

where the equality follows from the fact that each column of $\mathbf{D}$ sums to 1 because $\mathbf{D}$ is a distribution matrix. Contradiction! □

Proof of Lemma 7.3.11. Suppose $\mathbf{N}$ is of rank r, and columns $i_1, ..., i_r$ form a basis for its column space. Pick $v \in \ker(\mathbf{N})$ for which $\pi + v$ is supported only on coordinates $i_1, .., i_r$ so that $\mathbf{N}\pi = \mathbf{N}^{\{i_1,...,i_r\}}(\pi + v)$. Then

$$\|\mathbf{N}\pi\|_\infty \ge \sigma^\infty_{min}(\mathbf{N}^{\{i_1,...,i_r\}}) \cdot \|\pi + v\|_\infty > \frac{\varepsilon}{k} \cdot \sigma^\infty_{min}(\mathbf{N}^{\{i_1,...,i_r\}}). \tag{7.3}$$

Observe that $\sigma^\infty_{min}(\mathbf{N}^{\{i_1,...,i_r\}}) = \sigma^\infty_{min}(M)$ where M is the submatrix of $(\mathbf{M}_1 \| \mathbf{M}_2)$ given by columns $i_1, ..., i_r$. But M is a full-rank moment matrix of a mixture of at most k $\{0, 1/2, 1\}$-product distributions, so by (7.3) and Lemma 7.3.10, we have

$$\|\mathbf{N}\pi\|_\infty \ge \frac{\varepsilon}{k} \cdot k^{-c_{13}k} \ge \varepsilon \cdot k^{-c_{15}k}$$

as desired. □

For example, Lemma 7.3.11 tells us that in step 3a) of N-LIST, if $\mathcal{B}$ indexes a basis for the rows of $\mathbf{M}$ but we only have $\mathbb{E}_{\mathcal{D}}[x_S]$ up to $\varepsilon_{\text{samp}}$ sampling noise for every $S \in \mathcal{B}$, it's enough to run an L_∞ regression on the system (7.1) to get good approximations to the mixture weights π, as long as $\varepsilon_{\text{samp}} \le 2^{-c_{13}k^2} \cdot \varepsilon$.

The condition number bound on $\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J)}$ is a bit more subtle. By Observation 7.2.5,

$$\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(I)} = \mathbf{M}|_{\mathcal{R}'(I)} \cdot \text{diag}(\pi) \cdot (\mathbf{M}|_{\mathcal{B}})^\top$$

for any mixing weights π and moment matrix $\mathbf{M}$ realizing $\mathcal{D}$, so if π contains small entries, the condition number bound we want doesn't hold a priori. This is unsurprising: if a mixture $\mathcal{D}$ has a subcube with negligible mixture weight, our algorithm shouldn't be able to distinguish between $\mathcal{D}$ and the mixture obtained by removing that subcube and renormalizing the remaining mixture weights.

The upshot, it would seem, is that if $\mathbf{C}$ is badly conditioned because of small mixture weights, we might as well pretend we never see samples from the corresponding subcubes. Unfortunately, to get the desired level of precision in our learning algorithm, we will end up taking enough samples that we will see samples from those rarely occurring product distributions.

The key insight is that if there exist mixture weights small enough that omitting the corresponding subcubes and renormalizing the remaining mixture weights yields a distribution $\mathcal{D}'$ for which $d_{\text{TV}}(\mathcal{D}, \mathcal{D}') \le O(\varepsilon)$, then $\mathbf{C}$ morally behaves as if it had rank equal not to k, but to $\text{rank}(\mathbf{M}')$ where $\mathbf{M}'$ is the moment matrix for some realization of $\mathcal{D}'$. We then just need that all other mixing weights are not too small in order for $\tilde{\mathbf{C}}_{\mathcal{D}'}$ to be well-conditioned.

Definition 7.3.13. *Mixing weights π and marginals matrix $\mathbf{m}$ constitute a $[\tau_{small}, \tau_{big}]$-avoiding realization* of $\mathcal{D}$ *if $\pi^i \notin [\alpha, \beta]$ for all i.*

By a standard windowing argument, it will be enough to consider $\mathcal{D}$ which have $[\tau_{small}, \tau_{big}]$-avoiding realizations for some thresholds $0 < \tau_{small} < \tau_{big} < 1$. Let $\tau_{small} = \rho \cdot \tau_{big}$ where $\rho \triangleq k^{-c_{16}k^2}$ for some large absolute constant $c_{16} > 0$ to be specified later.

Below, given a moment matrix $\mathbf{M}$ with corresponding mixture weights π, we will denote by $\mathbf{M}'$ the subset of columns i of $\mathbf{M}$ for which $\pi^i > \tau_{big}$.

Lemma 7.3.14. *Let π and $\mathbf{M}$ be the mixing weights and moment matrix of a $[\tau_{small}, \tau_{big}]$-avoiding rank-$k$ realization of $\mathcal{D}$, and denote the number of columns of $\mathbf{M}'$ by k'. Let $\mathcal{B}$ be any collection of $r \le k'$ columns of $\mathbf{C}$ for which the corresponding r rows of $\mathbf{M}'|_{\mathcal{B}}$ are linearly independent, $J = \cup_{T\in\mathcal{B}} T$ satisfies $|J| \le k'$, and $\mathrm{rank}(\mathbf{M}'|_{\mathcal{R}'(J)}) = k'$. Then $\sigma^{\infty}_{\min}(\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J)}) \ge k^{-c_{17}k^2}\tau_{big}$ for some sufficiently large constant c_{17}.*

In particular, for any $\tilde{E}$ for which $\|\tilde{E} - \mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J)}\|_{\max} \le \frac{1}{2} \cdot k^{-c_{17}k^2-1}\tau_{big}$, we have that $\sigma^{\infty}_{\min}(\tilde{E}) \ge \frac{1}{2} \cdot k^{-c_{17}k^2}\tau_{big}$.

Proof. Because $\mathbf{M}$ is full-rank, $\mathbf{M}'$ is full-rank. Pick out a collection $\mathcal{R}^* \subseteq \mathcal{R}'(J)$ of k' row indices for which $\mathbf{M}'|_{\mathcal{R}^*}$ is still of rank k'. Obviously $\sigma^{\infty}_{\min}(\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J)}) \ge \sigma^{\infty}_{\min}(\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}^*})$.

Note that we have the decomposition

$$\begin{aligned}\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}^*} &= \mathbf{M}|_{\mathcal{R}^*} \cdot \mathrm{diag}(\pi) \cdot (\mathbf{M}_{\mathcal{B}})^\top \\ &= \mathbf{M}'|_{\mathcal{R}^*} \cdot \mathrm{diag}(\pi^1, ..., \pi^{k'}) \cdot (\mathbf{M}'|_{\mathcal{B}})^\top + \mathbf{M}|^{\{k'+1,...,k\}}_{\mathcal{R}^*} \cdot \mathrm{diag}(\pi^{k'+1}, ..., \pi^k) \cdot (\mathbf{M}|^{\{k'+1,...,k\}}_{\mathcal{B}})^\top.\end{aligned}$$

We know $\mathrm{diag}(\pi^{k'+1}, ..., \pi^k) \le \tau_{small}$ by assumption.

We already know $\mathbf{M}'|_{\mathcal{R}^*}$ is full-rank, and $(\mathbf{M}'|_{\mathcal{B}})^\top$ has linearly independent columns by assumption. So by Lemma 7.3.10, $\sigma_{\min}(\mathbf{M}'|_{\mathcal{R}^*}) \ge k^{-c_{13}k}, \sigma_{\min}((\mathbf{M}'|_{\mathcal{B}})^\top) \ge 2^{-c_{14}k^2}$, and because $\sigma^{\infty}_{\min}$ is super-multiplicative,

$$\sigma^{\infty}_{\min}\left(\mathbf{M}'|_{\mathcal{R}^*} \cdot \mathrm{diag}(\pi^1, ..., \pi^{k'}) \cdot (\mathbf{M}'|_{\mathcal{B}})^\top\right) \ge 2^{-c_{18}k^2}\pi^{k'}$$

for some constant $c_{18} > 0$. On the other hand,

$$\|\mathbf{M}|^{\{k'+1,...,k\}}_{\mathcal{R}^*} \cdot \mathrm{diag}(\pi^{k'+1}, ..., \pi^k) \cdot (\mathbf{M}|^{\{k'+1,...,k\}}_{\mathcal{B}})^\top\|_\infty \le (k-k')^2 \cdot \pi^{k'+1}$$

by super-mutiplicativity of the L^∞ norm. So we conclude that

$$\sigma^{\infty}_{\min}(N) \ge 2^{-2c_{18}k^2}\pi^{k'} - (k-k')^2 \cdot \pi^{k'+1} \ge k^{-c_{17}k^2}\tau_{big}$$

for some $c_{17} > c_{18}$, where the second inequality follows from the fact that $\pi^{k'+1} \le \tau_{small} < k^{-c_{16}k^2} \cdot \tau_{big} \le k^{-c_{16}k^2}\pi^{k'}$ for sufficiently large $c_{16} > 0$.

The last part of the lemma just follows by the triangle inequality. □

In Appendix 7.7, we use Lemmas 7.3.10 and 7.3.14 to prove analogues of the key lemmas in the preceding sections when we drop the assumption of zero sampling noise.

7.4 An $n^{\Omega(\sqrt{k})}$ Statistical Query Lower Bound

In this section we prove the following unconditional lower bound for statistical query learning mixtures of product distributions.

Theorem 7.4.1. *Let $\varepsilon < (2k)^{-\sqrt{k}}/4$. Any SQ algorithm with SQ access to a mixture of k product distributions $\mathcal{D}$ in $\{0,1\}^n$ and which outputs a distribution $\overline{\mathcal{D}}$ with $d_{TV}(\mathcal{D},\overline{\mathcal{D}}) \le \varepsilon$ requires at least $\Omega(n/k)^{\sqrt{k}}$ calls to $STAT(\Omega(n^{-\sqrt{k}/3}))$ or $VSTAT(O(n^{\sqrt{k}/3}))$.*

7.4.1 Statistical Query Learning of Distributions

In this subsection we review basic notions about statistical query (SQ) learning. Introduced in [Kea98], SQ learning is a restriction of PAC learning [Val84] to the setting where the learner has access to an oracle that answers statistical queries about the data, instead of access to the data itself. In [FGR$^+$17], this model was extended to learning of distributions, where for our purposes of learning distributions over $\{0,1\}^n$ the relevant SQ oracles are defined as follows:

Definition 7.4.2. *Fix a distribution $\mathcal{D}$ over $\{0,1\}^n$. For tolerance parameter $\tau > 0$, the $STAT(\tau)$ oracle answers any query $h : \{0,1\}^n \to [-1,1]$ with a value v such that*

$$|\mathbb{E}_{x\sim\mathcal{D}}[h(x)] - v| \le \tau.$$

For sample size parameter $t > 0$, the $VSTAT(t)$ oracle answers any query $h : \{0,1\}^n \to [0,1]$ with a value v for which

$$|\mathbb{E}_{x\sim\mathcal{D}}[h(x)] - v| \le \max\left\{\frac{1}{t}, \sqrt{\frac{\mathbb{V}_{x\sim\mathcal{D}}[h(x)]}{t}}\right\}$$

The prototypical approach to proving unconditional SQ lower bounds is by bounding the SQ dimension of the concept class, defined in [BFJ+94] for learning Boolean functions and extended in [FGR+17] to learning distributions.

Definition 7.4.3. *Let $\mathcal{D}$ be a class of distributions over $\{0,1\}^n$ and $\mathcal{F}$ be a set of* solution distributions *over $\{0,1\}^n$. For any map $\mathcal{Z}: \mathcal{D} \to 2^{\mathcal{F}}$, the* distributional search problem $\mathcal{Z}$ *over $\mathcal{D}$ and $\mathcal{F}$ is to find some $f \in \mathcal{Z}(\mathcal{D})$ given some form of access to $\mathcal{D} \in \mathcal{D}$.*

Definition 7.4.4. *Let U be a distribution over $\{0,1\}^n$ whose support S contains the support of distributions $\mathcal{D}_1, \mathcal{D}_2$. Then*

$$\chi_U(\mathcal{D}_1, \mathcal{D}_2) \triangleq -1 + \sum_{x \in S} \frac{\mathcal{D}_1(x)\mathcal{D}_2(x)}{U(x)}$$

is the pairwise correlation *of $\mathcal{D}_1, \mathcal{D}_2$ with respect to U. When $\mathcal{D}_1 = \mathcal{D}_2$, the pairwise correlation is merely the χ^2-divergence between $\mathcal{D}_1$ and U, denoted $\chi^2(\mathcal{D}_1, U) = -1 + \sum_{x\in S} \mathcal{D}_1(x)^2/U(x)$.*

Definition 7.4.5. *A set of distributions $\mathcal{D}_1, ..., \mathcal{D}_m$ over $\{0,1\}^n$ is (γ,β)-correlated relative to distribution U over $\{0,1\}^n$ if*

$$|\chi_U(\mathcal{D}_i, \mathcal{D}_j)| \le \begin{cases} \gamma, & i \neq j \\ \beta, & i = j. \end{cases}$$

Definition 7.4.6. *For $\beta, \gamma > 0$ and a distributional search problem $\mathcal{Z}$ over $\mathcal{D}$ and $\mathcal{F}$, the* SQ dimension *$SD(\mathcal{Z},\gamma,\beta)$ is the maximum d for which there exists a reference distribution U over $\{0,1\}^n$ and distributions $\mathcal{D}_1, ..., \mathcal{D}_m \in \mathcal{D}$ such that for any $\mathcal{D} \in \mathcal{F}$, the set $\mathcal{D}_f$ of $\mathcal{D}_i$ outside of $\mathcal{Z}^{-1}(\mathcal{D})$ is of size at least d and is (γ,β)-correlated relative to U.*

Lemma 7.4.7 (Corollary 3.12 in [FGR+17])**.** *For $\gamma' > 0$ and $\mathcal{Z}$ a distributional search problem $\mathcal{Z}$ over $\mathcal{D}$ and $\mathcal{F}$, any SQ algorithm for $\mathcal{Z}$ requires at least $SD(\mathcal{Z},\gamma,\beta)\cdot\gamma'/(\beta-\gamma)$ queries to $STAT(\sqrt{\gamma+\gamma'})$ or $VSTAT(1/3(\gamma+\gamma'))$.*

In our setting, $\mathcal{D}$ is the set of mixtures of product distributions over $\{0,1\}^n$, $\mathcal{F}$ is the set of all distributions over $\{0,1\}^n$, and $\mathcal{Z}$ sends any mixture $\mathcal{D}$ to the set of all distributions over

$\{0,1\}^n$ which are ε-close to $\mathcal{D}$ in total variation distance, and distributional search problem is to recover any such distribution given sample access to $\mathcal{D}$. Our approach will thus be to bound the SQ dimension of $\mathcal{Z}$ for appropriately chosen β, γ.

7.4.2 Embedding Interesting Coordinates

The SQ lower bound instance for mixtures of subcubes given in [FOS05] is the class of all k-leaf decision trees over $\{0,1\}^n$. The SQ lower bound for learning k-leaf decision trees stems from the SQ lower bound for learning $\log k$-sparse parities, for which the idea is that U_n and the uniform distribution over positive examples of $\log k$-sparse parity agree on all moments of degree less than $\log k$ and differ on exactly one moment of degree $\log k + 1$, corresponding to the coordinates of the parity. The observation that leads to our SQ lower bound is that for general mixtures of k product distributions, we can come up with much harder instances which agree with U_n even on moments of degree at most $O(\sqrt{k})$.

We begin with a mixture A of k product distributions in $\{0,1\}^m$, for appropriately chosen $m < n$, whose moments of degree at most $m-1$ are exactly equal to those of U_m, but whose m-th moment differs (we construct such an A in the next section). We then pick a subset of "interesting coordinates" $I \subseteq [n]$ of size m and embed A into U_n on those coordinates in the same way we would embed a sparse parity into U_n. Formally, we have the following construction, which is reminiscent of the blueprint for proving SQ lower bounds for learning sparse parities [Kea98] and mixtures of Gaussians [DKS17]:

Definition 7.4.8 (High-dimensional hidden interesting coordinates distribution)**.** *Let A be a mixture of k product distributions with mixing weights $\pi \in \mathbb{R}^k$ and marginals matrix $\mathbf{m} \in [0,1]^{m\times k}$. For $I \subseteq [n]$, define $\mathcal{D}_I$ to be the mixture of k product distributions in $\{0,1\}^m$ with mixing weights π and marginals matrix $\mathbf{m}^* \in [0,1]^{n\times k}$ defined by $\mathbf{m}^*|_I = \mathbf{m}|_I$ and $(\mathbf{m}^*)^j_i = 1/2$ for all $i \notin I$ and $j \in [k]$. In other words, $\mathcal{D}_I$ is the product distribution $A \times U_{[n]\backslash I}$ where $U_{[n]\backslash I}$ is the uniform distribution over coordinates $[n]\backslash I$.*

Remark 7.4.9. *In fact, we have much more flexibility in our lower bound construction. We can construct a mixture A matching moments with any single product distribution and embed it in any single product distribution over $\{0,1\}^n$ whose marginals in coordinates I agree with* *those of A, but for transparency we will focus on U_n.*

Let $\delta(A) = A(1^m) - 1/2^m$. A and U_m only disagree on their top-degree moment, and $\delta(A)$ is simply the extent to which they differ on this moment. The following simple fact will be useful in proving correlation bounds.

Observation 7.4.10. *If A and U_m agree on all moments of degree less than m, then $A(x) = 1/2^m + (-1)^{z(x)}\delta(A)$, where $z(x)$ is the number of zero bits in x.*

The main result of this section is

Proposition 7.4.11. *Fix n. Suppose there exists an $m \in \mathbf{Z}_+$ and distribution A on $\{0,1\}^m$ such that $A$and U_m agree on all moments of degree less than m, and consider the set of distributions $\{\mathcal{D}_I\}_{I\subseteq[n],|I|=m}$. Let $\varepsilon < \delta(A)\cdot 2^{m-2}$. Any SQ algorithm which, given an SQ oracle for some $\mathcal{D}_I$, outputs a distribution $\mathcal{D}$ for which $d_{TV}(\mathcal{D},\mathcal{D}_I) \le \varepsilon$ requires at least $\Omega(n)^{m/3}/\delta(A)^2$ queries to STAT$(\Omega(n^{-m/3})$ or VSTAT$(O(n^{m/3}))$.*

To invoke Lemma 7.4.7 to prove Proposition 7.4.11, we need to prove correlation bounds on the set of distributions $\{\mathcal{D}_I\}_{I\subseteq[n],|I|=m}$.

Lemma 7.4.12. *Suppose A and U_m agree on all moments of degree less than m. For distinct $I, J \subseteq [n]$ of size m, $\chi_{U_n}(\mathcal{D}_I,\mathcal{D}_J) = 0$.*

Proof. Let $S = I\cap J$, $T = [n]\backslash(I\cup J)$, $I' = I\backslash S$, and $J' = J\backslash S$. Decompose any $x \in \{0,1\}^n$ as $x_T \circ x_S \circ x_{I'} \circ x_{J'}$ in the natural way. We can write

$$\begin{aligned}
1 + \chi_{U_n}(\mathcal{D}_I,\mathcal{D}_J) &= 2^n \cdot \sum_{x\in\{0,1\}^n} \frac{A(x_I)}{2^{n-m}}\cdot\frac{A(x_J)}{2^{n-m}} \\
&= 2^{|S|}\cdot \sum_{x_S,x_{I'},x_{J'}} A(x_{I'\cup S})\cdot A(x_{J'\cup S}) \qquad (7.4)
\end{aligned}$$

For fixed x_S it is easy to see that

$$\sum_{x_{I'},x_{J'}} A(x_{I'\cup S})\cdot A(x_{J'\cup S}) = 2^{2m-2|S|-2}\left(\frac{1}{2^m}+\delta(A)+\frac{1}{2^m}-\delta(A)\right)^2 = 2^{-2|S|},$$

so (7.4) reduces to 1 and the claim follows. □

Lemma 7.4.13. *Suppose A and U_m agree on all moments of degree less than m. Then $\chi^2(\mathcal{D}_I, U_n) = \delta(A)^2 \cdot 4^m$.*

Proof. Decompose any $x \in \{0,1\}^n$ as $x_{I^c} \circ x_I$. Then

$$\begin{aligned}
1 + \chi^2(\mathcal{D}_I, U_n) &= 2^n \sum_{x \in \{0,1\}^n} \frac{A(x_I)^2}{2^{2n-2m}} \\
&= 2^m \sum_{x_I \in \{0,1\}^m} A(x_I)^2 \\
&= 2^m \left(\sum_{x_I : z(x_I) \text{ even}} \left(\frac{1}{2^m} + \delta(A)\right)^2 + \sum_{x_I : z(x_I) \text{ odd}} \left(\frac{1}{2^m} - \delta(A)\right)^2 \right) \\
&= 2^m \cdot 2^{m-1} \cdot \left(\frac{1}{2^{2m-1}} + 2\delta(A)^2\right) = 1 + \delta(A)^2 \cdot 4^m,
\end{aligned}$$

and the claim follows. □

Lemma 7.4.14. *Suppose A and U_m agree on all moments of degree less than m. For distinct $I, J \subseteq [n]$ of size m, $d_{TV}(\mathcal{D}_I, \mathcal{D}_J) = \delta(A) \cdot 2^{m-1}$.*

Proof. For any $x \in \{0,1\}^n$, $\mathcal{D}_I(x) = \frac{1}{2^{n-m}} \cdot A(x_I) = \frac{1}{2^n} + \frac{1}{2^{n-m}}(-1)^{z(x_I)}\delta$ by Observation 7.4.10. So $|\mathcal{D}_I(x) - \mathcal{D}_J(x)|$ is zero if $z(x_I)$ and $z(x_J)$ are the same parity, and $\delta(A)/2^{n-m-1}$ otherwise. When I and J are distinct, the probability that $z(x_I)$ and $z(J)$ are of different parities for $x \sim U_n$ is $1/2$, so

$$d_{\mathrm{TV}}(\mathcal{D}_I, \mathcal{D}_J) = \frac{1}{2} \sum_{x \in \{0,1\}^n} |\mathcal{D}_I(x) - \mathcal{D}_J(x)| = \frac{1}{2} \cdot \frac{\delta(A)}{2^{n-m-1}} \cdot 2^{n-1} = \delta(A) \cdot 2^{m-1}$$

as desired. □

Proof of Proposition 7.4.11. Given unknown $\mathcal{D}_I$, the distributional search problem $\mathcal{Z} : \mathcal{D} \to 2^{\mathcal{F}}$ is to find any distribution $\mathcal{D} \in \mathcal{F}$ for which $d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}_I) \le \varepsilon$, where $\mathcal{D} = \{\mathcal{D}_I\}_{I \subseteq [n], |I| = m}$ and $\mathcal{F}$ is the set of all distributions over $\{0,1\}^n$. Because we assume in Proposition 7.4.11 that $\varepsilon < \delta(A) \cdot 2^{m-2}$ and we know by Lemma 7.4.14 that $d_{\mathrm{TV}}(\mathcal{D}_I, \mathcal{D}_J) = \delta(A) \cdot 2^{m-1} > 2\varepsilon$ for any $J \ne I$, we see that for any $\mathcal{D} \in \mathcal{F}$, $\mathcal{Z}^{-1}(\mathcal{D})$ is just $\mathcal{D}_I$. So in the language of Definition 7.4.6, $\mathcal{D}_f$ consists of all $\mathcal{D}_J$ for $J \ne I$. Moreover, by Lemmas 7.4.12 and 7.4.13,

$\mathcal{D}_f$ is $(0, \delta(A)^2 \cdot 4^m)$-correlated. So $\mathrm{SD}(\mathcal{Z}, 0, \delta(A)^2 \cdot 4^m) = \binom{n}{m} - 1$. Applying Lemma 7.4.7, we conclude that for any $\gamma' > 0$, the number of queries to $\mathrm{STAT}(\sqrt{\gamma'})$ or $\mathrm{VSTAT}(1/3\gamma')$ to solve $\mathcal{Z}$ is at least

$$\frac{(\binom{n}{m} - 1) \cdot \gamma'}{\delta(A)^2 \cdot 4^m} = \frac{\Omega(n)^m \cdot \gamma'}{\delta(A)^2}.$$

We're done when we take $\gamma' = 1/O(n)^{-2m/3}$. □

7.4.3 A Moment Matching Example

It remains to construct for some $m \in \mathbf{Z}_+$ a distribution A over $\{0,1\}^m$ for which A and U_m agree only on moments of degree less than m and obtain bounds on $\delta(A)$.

Definition 7.4.15. *Given $\pi \in \Delta^k$, a collection of vectors $v_1, ..., v_m \in \mathbb{R}^k$ is d-*wise superorthogonal *with respect to π if for any $S \subseteq [m]$ of size at most d, $\langle \bigodot_{i \in S} v_i, \pi \rangle = 0$. Note that if $\pi = \frac{1}{k} \cdot \mathbf{1}$ and $d = 2$, this is just the usual notion of orthogonality.*

Lemma 7.4.16. *Let $d \leq m$ and suppose A is a mixture of product distributions with mixing weights π and marginals matrix $\mathbf{m}$. Then A and U_m agree on moments of degree at most d if and only if the rows of $\mathbf{m} - \frac{1}{2} \cdot \mathbf{J}_{m \times k}$ are d-wise superorthogonal with respect to π, where $\mathbf{J}_{m \times k}$ is the $m \times k$ all-ones matrix.*

Proof. For any $S \subseteq [m]$ of size at most d,

$$\left\langle \bigodot_{i \in S} \left(\mathbf{m}_i - \frac{1}{2} \cdot \mathbf{1} \right), \pi \right\rangle = \sum_{T \subseteq S} (-1/2)^{|S|-|T|} \langle \mathbf{M}_T, \pi \rangle. \tag{7.5}$$

So if A and U_m agree on moments of degree at most d so that $\langle \mathbf{M}_T, \pi \rangle = 1/2^{|T|}$ for all $|T| \leq d$, this is equal to $(1/2)^{|S|} \cdot \sum_{T \subseteq S} (-1)^{|S|-|T|} = 0$. Conversely, if the rows of $\mathbf{m} - \frac{1}{2} \cdot \mathbf{J}_{m \times k}$ are indeed d-wise superorthogonal with respect to π, then by induction on degree, the fact that (7.5) vanishes forces $\langle \mathbf{M}_S, \pi \rangle$ to be $2^{|S|}$. □

Because we insist that A and U_m agree on their moments of degree less than m and differ on their m-th moment, Lemma 7.4.16 reduces the task of constructing A to that of constructing a collection of vectors that is $(m-1)$-wise but not m-wise superorthogonal with respect to π.

Definition 7.4.17. *A collection of vectors $v_1, ..., v_\ell \in \mathbb{R}^k$ is* non-top-degree-vanishing *if $v_1 \odot \cdots \odot v_\ell$ does not lie in the span of $\{\odot_{i \in S} v_i\}_{S \subsetneq [\ell]}$.*

Observation 7.4.18. *Suppose $v_1, ..., v_{m-1} \in \mathbb{R}^k$ are $(m-1)$-wise superorthogonal with respect to π and non-top-degree-vanishing. Denote the span of $\{\odot_{i \in S} v_i\}_{S \subsetneq [m-1]}$ by V. If $v_m \in \mathbb{R}^k$ satisfies*

$$v_m \cdot \mathit{diag}(\pi) \cdot v^\top = 0 \; \forall \; v \in V \tag{7.6}$$

$$v_m \cdot \mathit{diag}(\pi) \cdot (v_1 \odot \cdots \odot v_{m-1})^\top \neq 0, \tag{7.7}$$

then $v_1, ..., v_m$ are $(m-1)$- but not m-wise superorthogonal with respect to π.

Note that any collection of vectors that are $(m-1)$-wise but not m-wise superorthogonal with respect to π must arise in this way. By Observation 7.4.18, we can focus on finding the largest ℓ for which there exist vectors $v_1, ..., v_\ell$ which are ℓ-wise superorthogonal and non-top-degree-vanishing.

Construction 1. *Let $k = (\ell+1)^2$ and $\pi = \frac{1}{k}\mathbf{1}$, and fix any distinct scalars $x_1, ..., x_{\ell+1} \in \mathbb{R}$. Define matrices*

$$\mathbf{a} = \begin{pmatrix} x_1 & x_2 & \cdots & x_{\ell+1} \\ x_1 & x_2 & \cdots & x_{\ell+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_1 & x_2 & \cdots & x_{\ell+1} \end{pmatrix} \qquad \mathbf{b_i} = \begin{pmatrix} -x_i & 0 & 0 & \cdots & 0 \\ x_i & -2x_i & 0 & \cdots & 0 \\ x_i & x_i & -3x_i & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_i & x_i & x_i & \cdots & -\ell x_i \end{pmatrix}$$

with ℓ rows each. Define the $\ell \times k$ matrix

$$\mathcal{E}(x_1, ..., x_{\ell+1}) \triangleq (\mathbf{a} \| \mathbf{b_1} \| \cdots \| \mathbf{b_{\ell+1}}).$$

Remark 7.4.19. *In fact there are more efficient constructions that save constant factors on k as a function of ℓ, but we choose not to discuss these to maximize transparency of the proof.*

Lemma 7.4.20. *The rows of $\mathcal{E}(x_1, ..., x_{\ell+1})$ are ℓ-wise superorthogonal and non-top-degree-*

vanishing.

Proof. Denote the matrices whose rows consist of entrywise products of rows of $\mathbf{a}$ and rows of $\mathbf{b_i}$ respectively by $\mathbf{A}$ and $\mathbf{B_i}$. Superorthogonality just follows from the fact that the entries of any row $(\mathbf{B_i})_S$ sum to $-x_i^{|S|}$, while the entries of any row $(\mathbf{A})_S$ sum to $\sum_{i=1}^{\ell+1} x_i^{|S|}$.

To show that the rows of $\mathcal{E}(x_1, ..., x_{\ell+1})$ are non-top-degree-vanishing, it's enough to show that the rows of $\mathbf{a}$ are non-top-degree-vanishing. The latter is true because the rows of A are copies of rows of an $(\ell+1) \times (\ell+1)$ Vandermonde matrix, and $\mathbf{A}_{[\ell]}$ is the unique row of $\mathbf{A}$ equal to $(x_1^\ell \ \cdots \ x_{\ell+1}^\ell)$. □

Henceforth let $m = \ell + 1$. To pass from $\mathcal{E}(x_1, ..., x_m)$ to the desired mixture of product distributions A with mixing weights $\pi = \frac{1}{k} \cdot \mathbf{1}$: solve (7.6) and (7.7) in v_m, append this as a row to $\mathcal{E}(x_1, ..., x_m)$, scale all rows so that the entries all lie in $[-1/2, 1/2]$, and add the resulting matrix to $\frac{1}{2} \cdot \mathbf{J}_{m\times k}$ to get the marginals matrix for A.

It remains to choose $x_1, ..., x_m$ so that $\delta(A)$ is reasonably large (we make no effort to optimize this choice). It turns out that simply choosing $x_1, ..., x_m$ to be an appropriately scaled arithmetic progression works, and the remainder of the section is just for verifying this.

We first collect some standard facts about Vandermonde matrices. Define

$$V_m = \begin{pmatrix} x_1 & \cdots & x_m \\ x_1^2 & \cdots & x_m^2 \\ \vdots & \ddots & \vdots \\ x_1^{m-1} & \cdots & x_m^{m-1} \end{pmatrix}.$$

Lemma 7.4.21. *For distinct $x_1, ..., x_m$, the right kernel of V_m is the line through $u \in \mathbb{R}^m$ given by*

$$u_i = (-1)^{i+1} \left(\prod_{j \neq i} x_j \right) \cdot \prod_{j<k: j,k \neq i} (x_j - x_k)$$

for each $i \in [m]$.

Proof. For any row index $1 \leq d \leq m-1$, observe that

$$\langle (V_m)_d, u \rangle = \begin{vmatrix} x_1 & \cdots & x_m \\ x_1^2 & \cdots & x_m^2 \\ \vdots & \ddots & \vdots \\ x_1^{m-1} & \cdots & x_m^{m-1} \\ x_1^d & \cdots & x_m^d \end{vmatrix} = 0.$$

□

Corollary 7.4.22. *If $(x_1, x_2, ..., x_m) = (\lambda, 2\lambda, ..., m\lambda)$, then the right kernel of V_m is the line through $v \in \mathbb{R}^m$ given by $v_i = (-1)^i \binom{m}{i}$.*

Proof. Let $u \in \mathbb{R}^m$ be a point on the line corresponding to the right kernel of V_m. Define $v = u/Z$ where $Z = (-1)^m m! \prod_{j=1}^m x_j \cdot \prod_{1 \leq j < k \leq m}(x_j - x_k)$, giving

$$\begin{aligned} v_i &= \frac{(-1)^{m+i+1} m!}{x_i \cdot (x_1 - x_i) \cdots (x_{i-1} - x_i) \cdot (x_{i+1} - x_i) \cdots (x_i - x_m)} \\ &= \frac{(-1)^{m+i+1} m!}{(-1)^{m-1} \cdot \lambda^m i (i-1)!(m-i)!} \\ &= (-1)^i \binom{m}{i} \end{aligned}$$

as desired. □

Observation 7.4.23. *Let V be the span of all entrywise products of rows of $\mathcal{E}(x_1, ..., x_m)$. For $1 \leq i, d \leq m-1$ let $v(d,i) \in \mathbb{R}^k$ be the vector defined by*

$$v(d,i)_j = \begin{cases} x_j^d, & j \leq m \\ x_s^d, & j = 1 + i + (m-1)s \text{ for } s \in [m] \\ 0, & \text{otherwise} \end{cases} \cdot$$

The set $\{v(d,i)\}_{1 \leq d \leq m-1, i+d \leq m}$ together with $\mathbf{1}$ form a basis for V.

Proof. This just follows by elementary row operations applied to entrywise products of d rows of $\mathcal{E}(x_1, ..., x_m)$ for each $d \in [m-1]$. □

Corollary 7.4.24. *Let $v_i = (-1)^i\binom{m}{i}$. The space of solutions to (7.6) contains the space of vectors parametrized by*

$$(a_1, ...a_m, a_1+\lambda_1 v_1, ..., a_1+\lambda_{m-1}v_1, a_2+\lambda_1 v_2, ..., a_2+\lambda_{m-1}v_2, ..., a_m+\lambda_1 v_m, ..., a_m+\lambda_{m-1}v_m) \tag{7.8}$$

for $a_1, ..., a_m, \lambda_1, ..., \lambda_{m-1} \in \mathbb{R}$ satisfying

$$m(a_1 + \cdots + a_m) = \lambda_1 + \cdots + \lambda_{m-1}. \tag{7.9}$$

Proof. The space of vectors parametrized by (7.8) is precisely those which are orthogonal to the span of $\{v(d,i)\}_{1\le i,d\le m-1}$. This span together with $\mathbf{1}$ has orthogonal complement which is a strict subspace of the space of solutions to (7.6). The sum of the entries in (7.8) is

$$m(a_1+\cdots+a_m)+(\lambda_1+\cdots+\lambda_{m-1})(v_1+\cdots+v_m) = m(a_1+\cdots+a_m)-(\lambda_1+\cdots+\lambda_{m-1})$$

because $\sum_{i=1}^m v_i = -1$, so (7.9) is just the constraint that any solution to (7.6) is orthogonal to $\mathbf{1}$. □

Lemma 7.4.25. *Let $\pi = \frac{1}{k}\cdot\mathbf{1}$ and let A' be the $m\times k$ matrix obtained by concatenating $\mathcal{E}(x_1, ..., x_m)$ with a row vector of the form (7.8), where $x_i = i/2m^2$ for all $i \in [m]$, $\lambda_2 = -\lambda_1 = -2^m$, $\lambda_3 = \cdots = \lambda_{m-1} = 0$, and $a_1 = \cdots = a_m = 0$. Define $A = A' + \frac{1}{2}\cdot\mathbf{J}_{m\times k}$. Then if $m+1$ is a prime, $|\delta(A)| \ge (2m)^{-2m}$.*

Proof. One can check that

$$\delta(A) = -\lambda_1(v_1 x_1^m + \cdots + v_m x_m^m). \tag{7.10}$$

By selecting $\lambda_2 = -\lambda_1$, $\lambda_3 = \cdots = \lambda_{m-1} = 0$, and $a_1 = \cdots = a_m = 0$, we satisfy (7.9). Furthermore, the only nonzero entries of (7.8) are $\pm\lambda_1 v_i$ for all $i \in [m]$. In particular by taking, e.g., $\lambda_1 = 2^m$, we ensure that all entries of (7.8) are in $[-1/2, 1/2]$. If we then take

$x_i = i/2m^2$ for all $i \in [m]$, we get from (7.10) that

$$\delta(A) = -\frac{1}{2^m}\sum_{i=1}^{m}(-1)^i\binom{m}{i}\left(\frac{i}{2m^2}\right)^m = -\frac{1}{(2m)^{2m}}\sum_{i=1}^{m}(-1)^i\binom{m}{i}i^m.$$

$\sum_{i=1}^{m}(-1)^i\binom{m}{i}$ is an integer, so it's enough to show that it's nonzero to get that $|\delta(A)| \geq \frac{1}{(2m)^{2m}}$. Without loss of generality suppose $m+1$ is a prime, in which case

$$\sum_{i=1}^{m}(-1)^i\binom{m}{i}i^m \equiv \sum_{i=1}^{m}(-1)^i\binom{m}{i} \equiv 1 \pmod{m+1}$$

by Fermat's little theorem. □

Theorem 7.4.1 then follows from Proposition 7.4.11 by taking $m = \sqrt{k}$.

7.5 Learning Mixtures of Product Distributions in $n^{O(k^2)}$ Time

We now use ideas similar to those of Section 7.3 to prove Theorem 7.1.6. Specifically, we give an algorithm that outputs a list of at most $(n/\varepsilon)^{O(k^2)}$ candidate distributions, at least one of which is $O(\varepsilon)$-close in total variation distance to $\mathcal{D}$. By standard results about hypothesis selection, e.g. Scheffe's tournament method [DL01], we can then pick out a distribution from this list which is $O(\varepsilon)$-close to $\mathcal{D}$ in time and samples polynomial in the size of the list.

Unlike in the case of learning mixtures of subcubes where we insisted on running in time $n^{O(\log k)}$, here we can afford to simply brute-force search for a basis for $\mathbf{M}$ for any realization of $\mathcal{D}$. In fact our strategy will be: 1) brute-force search for a row basis $J = \{i_1, ..., i_r\}$ for $\mathbf{m}$, 2) use $\mathbf{C}$ together with Lemma 7.2.10 to find coefficients expressing the remaining rows of $\mathbf{m}$ as linear combinations of the basis elements, and 3) brute-force search for the mixing weights and entries of $\mathbf{m}$ in rows $i_1, ..., i_r$. Each attempt in our brute-force procedure will correspond to a candidate distribution in the list on which our algorithm ultimately runs hypothesis selection. We outline this general approach in the first three subsections.

While the task of obtaining a basis for the rows of $\mathbf{M}$ is simpler here than in learning mixtures of subcubes, the issue of ill-conditioned matrices is much more subtle. Whereas for mixtures of subcubes, Lemma 7.3.10 guarantees that the matrices we deal with are all either well-conditioned or not full rank, for general mixtures of product distributions the matrices we deal with can be arbitrarily badly conditioned. This already makes it much trickier to prove robust low-degree identifiability, which we do in Section 7.5.4.

Once we have robust low-degree identifiability, we can adapt the ideas of Section 7.3 for handling $\mathbf{M}|_{\mathcal{R}'(J)}$ not being full rank to handle $\mathbf{M}|_{\mathcal{R}'(J)}$ being ill-conditioned.[7] Analogous to arguing that $\mathcal{D}|x_J = s$ can be realized as a mixture of fewer product distributions when $\mathbf{M}|_{\mathcal{R}'(J)}$ is not full rank, we argue that $\mathcal{D}|x_J = s$ is *close* to a mixture of fewer product distributions when $\mathbf{M}|_{\mathcal{R}'(J)}$ is ill-conditioned.

After describing our algorithm in greater detail, we summarize in Section 7.5.6 how our algorithm manages to improve upon that of [FOS05].

7.5.1 Parameter Closeness Implies Distributional Closeness

We first clarify what we mean by brute-force searching for the underlying parameters of $\mathcal{D}$. In a general mixture $\mathcal{D}$ of product distributions realized by mixing weights π and marginals matrix $\mathbf{m}$, the entries of π and $\mathbf{m}$ can take on any values in $[0,1]$. The following lemmas show that it's enough to recover π and $\mathbf{m}$ to within some small entrywise error ε'. So for instance, instead of searching over all choices $\{0,1/2,1\}^{n\times k}$ for $\mathbf{m}$ as in the subcubes setting, we can search over all choices $\{0,\varepsilon',2\varepsilon',...,\lfloor 1/\varepsilon'\rfloor\varepsilon\}^{n\times k}$.

Lemma 7.5.1. *If $\mathcal{D}$ and $\overline{\mathcal{D}}$ are mixtures of at most k product distributions over $\{0,1\}^n$ with the same mixing weights π and marginals matrices $\mathbf{m}$ and $\overline{\mathbf{m}}$ respectively such that $|\mathbf{m}^i_j - \overline{\mathbf{m}}^i_j| \le \varepsilon/2kn$ for all $i,j \in [k]\times[n]$, then $d_{TV}(\mathcal{D},\overline{\mathcal{D}}) \le \varepsilon$.*

Proof. Consider $\mathcal{D}$ and $\overline{\mathcal{D}}$ whose marginals matrices are equal except in the (i,j)-th entry

[7] Here, recall that $\mathcal{R}(J) = 2^{[n]\backslash J}$. Lemma 7.3.2 implies that

$$\text{rank}(\mathbf{M}|_{\mathcal{R}(J)}) = \text{rank}(\mathbf{M}|_{\mathcal{R}'(J)})$$

where in our discussion of general mixtures of product distributions $\mathcal{R}'(J)$ is the set of all subsets $T \subseteq [n]\backslash J$ with $|T| = k$. In fact, for technical reasons that we defer to Appendix 7.8, we will actually need to use subsets of size up to $O(k^2)$. This will not affect our runtime, but for the discussion in this section it is fine to ignore this detail.

where they differ by $\le \varepsilon/2kn$. Then it is clear that $d_{\mathrm{TV}}(\mathcal{D}, \overline{\mathcal{D}}) \le \varepsilon/kn$. So by a union bound, if $\mathcal{D}$ and $\overline{\mathcal{D}}$ have marginals matrices differing entrywise by $\le \varepsilon/2kn$, then $d_{\mathrm{TV}}(\mathcal{D}, \overline{\mathcal{D}}) \le \varepsilon$ □

Lemma 7.5.2. *If $\mathcal{D}$ and $\overline{\mathcal{D}}$ are mixtures of at most k product distributions over $\{0,1\}^n$ with the same marginal matrices $\mathbf{m}$ and mixing weights π and $\overline{\pi}$ respectively such that $|\pi^i - \overline{\pi}^i| \le 2\varepsilon/k$ for all $i \in [k]$, then $d_{TV}(\mathcal{D}, \overline{\mathcal{D}}) \le \varepsilon$.*

Proof. Denote the probability that the i-th center of either $\mathcal{D}$ or $\overline{\mathcal{D}}$ takes on the value s by p_i. For any $s \in \{0,1\}^n$,

$$|\Pr_{\mathcal{D}}[s] - \Pr_{\overline{\mathcal{D}}}[s]| = |\langle (p_1 \ \cdots \ p_k), \pi - \overline{\pi}\rangle| \le k \cdot (2\varepsilon/k) = 2\varepsilon,$$

We conclude that $d_{\mathrm{TV}}(\mathcal{D}, \overline{\mathcal{D}}) \le \varepsilon$ as desired. □

The next lemma says that we can get away with not recovering product distributions in the mixture that have sufficiently small mixing weights.

Lemma 7.5.3. *Let $\mathcal{D}$ be a mixture of k product distributions over $\{0,1\}^n$ with mixing weights π and marginals matrix $\mathbf{m}$. Denote by $S \subseteq [k]$ the coordinates i of π for which $\pi^i \ge \varepsilon/k$, and let $Z = \sum_{i \in S} \pi^i$. Then the mixture $\overline{\mathcal{D}}$ of $|S|$ product distributions over $\{0,1\}^n$ realized by $(\frac{1}{Z}\pi|^S, \mathbf{m}|^S)$ satisfies $d_{TV}(\mathcal{D}, \overline{\mathcal{D}}) \le \varepsilon$.*

Proof. We can regard $\overline{\mathcal{D}}$ as a distribution which with probability Z samples from one of the centers of $\mathcal{D}$ indexed by $i \in S$ with probability proportional to π, and with probability $1 - Z$ samples from some other distribution. We can regard $\mathcal{D}$ in the same way. Then their total variation distance is bounded above by $1 - Z \le \varepsilon/k \cdot k = \varepsilon$. □

7.5.2 Barycentric Spanners

To control the effect that sampling noise in our estimates for moments of $\mathcal{D}$ has on the approximation guarantees of our learning algorithm, it is not enough simply to find a row basis for $\mathbf{m}$ for any realization of $\mathcal{D}$, but rather one for which the coefficients expressing the remaining rows of $\mathbf{m}$ in terms of this basis are small. The following, introduced in [AK08], precisely captures this notion.

Definition 7.5.4. *Given a collection of vectors $V = \{v_1, ..., v_n\}$ in $\mathbb{R}^k$, $S \subseteq V$ is a* barycentric spanner *if every element of V can be expressed as a linear combination of elements of S using coefficients in $[-1, 1]$.*

Lemma 7.5.5 (Proposition 2.2 in [AK08])**.** *Every finite collection of vectors $V = \{v_1, ..., v_n\} \subseteq \mathbb{R}^k$ has a barycentric spanner.*

Proof. Without loss of generality suppose that V spans all of $\mathbb{R}^k$. Pick $v_{i_1}, ..., v_{i_k}$ for which $|\det(v_{i_1}, ..., v_{i_k})|$ is maximized. Take any $v \in V$ and write it as $\sum_j \alpha_j v_{i_j}$. Then for any $j \in [n]$, $|\det(v_{i_1}, ..., v_{i_{j-1}}, v, v_{i_{j+1}}, ..., v_{i_k})| = |\alpha_j| \cdot |\det(v_{i_1}, ..., v_{i_k})|$. By maximality, $|\alpha_j| \leq 1$, so $v_{i_1}, ..., v_{i_k}$ is a barycentric spanner. □

7.5.3 Gridding the Basis and Learning Coefficients

In time $n^{O(k)}$ we can brute-force find a barycentric spanner $J = \{i_1, ..., i_r\}$ for the rows of $\mathbf{m}$. We can then $\frac{\varepsilon}{4k^2 n}$-grid the entries of $\mathbf{m}|_J$ in time $(n/\varepsilon)^{k^2}$ to get an entrywise $\frac{\varepsilon}{4k^2 n}$-approximation $\overline{\mathbf{m}}|_J$ of $\mathbf{m}$. Now suppose for the moment that we had exact access to the entries of $\mathbf{C}$. We can try solving

$$\mathbf{C}|^{\{i_1\},...,\{i_r\}}_{\mathcal{R}'(J\cup\{i\})} \alpha_i = \mathbf{C}|^{\{i\}}_{\mathcal{R}'(J\cup\{i\})} \tag{7.11}$$

in $\alpha_i \in \mathbb{R}^r$ for every $i \notin J$.

If $\text{rank}(\mathbf{M}|_{\mathcal{R}'(J\cup\{i\})}) = k$ for all $i \notin J$ and realizations of $\mathcal{D}$, then by Lemma 7.2.10, the coefficient vectors α_i also satisfy $\alpha_i \cdot \mathbf{m}|_J = \mathbf{m}_i$. Because J is a barycentric spanner so that $\alpha_i \in [-1, 1]^r$, if we define $\overline{\mathbf{m}}_i$ by

$$\overline{\mathbf{m}}_i = \alpha_i \cdot \overline{\mathbf{m}}|_J, \tag{7.12}$$

then $\overline{\mathbf{m}}$ is an entrywise $\frac{\varepsilon}{4k^2 n} \cdot k = \frac{\varepsilon}{4kn}$-approximation of $\mathbf{m}$. We can then $\frac{\varepsilon}{2k}$-grid mixture weights $\bar{\pi}$, and by Lemmas 7.5.1, 7.5.2, and 7.5.3 we have learned a mixture of product distributions $\overline{\mathcal{D}}$ for which $d_{\text{TV}}(\mathcal{D}, \overline{\mathcal{D}}) \leq \varepsilon$.

As usual, the complication is that it may be that $\text{rank}(\mathbf{M}|_{\mathcal{R}'(J\cup\{i\})}) < k$ for some realization of $\mathcal{D}$, but as in our algorithm for learning mixtures of subcubes, we can handle this by conditioning on $J \cup \{i\}$ and recursing.

As we alluded to at the beginning of the section, a more problematic issue that comes up here but not in the subcube setting is that $\mathbf{M}|_{\mathcal{R}'(J\cup\{i\})}$ might be full rank but very badly conditioned. Indeed, in reality we only have $\varepsilon_{\mathrm{samp}}$-close estimates $\tilde{\mathbf{C}}$ to the accessible entries of $\mathbf{C}$, so instead of solving (7.11), we solve the analogous L_∞ regression

$$\tilde{\alpha}_i \triangleq \operatorname*{argmin}_{\alpha\in[-1,1]^r} \|\tilde{\mathbf{C}}|^{\{i_1\},\dots,\{i_r\}}_{\mathcal{R}'(J\cup\{i\})}\alpha - \tilde{\mathbf{C}}|^{\{i\}}_{\mathcal{R}'(J\cup\{i\})}\|_\infty. \tag{7.13}$$

If $\sigma^\infty_{\min}(\mathbf{M}|_{\mathcal{R}'(J\cup\{i\})}$ is badly conditioned, then we cannot ensure that the resulting $\tilde{\alpha}_i$ lead to $\overline{\mathbf{m}}_i = \tilde{\alpha}_i \cdot \overline{\mathbf{m}}|_J$ in (7.12) which are close to the true $\mathbf{m}_i$.

We show in the next subsections that this issue is not so different from when $\operatorname{rank}(\mathbf{M}|_{\mathcal{R}'(J\cup\{i\})}) < k$ for some realization of $\mathcal{D}$, and we can effectively treat ill-conditioned moment matrices as degenerate-rank moment matrices. As we will see, the technical crux underlying this is the fact that mixtures of product distributions are robustly identified by their $O(k)$-degree moments.

7.5.4 Robust Low-degree Identifiability

Lemma 7.3.2 is effectively an exact identifiability result that implies that if a mixture of k_1 product distributions *exactly* agrees with a mixture of k_2 product distributions on all moments of degree at most $k_1 + k_2$, then they are identical as distributions. The following is a robust identifiability lemma saying that if instead the two mixtures are only close on moments of degree at most k_1+k_2, then they are close in total variation distance. Recall that we showed a similar lemma for mixtures of subcubes, but there it was much easier to extend exact identifiability to robust identifiability because full rank moment matrices are always well-conditioned, something that does not always hold for mixtures of product distributions.

Lemma 7.5.6. *Let $\mathcal{D}_1, \mathcal{D}_2$ respectively be mixtures of k_1 and k_2 product distributions in $\{0,1\}^n$ for k_1, k_2. If $d_{TV}(\mathcal{D}_1,\mathcal{D}_2) > \varepsilon$, there is some S for which $|S| < k_1+k_2$ and $|\mathbb{E}_{\mathcal{D}_1}[x_S] - \mathbb{E}_{\mathcal{D}_2}[x_S]| > \eta$ for some $\eta = \exp(-O(k_1+k_2)^2)\cdot \operatorname{poly}(k_1+k_2,n,\varepsilon)^{-k_1-k_2}$.*

We will prove the contrapositive by induction on k_1+k_2. Suppose $|\mathbb{E}_{\mathcal{D}_1}[x_S]-\mathbb{E}_{\mathcal{D}_2}[x_S]| \le \eta$ for all $S \subseteq [n]$ with $|S| < k_1 + k_2$. Define $\delta = \varepsilon/2kn$. Henceforth suppose $\mathcal{D}_1$ and $\mathcal{D}_2$ are realized by $(\pi_1, \mathbf{m}_1)$ and $(\pi_2, \mathbf{m}_2)$ respectively for $\pi^1_1 \ge \cdots \ge \pi^{k_1}_1$ and $\pi^1_2 \ge \cdots \ge \pi^{k_2}_2$. For

$i \in [n]$ let u_i, ℓ_i denote the largest and smallest value in row i of either $\mathbf{m}_1$ or $\mathbf{m}_2$.

The following simple observation, similar in spirit to Lemma 7.2.8, drives our induction:

Observation 7.5.7. *For any $i \in [n]$ and each $j = 1, 2$, there exists a mixture of product distributions $\mathcal{D}^\ell_j$ over $\{0,1\}^{n-1}$ such that for any $S \subseteq [n]\backslash\{i\}$,*

$$\mathbb{E}_{\mathcal{D}_1}[(x_i - \ell_i) \cdot x_S] = \mathbb{E}_{\mathcal{D}_1}[x_i - \ell_i] \cdot \mathbb{E}_{\mathcal{D}^\ell_1}[x_S]. \tag{7.14}$$

If ℓ_i is an entry of $(\mathbf{m}_1)_i$, then $\mathcal{D}^\ell_1$ and $\mathcal{D}^\ell_2$ are mixtures of at most $k_1 - 1$ and k_2 product distributions respectively. If we replace $(x_i - \ell_i)$ with $(u_i - x_i)$, the analogous statement holds for mixtures $\mathcal{D}^u_1, \mathcal{D}^u_2$.

Proof. For each j, $\mathcal{D}^\ell_j$ is obviously realized by

$$\left(\frac{1}{Z_j}\pi_1 \odot ((\mathbf{m}_j)_i - \ell_i \cdot \mathbf{1}), (\mathbf{m}_j)|_{[n]\backslash\{i\}}\right),$$

where $Z_j = \mathbb{E}_{\mathcal{D}_j}[x_i - \ell_i]$. But for $j = 1$, $\pi_1 \odot ((\mathbf{m}_1)_i - \ell_i \cdot \mathbf{1})$ has a zero in the entry corresponding to where ℓ_i is in $(\mathbf{m}_1)_i$. So $\mathcal{D}^\ell_1$ is in fact realized by the mixture weight vector consisting of all nonzero entries of $\pi_1 \odot ((\mathbf{m}_1)_i - \ell_i \cdot \mathbf{1})$ together with the corresponding at most $k_1 - 1$ columns of $(\mathbf{m}_1)_{[n]\backslash\{i\}}$. □

One subtlety is that we need to pick a row i such that $\mathbb{E}_{\mathcal{D}_j}[(x_i - \ell_i)]$ and $\mathbb{E}_{\mathcal{D}_j}[(u_i - x_i)]$ is sufficiently large that when we induct on the pairs of mixtures $\mathcal{D}^\ell_1, \mathcal{D}^\ell_2$ and $\mathcal{D}^u_1, \mathcal{D}^u_2$, the assumption that the pair of mixtures $\mathcal{D}_1, \mathcal{D}_2$ is close on low-degree moments carries over to these pairs. In the following lemma we argue that if no such row i exists, then $\mathcal{D}_1$ and $\mathcal{D}_2$ are both close to a single product distribution and therefore close in total variation distance to each other.

Lemma 7.5.8. *If there exists no $i \in [n]$ for which $\mathbb{E}_{\mathcal{D}_1}[x_i - \ell_i] \geq \delta\varepsilon/4k$ and $\mathbb{E}_{\mathcal{D}_1}[u_i - x_i] \geq \delta\varepsilon/4k$, then $d_{TV}(\mathcal{D}_1, \Pi) \leq \varepsilon$, where Π is the single product distribution with i-th marginal ℓ_i if $\mathbb{E}_{\mathcal{D}_1}[x_i - \ell_i] \leq \delta\varepsilon/4k$ and u_i if $\mathbb{E}_{\mathcal{D}_1}[u_i - x_i] \geq \delta\varepsilon/4k$. In particular, if there exists no $i \in [n]$ for which $\mathbb{E}_{\mathcal{D}_1}[x_i - \ell_i] \geq \delta\varepsilon/9k$ and $\mathbb{E}_{\mathcal{D}_1}[u_i - x_i] \geq \delta\varepsilon/9k$, then $d_{TV}(\mathcal{D}_1, \mathcal{D}_2) \leq \varepsilon$.*

Proof. Let $k'_1 \leq k_1$ be the largest index for which $\pi_1^{k'_1} \geq \varepsilon/2k$. If there exists no $i \in [n]$ for which $\mathbb{E}_{\mathcal{D}_1}[x_i - \ell_i] \geq \delta\varepsilon/4k$ and $\mathbb{E}_{\mathcal{D}_1}[u_i - x_i] \geq \delta\varepsilon/4k$, then for every $i \in [n]$ and $1 \leq j \leq k'_1$, $\mathbf{m}^j_i \in [\ell_i, \ell_i + \delta/2] \cup [u_i - \delta/2, u_i]$, so by Lemmas 7.5.1 and Lemma 7.5.3, $d_{\mathrm{TV}}(\mathcal{D}_1, \Pi) \leq \varepsilon$.

For the second statement in the lemma, note that the argument above obviously also holds if $\mathcal{D}_1$ is replaced with $\mathcal{D}_2$. If $\mathbb{E}_{\mathcal{D}_1}[x_i - \ell_i] \geq \delta\varepsilon/9k$, then by the assumption that $\mathcal{D}_1$ and $\mathcal{D}_2$ are η-close on all low-order moments, $\mathbb{E}_{\mathcal{D}_2}[x_i - \ell_i] \leq \delta\varepsilon/9k + \eta \leq \delta\varepsilon/8k$ and we conclude by invoking the first part of the lemma on both $\mathcal{D}_1$ and $\mathcal{D}_2$ to conclude that $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) \leq d_{\mathrm{TV}}(\mathcal{D}_1, \Pi) + d_{\mathrm{TV}}(\mathcal{D}_2, \Pi) \leq \varepsilon$. □

Finally, before we proceed with the details of the inductive step, we check the base case when at least one of k_1, k_2 is 1.

Lemma 7.5.9. *Let $\mathcal{D}_1$ be a single product distribution over $\{0,1\}^n$ and $\mathcal{D}_2$ a mixture of k product distributions over $\{0,1\}^n$. If $d_{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$, there is some S for which $|S| \leq k+1$ and $|\mathbb{E}_{\mathcal{D}_1}[x_S] - \mathbb{E}_{\mathcal{D}_2}[x_S]| > \eta$ for $\eta = \frac{\varepsilon^3}{648 \cdot 2^k n^2}$.*

Proof. Let $p_1, ..., p_n$ be the marginals of $\mathcal{D}_1$ and let π and $\mathbf{m}$ be mixing weights and marginals matrix realizing $\mathcal{D}_2$. For each $i \in [n]$ define $v_i = \mathbf{m}_i - p_i \cdot \mathbf{1}$. For $i \neq j$, observe that

$$\begin{aligned}\langle \pi, v_i \odot v_j \rangle| &= |\langle \pi, \mathbf{m}_i \odot \mathbf{m}_j + p_i \cdot p_j \cdot \mathbf{1} - p_i \cdot \mathbf{m}_j - p_j \cdot \mathbf{m}_i \rangle| \\ &= \left| \mathbb{E}_{\mathcal{D}_2}[x_{\{i,j\}}] + \mathbb{E}_{\mathcal{D}_1}[x_{\{i,j\}}] - (\mathbb{E}_{\mathcal{D}_1}[x_{\{i,j\}}] \pm \eta) - (\mathbb{E}_{\mathcal{D}_1}[x_{\{i,j\}}] \pm \eta) \right| \leq 3\eta.\end{aligned}$$

Pick out a barycentric spanner $J \subseteq [n]$ for $\{v_1, ..., v_n\}$ so that for all $i \notin J$, there exist coefficients $\lambda^i_j \in [-1,1]$ for which $v_i = \sum_{j \in J} \lambda_j v_j$. From this we get

$$\langle \pi, v_i \odot v_i \rangle = |\langle \pi, v_i \odot v_i \rangle| \leq \sum_{j \in J} |\lambda_j| \cdot |\langle \pi, v_i \odot v_j \rangle| \leq 3\eta k.$$

All entries of $v_i \odot v_i$ are obviously nonnegative, so for $\tau = \varepsilon/6k$ to be chosen later, we find that $|v^\ell_i| \leq \sqrt{3\eta k/\tau} \leq \varepsilon/6nk$ for all $\ell \in [k]$ for which $\pi^\ell > \tau$. Denote the set of such ℓ by $S \subseteq [k]$.

By restricting to entries of π in S, normalizing, and restricting the columns of $\mathbf{m}$ to S, we get a new mixture of product distributions $\mathcal{D}'$ with marginals matrix $(\pi', \mathbf{m}|^S)$ which is $\tau k = \varepsilon/6$-close to $\mathcal{D}_2$. For all $i \notin J$ and $\ell \in S$, because $|v^\ell_i| \leq \varepsilon/6nk$, if we replace every such

(i,ℓ)-th entry of $\mathbf{m}|^S$ by p_i to get $\mathbf{m}'$, then the mixture of product distributions $\mathcal{D}''$ realized by $(\pi', \mathbf{m}')$ is $\varepsilon/6$-close to $\mathcal{D}'$.

For a distribution D let $D|_J$ denote its restriction to coordinates J. Total variation distance is nonincreasing under this restriction operation, so $d_{\mathrm{TV}}(\mathcal{D}_2|_J, \mathcal{D}''|_J) \le d_{\mathrm{TV}}(\mathcal{D}_2, \mathcal{D}'')$. Furthermore, note that $d_{\mathrm{TV}}(\mathcal{D}_1|_J, \mathcal{D}_2|_J) \le 2^{2k}\eta < \varepsilon/3$, because $|J| \le k$ and any event on $\{0,1\}^k$ can obviously be expressed in terms of at most 2^{2k} moments of $\mathcal{D}_1|_J$ and $\mathcal{D}_2|_J$. By the triangle inequality, $d_{\mathrm{TV}}(\mathcal{D}_1|_J, \mathcal{D}''|_J) \le 2\varepsilon/3$.

Finally, define Π to be the product distribution over $\{0,1\}^{n-|J|}$ with marginals $\{p_i\}_{i \notin J}$. By design, $\mathcal{D}_1 = \mathcal{D}_1|_J \times \Pi$ and $\mathcal{D}'' = \mathcal{D}''|_J \times \Pi$. Because Π is a single product distribution, $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}'') = d_{\mathrm{TV}}(\mathcal{D}_1|_J, \mathcal{D}''|_J) \le 2\varepsilon/3$. By the triangle inequality, we get that $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) \le \varepsilon$. □

We are now ready to complete the inductive step in the proof of Lemma 7.5.6.

Proof of Lemma 7.5.6. Pick $\eta = 5^{-2(k_1+k_2)^2} \cdot \left(\frac{(\delta\varepsilon/k)^2}{162}\right)^{k_1+k_2}$. For $k_1 = 1$ or $k_2 = 1$, we certainly have $\eta < \frac{\varepsilon^3}{648 \cdot 2^k n^2}$, so the base case follows by Lemma 7.5.9.

Now consider the case where $k_1, k_2 > 1$. Suppose

$$|\mathbb{E}_{\mathcal{D}_1}[x_S] - \mathbb{E}_{\mathcal{D}_2}[x_S]| \le \eta \tag{7.15}$$

for all $|S| < k_1 + k_2$. By Lemma 7.5.8 we may assume that there exists an i for which $\mathbb{E}_{\mathcal{D}_1}[x_i - \ell_i] \ge \delta\varepsilon/9k$ and $\mathbb{E}_{\mathcal{D}_1}[u_i - x_i] \ge \delta\varepsilon/9k$. Because Lemma 7.5.8 also holds for $\mathcal{D}_2$, we may assume without loss of generality that ℓ_i is an entry of $(\mathbf{m}_1)_i$. Take any $T \subseteq [n]\backslash\{i\}$ for $|T| < k_1 + k_2 - 1$. By (7.15) we have that

$$\left|\mathbb{E}_{\mathcal{D}_1}[(x_i - \ell_i) \cdot x_{T\cup\{i\}}] - \mathbb{E}_{\mathcal{D}_1}[(x_i - \ell_i) \cdot x_{T\cup\{i\}}]\right| \le 2\eta.$$

By (7.14) we have that

$$\begin{aligned}\left|\mathbb{E}_{\mathcal{D}_1^\ell}[x_T] - \mathbb{E}_{\mathcal{D}_2^\ell}[x_T]\right| &= \left|\frac{\mathbb{E}_{\mathcal{D}_1}[(x_i - \ell_i) \cdot x_{T\cup\{i\}}]}{\mathbb{E}_{\mathcal{D}_1}[x_i - \ell_i]} - \frac{\mathbb{E}_{\mathcal{D}_2}[(x_i - \ell_i) \cdot x_{T\cup\{i\}}]}{\mathbb{E}_{\mathcal{D}_2}[x_i - \ell_i]}\right| \\ &= \left|\frac{\pm\eta \cdot \mathbb{E}_{\mathcal{D}_1}[(x_i - \ell_i) \cdot x_{T\cup\{i\}}] \pm 2\eta \cdot \mathbb{E}_{\mathcal{D}_1}[x_i - \ell_i]}{\mathbb{E}_{\mathcal{D}_1}[x_i - \ell_i] \cdot \mathbb{E}_{\mathcal{D}_2}[x_i - \ell_i]}\right|\end{aligned}$$

$$\leq \frac{2\eta}{\delta\varepsilon/9k} + \frac{\eta}{(\delta\varepsilon/9k)^2} \leq \frac{2\eta}{(\delta\varepsilon/9k)^2} \leq 5^{-2(k_1+k_2-1)^2}\left(\frac{(\delta\varepsilon/5k)^2}{162}\right)^{k_1+k_2-1}$$

Because $\mathcal{D}_1^\ell$ is a mixture of fewer than k_1 product distributions and $\mathcal{D}_2^\ell$ is a mixture of at most k_2 product distributions, we inductively have that $d_{\mathrm{TV}}(\mathcal{D}_1^\ell, \mathcal{D}_2^\ell) \leq \varepsilon/5$. In the exact same way we can show that we inductively have that $d_{\mathrm{TV}}(\mathcal{D}_1^u, \mathcal{D}_2^u) \leq \varepsilon/5$.

Now consider any event $\mathcal{S} \subseteq \{0,1\}^n$. We wish to bound

$$\left|\sum_{s\in\mathcal{S}}(\Pr_{\mathcal{D}_1}[s] - \Pr_{\mathcal{D}_2}[s])\right| \leq \left|\sum_{s\in\mathcal{S}:s_i=0}(\Pr_{\mathcal{D}_1}[s] - \Pr_{\mathcal{D}_2}[s])\right| + \left|\sum_{s\in\mathcal{S}:s_i=1}(\Pr_{\mathcal{D}_1}[s] - \Pr_{\mathcal{D}_2}[s])\right| \tag{7.16}$$

Because $x = \alpha_{1,i}(x-\ell_i) + \beta_{1,i}(u_i - x)$ for $\alpha_{1,i} = \frac{u_i}{u_i-\ell_i}$ and $\beta_{1,i} = \frac{\ell_i}{u_i-\ell_i}$, and $1-x = \alpha_{0,i}(x-\ell_i) + \beta_{0,i}(u_i-x)$ for $\alpha_{0,i} = \frac{1-u_i}{u_i-\ell_i}$ and $\beta_{0,i} = \frac{1-\ell_i}{u_i-\ell_i}$. For $b = 0,1$, we can thus use (7.14) to express $\Pr_{\mathcal{D}_j}[s]$ for $s_i = b$ as

$$\Pr_{\mathcal{D}_j}[s] = \alpha_{b,i}\cdot\mathbb{E}_{\mathcal{D}_j}[x_i]\cdot\Pr_{\mathcal{D}_j^\ell}[s'] + \beta_{b,i}\cdot\mathbb{E}_{\mathcal{D}_j}[x_i]\cdot\Pr_{\mathcal{D}_j^u}[s']$$

where s' denotes the substring of s outside of coordinate i. From this we see that

$$\begin{aligned}\Pr_{\mathcal{D}_1}[s] - \Pr_{\mathcal{D}_2}[s] &= \alpha_{b,i}\left(\mathbb{E}_{\mathcal{D}_1}[x_i]\Pr_{\mathcal{D}_1^\ell}[s'] - \mathbb{E}_{\mathcal{D}_2}[x_i]\Pr_{\mathcal{D}_2^\ell}[s']\right) + \beta_{b,i}\left(\mathbb{E}_{\mathcal{D}_1}[x_i]\Pr_{\mathcal{D}_1^u}[s'] - \mathbb{E}_{\mathcal{D}_2}[x_i]\Pr_{\mathcal{D}_2^u}[s']\right)\\ &= \alpha_{b,i}\cdot\mathbb{E}_{\mathcal{D}_1}[x_i]\left(\Pr_{\mathcal{D}_1^\ell}[s'] - \Pr_{\mathcal{D}_2^\ell}[s']\right) + \beta_{b,i}\cdot\mathbb{E}_{\mathcal{D}_1}[x_i]\left(\Pr_{\mathcal{D}_1^u}[s'] - \Pr_{\mathcal{D}_2^u}[s']\right) \pm \alpha_{b,i}\eta\Pr_{\mathcal{D}_2^\ell}[s'] \pm \beta_{b,i}\eta\Pr_{\mathcal{D}_2^u}[s'].\end{aligned}$$

Note that

$$\alpha_{b,i}\cdot\mathbb{E}_{\mathcal{D}_1}[x_i], \beta_{b,i}\cdot\mathbb{E}_{\mathcal{D}_1}[x_i] \leq 1$$

because $u_i - \ell_i$ is an obvious upper bound on $\mathbb{E}_{\mathcal{D}_1}[x_i]$. We can thus bound (7.16) by

$$2d_{\mathrm{TV}}(\mathcal{D}_1^\ell, \mathcal{D}_2^\ell) + 2d_{\mathrm{TV}}(\mathcal{D}_1^u, \mathcal{D}_2^u) + \eta(\alpha_{0,i} + \alpha_{1,i} + \beta_{0,i} + \beta_{1,i}) \leq 4\varepsilon/5 + \frac{4\eta}{\delta\varepsilon/9k} \leq \varepsilon,$$

thus completing the induction. □

Henceforth fix $\eta(n, k_1+k_2, \varepsilon)$ to be the η in Lemma 7.5.6.

7.5.5 Collapsing Ill-conditioned Moment Matrices

Lastly, we illustrate how to use Lemma 7.5.6 to implement the same recursive conditioning strategy that we used in N-List to learn mixtures of subcubes, deferring the details to Appendix 7.8. Just as we showed in Lemma 7.2.7 that we can collapse mixtures of k product distributions to mixtures of fewer product distributions provided their moment matrices are of rank less than k, here we show that we can do the same if their moment matrices are ill-conditioned.

Lemma 7.5.10. *The following holds for any $\eta > 0$. Let $\mathcal{D}$ be a mixture of k product distributions realized by mixing weights π and marginals matrix $\mathbf{m}$ such that*

$$\sigma^{\infty}_{\min}(\mathbf{M}) \leq \frac{\eta \cdot \sqrt{2}}{3k^2}.$$

Then there exists $\mathcal{D}'$ a mixture of at most $k-1$ product distributions realized by mixing weights π' and marginals matrix $\mathbf{m}'$ such that $|\mathbb{E}_{\mathcal{D}}[x_S] - \mathbb{E}_{\mathcal{D}'}[x_S]| \leq \eta$ for all $|S| \leq k$. In particular, if we take $\eta = \eta(n, 2k, \varepsilon)$, then by Lemma 7.5.6, $d_{TV}(\mathcal{D}, \mathcal{D}') \leq \varepsilon$.

To prove this, we require the following basic fact similar in spirit to the proof of Lemma 7.2.7.

Lemma 7.5.11. *For any $v \in \mathbb{R}^k$, there exists $t \in \mathbb{R}$ with $|t| \leq \sqrt{k}/\|v\|_2$ for which $\pi - t \cdot v$ has a zero entry and lies in $[0,1]^k$.*

Proof. If π already has a zero entry, then we are done. Otherwise π lies in the interior of the box $[0,1]^k$. Consider the line through π given by $\{\pi - t \cdot v\}_{t \in \mathbb{R}}$. This will intersect the boundary of the box in two points, which correspond to values t for which $\pi - t \cdot v$ has a zero entry. The bound on $|t|$ follows from the fact that the diameter of $[0,1]^k$ is $\sqrt{k}$. □

We will move π in the direction of the minimal singular vector corresponding to $\sigma^{\infty}_{min}(\mathbf{M})$ and argue by Lemma 7.5.6 that the resulting mixture of at most $k-1$ product distributions is close to $\mathcal{D}$.

Proof of Lemma 7.5.10. Let $\sigma^{\infty}_{\min}(\mathbf{M}) = \tau$. Let $v \in \mathbb{R}^k$ be the vector for which $\|\mathbf{M} \cdot v\|_{\infty} = \tau$ and $\|v\|_{\infty} = 1$. Denote by $S_+, S_- \subseteq [k]$ the coordinates on which v is positive or negative respectively, and let $i \in [k]$ be the coordinate for which $v_i = 1$, without loss of generality.

Let $Z_+ = \sum_{j\in S_+} v_j$ and $Z_- = -\sum_{j\in S_-} v_j$ and note that $|Z_+ - Z_-| \le \tau$ because $\mathbf{1}$ is a row of $\mathbf{M}$ and $1 \le |Z_+| \le k$ because $v_i = 1$.

Define $\pi_+ = v_{S_+}/Z_+, \pi_- = -v_{S_-}/Z_-, \mathbf{m}_+ = \mathbf{m}|^{S_+}, \mathbf{m}_- = \mathbf{m}|^{S_-}$ and let $\mathcal{D}_+$ and $\mathcal{D}_-$ be the mixtures of $|S_+|$ and $|S_-|$ product distributions realized by $(\pi_+, \mathbf{m}_+)$ and $(\pi_-, \mathbf{m}_-)$.

We claim that it suffices to show that

$$\left|\mathbb{E}_{\mathcal{D}_+}[x_S] - \mathbb{E}_{\mathcal{D}_-}[x_S]\right| \le \eta \cdot \sqrt{2}/k \tag{7.17}$$

for all $|S| \le k$. Indeed, define v^* to be the rescaling of v by Z_+ in coordinates S_+ and by Z_- in coordinates S_- (i.e. the appropriate concatenation of π_1 and $-\pi_2$). By Cauchy-Schwarz, $\|v^*\| \le \sqrt{2k}$, so by Lemma 7.5.11 there exists a $t \in \mathbb{R}$ with $|t| \le \sqrt{k}/\|v^*\|_2 \le k/\sqrt{2}$ for which $\pi - t \cdot v^*$ has at most $k-1$ nonzero coordinates. Moreover, because the sum of the entries in v^* is zero by design, $\pi - t \cdot v^* \in \Delta^k$. Let $\pi' \in \mathbb{R}^{k-1}$ be the nonzero part of π and $\mathbf{m}'$ be the corresponding columns of $\mathbf{m}$, and let $\mathcal{D}'$ be the mixture of at most $k-1$ product distributions realized by $(\pi', \mathbf{m}')$. It is clear that

$$|\mathbb{E}_{\mathcal{D}}[x_S] - \mathbb{E}_{\mathcal{D}'}[x_S]| = t \cdot \left|\mathbb{E}_{\mathcal{D}_+}[x_S] - \mathbb{E}_{\mathcal{D}_-}[x_S]\right|, \tag{7.18}$$

so if (7.17) held, then by (7.18) and Lemma 7.5.6, $d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}') \le \varepsilon$ as desired.

It remains to show (7.17). We know that $\|\mathbf{M} \cdot v\|_\infty \le \tau$, and

$$\begin{aligned}\left|\mathbb{E}_{\mathcal{D}_+}[x_S] - \mathbb{E}_{\mathcal{D}_-}[x_S]\right| &= \left|\frac{1}{Z_+}(\mathbf{M}_+)_S v_{S_+} + \frac{1}{Z_-}(\mathbf{M}_-)_S v_{S_-}\right| \\ &= \left|\frac{1}{Z_+}\mathbf{M}_S v + \left(\frac{1}{Z_-} - \frac{1}{Z_+}\right)(\mathbf{M}_-)_S v_{S_+}\right| \\ &\le \tau + 2\tau k \le 3\tau k,\end{aligned}$$

so (7.17) holds as long as $\tau \le \frac{\eta\cdot\sqrt{2}}{3k^2}$. □

In Appendix 7.8 we show how to put all of these ingredients together to learn a mixture of product distributions given *arbitrary* $\varepsilon_{\mathrm{samp}}$-close estimates of its low-degree moments (not just estimates obtained by sampling), so in particular if the gridding procedure described in Section 7.5.3 fails because $\sigma^{\infty}_{\min}(\mathbf{M}|_{\mathcal{R}'(J\cup\{i\})})$ is small for some $i \notin J$, where J indexes a barycentric spanner for the rows of $\mathbf{m}$, then Lemma 7.5.10 tells us that we can learn $\mathcal{D}|x_{J\cup\{i\}} = s$ for each $s \in \{0,1\}^{|J\cup\{i\}|}$ by instead recursively learning distributions $\mathcal{D}_s$ which are mixtures of at most $k-1$ product distributions that are $\varepsilon_{\text{samp}}$-close in low-degree moments to $\mathcal{D}|x_{J\cup\{i\}} = s$ and ε-close in total variation distance.

7.5.6 Comparison to Feldman-O'Donnell-Servedio's Algorithm

The algorithm of Feldman, O'Donnell and Servedio [FOS05] also uses brute-force search to find a basis for the rows of $\mathbf{m}$. However instead of constructing a barycentric spanner they construct a basis that is approximately as well-conditioned as $\mathbf{m}$. Their algorithm proceeds by gridding the entries $\mathbf{m}|_J$. The key difference between their approach and ours is that their gridding requires granularity $O((\varepsilon/n)^k)$ while ours requires only $O(\varepsilon/n)$. The reason is that they try to solve for the other rows of $\mathbf{m}$ in the same way that we do in (7.2) when learning mixtures of subcubes, that is, by solving a system of equations for each $i \notin J$ with coefficients given by row $\mathbf{m}_i$. They require granularity $O((\varepsilon/n)^k)$ to account for $\mathbf{m}$ being ill-conditioned. Just as we showed we could assume in our algorithm for mixtures of subcubes that the mixture weights had a gap of $\rho = 2^{-O(k^2)}$, [FOS05] showed they can assume that $\mathbf{m}$ has a *spectral* gap of $O(\varepsilon/n)$ by brute-forcing singular vectors of $\mathbf{m}$ and appending them to $\mathbf{m}$ to make it better conditioned. Such a spectral gap corresponds in the worst case to an $\mathbf{m}$ that is $O((\varepsilon/n)^k)$-well-conditioned, which in turn ends up as the granularity in their gridding procedure. As a result, the bottleneck in the algorithm of [FOS05] is the $(n/\varepsilon)^{O(k^3)}$ time spent just to grid the entries of $\mathbf{m}|_J$.

In comparison, we save a factor of k in the exponent of the running time by only $O(\varepsilon/n)$-gridding the entries of $\mathbf{m}|_J$. The reason is that we solve for the remaining rows of $\mathbf{m}$ not by solving systems of equations with coefficients in the rows $\mathbf{m}_i$ for $i \notin J$, but by expressing these rows $\mathbf{m}_i$ as linear combinations of the rows of $\mathbf{m}|_J$, where the linear combinations have bounded coefficients. This leverages higher order multilinear moments to make the linear system better conditioned. We estimate these coefficients by solving the regression problem (7.13), and the coefficients are accurate so long as the sampling error is $O(\varepsilon/n)$ times the condition number of $\mathbf{M}|_{\mathcal{R}'(J\cup\{i\})}$ for J the barycentric spanner of the rows of $\mathbf{m}$ and any $i \notin J$. So in our algorithm, the bottlenecks leading to a k^2 dependence in the exponent are

(1) $O(\varepsilon/n)$-gridding all $O(k^2)$ entries of $\mathbf{m}|_J$, (2) brute-forcing $O(k)$ coordinates to condition in every one of the $\le k$ recursive steps, (3) using degree-$O(k^2)$ subsets in $\mathcal{R}'(J \cup \{i\})$ to ensure that when we condition on each of at most k subsequent subsets $J \cup \{i\}$, the resulting mixtures are all close in low-order moments to mixtures of fewer components.

7.6 Appendix: Learning via Sampling Trees

Recall that our algorithms for learning mixtures of subcubes and general mixtures of product distributions over $\{0,1\}^n$ both work by first running an initial subroutine that will successfully learn the distribution if certain non-degeneracy conditions are met (e.g. $\mathrm{rank}(\mathbf{M}|_{\mathcal{R}'(J\cup\{i\})}) = k$ or $\sigma^{\infty}_{\min}(\mathbf{M}|_{\mathcal{R}(J\cup\{i\})})$ is sufficiently large for all $i \notin J$ and all realizations of $\mathcal{D}$). If this initial subroutine fails, some non-degeneracy condition is not met, so we can condition on all assignments to a small set of coordinates and recursively learn the resulting conditional distributions which are guaranteed to be simpler. Before analyzing these algorithms in detail, we make this recursive procedure precise.

Definition 7.6.1. *A* sampling tree *$\mathcal{T}$ is a tree whose vertices $v_{S,s}$ correspond to tuples (S, s) for $S \subseteq [n]$ and $s \in \{0,1\}^{|S|}$, with the root being $v_{\emptyset,\emptyset}$. For every node $v_{S,s}$, either $v_{S,s}$ is a leaf corresponding to a distribution $\mathcal{D}_{S,s}$ over $\{0,1\}^{n-|S|}$, or there is a $W \subseteq [n]\backslash S$ for which $v_{S,s}$ is connected to children $v_{S\cup W, s\oplus t}$ for all $t \in \{0,1\}^{|W|}$ via edges of weight $w_{S,W,s,t}$. For any non-leaf vertex $v_{S,s}$, $\sum_{W,t} w_{S,W,s,t} = 1$.*

$\mathcal{T}$ gives an obvious procedure for sampling from $\{0,1\}^n$: randomly walk down the tree according to the edge weights, and sample from the distribution corresponding to the leaf you end up at. We call the resulting distribution the distribution associated to $\mathcal{T}$. *We can analogously define the distributions associated to (subtrees rooted at) vertices of $\mathcal{T}$.*

Given a mixture of product distributions $\mathcal{D}$, our learning algorithm will output a sampling tree $\mathcal{T}$ where for each $S \subseteq [n]$ and $s \in \{0,1\}^{|S|}$, the subtree rooted at $v_{S,s}$ corresponds to the distribution the algorithm recursively learns to approximate the posterior distribution $(\mathcal{D}|x_S = s)$. If $v_{S,s}$ is any vertex of $\mathcal{T}$, we can learn the subtree rooted at $v_{S,s}$ as follows. First use rejection sampling on $\mathcal{D}$ to get enough samples of $\mathcal{D}|x_S = s$ that all moment estimates are $\varepsilon_{\mathrm{samp}}$-close to their true values. We can then run our initial subroutine for learning non-degenerate mixtures.

It either outputs both a list $\mathcal{M}$ of candidate mixtures for $(\mathcal{D}|x_S = s)$ and a list $\mathcal{U}$ of subsets of coordinates $W \subseteq [n]\backslash S$ to condition on, or it outputs FAIL if we've already recursed r times and yet $(\mathcal{D}|x_S = s)$ is not close to or exactly realizable by a mixture of at most $k - r$ product distributions.

If the output is not FAIL, the guarantee is that either some mixture from $\mathcal{M}$ is $O(\varepsilon)$-close to $(\mathcal{D}|x_S = s)$, or some $W \in \mathcal{U}$ satisfies that $(\mathcal{D}|x_{S\cup W} = s \circ t)$ is "simpler" for every $t \in \{0,1\}^{|W|}$ (i.e. close to or exactly realizable as a mixture of fewer product distributions). In the latter case, the algorithm guesses W and tries to recursively learn each $(\mathcal{D}|x_{S\cup W} = s \circ t)$. For every guess W, the algorithm gets candidate sampling trees $\mathcal{D}_{v_{S\cup W}, s\circ t}$ to connect to $v_{S,s}$. Moreover, by guarantees we prove about the initial subroutine for learning non-degenerate mixtures, we do not need to recurse more than k more times from the root $v_{\emptyset,\emptyset}$.

If the output is FAIL, this means we incorrectly guessed W at some earlier recursive step.

So in total we get a pool of $|\mathcal{M}| + |\mathcal{U}|$ candidate distributions, one of which is guaranteed to be $O(\varepsilon)$-close to $(\mathcal{D}|x_S = s)$. It then remains to pick out a candidate which is $O(\varepsilon)$-close, which can be done via the following well-known fact.

Lemma 7.6.2 (Scheffé tournament, see e.g. [DL01]). *Given sample access to a distribution $\mathcal{D}$, and given a list $\mathcal{L}$ of distributions $\mathcal{D}'$ at least one of which satisfies $d_{TV}(\mathcal{D}, \mathcal{D}') \le \varepsilon$, there is an algorithm* Select*$(\mathcal{L}, \mathcal{D})$ which outputs a distribution $\mathcal{D}'' \in \mathcal{L}$ satisfying $d_{TV}(\mathcal{D}, \mathcal{D}'') \le 9.1\varepsilon$ using $O(\varepsilon^{-2} \log |\mathcal{L}|)$ samples from $\mathcal{D}$ and in time $O(\varepsilon^{-2}|\mathcal{L}|^2 \log |\mathcal{L}|T)$, where T is the time to evaluate the pdf of any distribution in $\mathcal{L}$ on a given point.*

Remark 7.6.3. *For mixtures of* subcubes*, our initial subroutine for learning non-degenerate mixtures has stronger guarantees: it outputs a* single *mixture which is guaranteed to be close to $(\mathcal{D}|x_S = s)$, a collection $\mathcal{U}$ of subsets W, or FAIL.*

One minor subtlety is that for certain S, s, $\Pr_{\mathcal{D}}[x_S = s]$ may be so small that rejection sampling will not give us enough samples from $(\mathcal{D}|x_S = s)$, and the subtree rooted at $v_{S,s}$ will end up looking very different from $(\mathcal{D}|x_S = s)$. But this is fine because in sampling from $\mathcal{T}$, we will reach $v_{S,s}$ so rarely that if $\mathcal{D}^*$ is the distribution associated to $\mathcal{T}$, $d_{\mathrm{TV}}(\mathcal{D}^*, \mathcal{D})$ is still very small.

The above discussion is summarized in Algorithm 24 below, where NonDegenerateLearn is the abovementioned initial subroutine for learning non-degenerate mixtures. Formally, it outputs a list $\mathcal{M}$ of candidate mixtures as well as a list $\mathcal{U}$ of subsets $W \subseteq [n]\backslash S$ to be conditioned on. The list might contain a distribution close to $\mathcal{D}$, but if not, $\mathcal{U}$ will contain some W such that conditioning on $x_W = s$ for any $s \in \{0,1\}^{|W|}$ will yield a "simpler" distribution.

Algorithm 24: N-List($\mathcal{D}, S, s, k$)

```
Input: Mixture of subcubes/product distributions $\mathcal{D}$, $S \subseteq [n], s \in \{0,1\}^{|S|}$, counter
       $k$
Output: List of sampling trees rooted at node $v_{S,s}$, one of which is guaranteed to
        be close to $(\mathcal{D}|x_S = s)$
$\mathcal{S} \leftarrow \emptyset$.
Draw $2N/\tau_{trunc}$ samples $y$ from $\mathcal{D}$ and keep those for which $y_S = s$ as samples from
  $(\mathcal{D}|x_S = s)$.
Run NonDegenerateLearn($\mathcal{D}|x_S = s$, $k$).
if output is FAIL then
    return FAIL.
else
    /* output is list $\mathcal{M}$ of candidate mixtures and/or list $\mathcal{U}$ of
       candidate subsets $W \subseteq [n]\backslash S$ to condition on                    */
    for each mixture in $\mathcal{M}$ do
        Add to $\mathcal{S}$ the sampling tree given by the single node $v_{S,s}$ with distribution
          equal to this mixture.
    if $k > 1$ then
        for $W \in \mathcal{U}$ do
            for $t \in \{0,1\}^{|W|}$ do
                Run N-List($\mathcal{D}, S \cup W, s \circ t, k-1$) to get some list of sampling trees $\mathcal{T}_t$
                  or FAIL.
                If we get FAIL for any $t$, skip to the next $W$.
            Empirically estimate $\mathbb{E}_{y \in \mathcal{D}}[y_W = t | y_S = s]$ to within $\delta_{edge}$ using the
              samples from $(\mathcal{D}|x_S = s)$.
            For each $\mathcal{T}_t$: connect $v_{S,s}$ to the root $v_{S \cup W, s \circ t}$ of $\mathcal{T}_t$ with edge weight
              $w_{S,W,s,t}$ for every $t \in \{0,1\}^{|W|}$ and add this sampling tree to $\mathcal{S}$.
return Select($\mathcal{S}, \mathcal{D}, \varepsilon_{select}$).
```

Our implementations of NonDegenerateLearn will interact solely with estimates of moments of the input distribution, so in our analysis it will be convenient to assume that these estimates are accurate.

Definition 7.6.4. *Let $\varepsilon_{samp}(\cdot) : \mathbf{Z}_+ \to [0,1]$ be a decreasing function. We say a run of* NonDegenerateLearn *on some counter k and some $(\mathcal{D}|x_S = s)$ is* $\varepsilon_{\mathrm{samp}}(k)$-sample-rich *if enough samples are drawn from $\mathcal{D}$ that all moment estimates used are $\varepsilon_{samp}(k)$-close to their true values.*

For $\delta_{edge}, \tau_{trunc} > 0$, we say a run of N-List *on distribution $\mathcal{D}$ is* $(\varepsilon_{\mathrm{samp}}(\cdot), \delta_{edge}, \tau_{trunc})$-sample-rich *if enough samples are drawn from $\mathcal{D}$ that every invocation of* NonDegenerateLearn *on counter k and $(\mathcal{D}|x_S = s)$ for which $Pr_{y\sim\mathcal{D}}[y_S = s] \geq \tau_{trunc}$ is $\varepsilon_{samp}(k)$-sample rich, and such that every transition probability computed in an iteration of Step 14 is estimated to within δ_{edge} error.*

Because the runtimes of our algorithms for learning mixtures of subcubes and mixtures of product distributions are rather different, the kinds of guarantees we need for NonDegenerateLearn are somewhat different. We therefore defer proofs of correctness of N-List for mixtures of subcubes and general mixtures to Appendix 7.7 and Appendix 7.8 respectively. We can however give a generic runtime analysis for N-List now. We will use the following basic facts.

Fact 7.6.5. *Suppose $\mathbb{E}_{\mathcal{D}}[x_S = s] \geq \tau_{trunc}$. Then if $2N/\tau_{trunc}$ samples are drawn from $\mathcal{D}$, with probability $1 - e^{-N/4}$ at least N samples x will satisfy $x_S = s$.*

Fact 7.6.6. *Fix $S \subseteq [m]$. If $(3/\varepsilon^2)\ln(2/\rho)$ samples are taken from a distribution $\mathcal{D}$ over $\{0,1\}^m$, then $\left|\widetilde{\mathbb{E}}_{\mathcal{D}}[x_S] - \mathbb{E}_{\mathcal{D}}[x_S]\right| > \varepsilon$ with probability at most ρ.*

Lemma 7.6.7. *Suppose* NonDegenerateLearn *on any input distribution and counter k always uses at most Z different moments, returns $\mathcal{M}$ of size at most M and $\mathcal{U}$ of size at most U and consisting of subsets of size at most S, and takes time at most $T(r)$. If $\delta_{edge} \leq \varepsilon_{samp}(k)$, then achieving an $(\varepsilon_{samp}(\cdot), \delta_{edge}, \tau_{trunc})$-sample-rich run of* N-List *on a given distribution and counter k with probability $1 - \delta$ requires*

$$O(\varepsilon_{samp}(k)^{-2}\ln(1/\delta)\ln(Z) + \varepsilon_{select}^{-2}\cdot\mathrm{poly}(n,k)\log(M+U))\cdot(2^{Sk}U^k)^{1+o(1)}/\tau_{trunc} + T(k)\cdot 2^{Sk}U^k$$

time and

$$O(\varepsilon_{samp}(k)^{-2}\ln(1/\delta)\ln(Z)+\varepsilon_{select}^{-2}\log(M+U))\cdot(2^{Sk}U^k)^{1+o(1)}/\tau_{trunc}$$

samples.

Proof. The only places where we need to take samples are to estimate N moments in each invocation of NonDegenerateLearn, to estimate transition probabilities $\Pr_{y\sim\mathcal{D}}[y_j = t|x_S = s]$ in each iteration of Step 14, and to run Select. Denote by $N_1(k)$, $N_2(k)$, $N_3(k)$ the maximum possible number of invocations of NonDegenerateLearn, estimations of transition probabilities, and the number of invocations of Select in a run of N-List on a distribution and a counter k. Then $N_1(k) \leq 1 + N_1(k-1)\cdot U \cdot 2^S$, $N_2(k) \leq U \cdot 2^S + N_2(k-1)\cdot U\cdot 2^S$, and $N_3(k) \leq 1 + N_3(k-1)\cdot U\cdot 2^S$. But $N_1(1), N_3(1) = 1$ and $N_2(1) = 0$, so unwinding the recurrences and noting that $2^S \cdot U \geq 2$, we get that $N_1(k), N_2(k), N_3(k) \leq 2^{Sk}\cdot U^k$.

For NonDegenerateLearn and the transition probabilities, we need to estimate at most Z moments of some $(\mathcal{D}|x_S = s)$ in each invocation of NonDegenerateLearn and $N_2(k)$ statistics of the form $\Pr_{y\sim(\mathcal{D}|x_S=s)}[y_T = t]$, and we require that for S, s such that $\Pr_{\mathcal{D}}[x_S = s] \geq \tau_{trunc}$, our estimates are $\varepsilon_{\mathrm{samp}}(k)$-close. For such S, s, by Fact 7.6.5 we can simulate N draws from $(\mathcal{D}|x_S = s)$ using $2N/\tau_{trunc}$ draws from $\mathcal{D}$ with probability at least $1 - e^{-N/4}$. By Fact 7.6.6, if we set $N = (3/\varepsilon_{\mathrm{samp}}(k)^2)\ln(2/\rho)$ for some $\rho > 0$, then we can estimate some $\Pr_{y\sim(\mathcal{D}|x_S=s)}[y_T = t]$ to within error $\varepsilon_{\mathrm{samp}}(k)$ with probability at least $1-\rho$. In this case, $N > 4\ln(1/\rho)$, so the probability that we fail to estimate this statistic to within error $\varepsilon_{\mathrm{samp}}(k)$ is at most 2ρ. By a union bound over all $Z\cdot N_1(k) + N_2(k)$ statistics, the probability we fail to get an $(\varepsilon_{\mathrm{samp}}(\cdot), \delta_{edge}, \tau_{trunc})$-sample-rich run of N-List is at most $2\rho(Z\cdot N_1(k) + N_2(k)) \leq 2\rho\cdot(Z+1)2^{Sk}U^k$, so by taking $\rho = \delta/(4(Z+1)2^{Sk}U^k)$, we ensure the run is $(\varepsilon_{\mathrm{samp}}(\cdot), \delta_{edge}, \tau_{trunc})$-rich with probability at least $1-\delta/2$. In total, this requires

$$(2N/\tau_{trunc})\cdot(N_1(k)+N_2(k)) = O((1/\varepsilon_{\mathrm{samp}}(k)^2)\cdot(2^{Sk}U^k)^{1+o(1)}\cdot\ln(Z)\ln(1/\delta)/\tau_{trunc})$$

samples. In addition to drawing samples for NonDegenerateLearn and the transition probabilities, we also need time at most $T(k)$ for each invocation of NonDegenerate-

Learn, for a total of $T(k) \cdot 2^{Sk}U^k$ time.

For Select, we need to use Lemma 7.6.2 $N_3(k)$ times. Note that the list of candidates is always at most $M+U$, so for each invocation of Select on $(\mathcal{D}|x_S = s)$ for which $\Pr_{\mathcal{D}}[x_S = s] \geq \tau_{trunc}$, we require $O(\varepsilon_{select}^{-2}\log(M+U))$ samples from $(\mathcal{D}|x_S = s)$, which can be done using $O(\varepsilon_{select}^{-2}\log(M+U)/\tau_{trunc})$ samples from $\mathcal{D}$. In total, this requires $N_3(k) \cdot O(\varepsilon_{select}^{-2}\log(M+U)/\tau_{trunc}) = O(\varepsilon_{select}^{-2}2^{Sk}U^k\log(M+U)/\tau_{trunc})$ samples. The time to evaluate the pdf of a sampling tree is obviously $\mathrm{poly}(n,k)$, so $N_3(k)$ invocations of Select requires time at most $O(\varepsilon_{select}^{-2}) \cdot (2^{Sk}U^{k+2}\log(M+U)) \cdot \mathrm{poly}(n,k)$.

Putting this all together gives the desired time and sample complexity. □

7.7 Appendix: Learning Mixtures of Subcubes

N-List and GrowByOne in Section 7.3 were described under the assumption that we had exact access to the accessible entries of $\mathbf{C}$, when in reality we only have access to them up to some sampling noise $\varepsilon_{\mathrm{samp}} > 0$ (we fix this parameter $\varepsilon_{\mathrm{samp}}$ later). In this section, we show how to remove the assumption of zero sampling noise and thereby give a complete description of the algorithm for learning mixtures of subcubes.

Throughout this section, we fix a $[\tau_{small}, \tau_{big}]$-avoiding rank-$k$ realization of $\mathcal{D}$ by mixing weights π and marginals matrix $\mathbf{m}$ such that $\mathbf{M}'$ has k' columns. Here, recall that $\mathbf{M}'$ denotes the submatrix of $\mathbf{M}$ of columns corresponding to mixing weights that are at least τ_{big}. We will use $\widetilde{\mathbb{E}}[x_S]$ to denote any $\varepsilon_{\mathrm{samp}}$-close estimate of $\mathbb{E}x_S]$ and $\tilde{\mathbf{C}}$ to denote a matrix consisting of $\varepsilon_{\mathrm{samp}}$-close estimates of the accessible entries of $\mathbf{C}$. Note that we only ever use particular submatrices of $\tilde{\mathbf{C}}$ of reasonable size in our algorithm, so at no point will we need to instantiate all entries of $\tilde{\mathbf{C}}$.

7.7.1 Robustly Building a Basis

Here we describe and prove guarantees for a sampling noise-robust implementation of GrowByOne. Recall that every time we reach step 8 of GrowByOne, we are appending to the basis $\mathcal{B} = \{T_1, ..., T_r\}$ a subset of $\{T_1 \cup \{i\}, ..., T_r \cup \{i\}\}$ so that the corresponding columns in $\mathbf{C}|_{\mathcal{R}'(J\cup\{i\})}$ form a basis for columns $T_1, ..., T_r, T_1 \cup \{i\}, ..., T_r \cup \{i\}$, where as usual

$J = T_1 \cup \cdots \cup T_r$.

One way to pick out the appropriate columns to add is to solve at most r linear systems of the following form. Suppose we have already added some indices to $\mathcal{B}$ so that the corresponding columns of $\mathbf{C}|_{\mathcal{R}'(J\cup\{i\})}$ span columns $T_1, ..., T_r, T_1 \cup \{i\}, ..., T_{m-1} \cup \{i\}$ for some $m \le r$.

To check whether to add some $T' \subseteq [n]$ to $\mathcal{B}$, we could simply check whether there exists $\alpha^{T'} \in \mathbb{R}^{|\mathcal{B}|}$ for which

$$\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J\cup T')}\alpha^{T'} = \mathbf{C}|^{T'}_{\mathcal{R}'(J\cup T')}. \tag{7.19}$$

In reality however, we only have access to $\tilde{\mathbf{C}}$, so instead of solving (7.19), we will solve the regression problem

$$\tilde{\alpha}^{T'} \triangleq \operatorname*{argmin}_{\alpha\in\mathbb{R}^{|\mathcal{B}|}} \|\tilde{\mathbf{C}}|^{\mathcal{B}}_{\mathcal{R}'(J\cup T')}\alpha - \tilde{\mathbf{C}}|^{T'}_{\mathcal{R}'(J\cup T')}\|_\infty. \tag{7.20}$$

Denote by $\varepsilon(\tilde{E}, T', \mathcal{B})$ the corresponding L^∞ error of the optimal solution; where the context is clear, we will refer to this as ε_{err}.

We can now give the following robust version of Lemma 7.2.9 and Lemma 7.2.10.

Lemma 7.7.1 (Robust version of Lemma 7.2.9 and Lemma 7.2.10)**.** *There exist large enough constants $c_{19}, c_{16} > 0$ for which the following holds. Fix a $[\tau_{small}, \tau_{big}]$-avoiding rank-$k$ realization of $\mathcal{D}$, and let $\varepsilon_{samp} < k^{-c_{19}k^2}\tau_{big}$ and $\rho = k^{-c_{16}k^2}$. Let $\mathcal{B} = \{T_1, ..., T_r\}$ be such that the rows of $\mathbf{M}'|_{\mathcal{B}}$ are linearly independent, and fix $T' \subseteq [n]$ for which $|J \cup T'| \le k'$, where k' is the number of columns of $\mathbf{M}'$ and $J = T_1 \cup \cdots \cup T_r$. Let $\tilde{C}_2|^{\mathcal{B}}_{\mathcal{R}'(J\cup T')}$ be any matrix of moment estimates satisfying $\|\tilde{\mathbf{C}}|^{\mathcal{B}}_{\mathcal{R}'(J\cup T')} - \mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J\cup T')}\|_{\max} \le \varepsilon_{samp}$.*

- *If $\operatorname{rank}(\mathbf{M}'|_{\mathcal{R}'(J\cup T')}) = k'$ and $\mathbf{M}'_{T'}$ is not in the span of $\{\mathbf{M}'_{T'}\}_{T\in\mathcal{B}}$, then $\varepsilon_{err} \ge \frac{1}{2}k^{-c_{17}k^2}\tau_{big}$.*

- *If $\mathbf{M}'_{T'}$ is in the span of $\{\mathbf{M}'_T\}_{T\in\mathcal{B}}$ so that there exists $\alpha^{T'} \in \mathbb{R}^{|\mathcal{B}|}$ for which*

$$\mathbf{M}'_{T'} = \sum_{T\in\mathcal{B}} \alpha^{T'}_T \mathbf{M}'_T, \tag{7.21}$$

  then $\varepsilon_{err} < k^{-c_{20}k^2}\tau_{big}$ for some $c_{20} > c_{17}$.

Proof. First suppose that $\mathrm{rank}(\mathbf{M}'|_{\mathcal{R}'(J\cup T')}) = k'$ and that there exists no coefficients $\alpha^{T'}$ for which (7.21) holds, so $F \triangleq (\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J\cup T')} \| \mathbf{C}|^{T'}_{\mathcal{R}'(J\cup T')})$ satisfies the hypotheses of Lemma 7.3.14. Also define $\tilde{F} \triangleq (\tilde{\mathbf{C}}|^{\mathcal{B}}_{\mathcal{R}'(J\cup T')} \| \tilde{\mathbf{C}}|^{T'}_{\mathcal{R}'(J\cup T')})$ and $\tilde{\alpha}'^{T'} = (\tilde{\alpha}^{T'} \| -1)$ so that $\varepsilon_{err} = \|\tilde{F}\tilde{\alpha}'^{T'}\|_\infty$. Applying Lemma 7.3.14 to $\tilde{F}$, we get $\sigma^\infty_{\min}(\tilde{F}) \geq \frac{1}{2}k^{-c_{17}k^2} \cdot \tau_{big}$ provided $\varepsilon_{\mathrm{samp}} \leq \frac{1}{2}k^{-c_{17}k^2-1}\tau_{big}$. So we can ensure that

$$\varepsilon_{err} = \|\tilde{F}\tilde{\alpha}'^{T'}\|_\infty \geq \sigma^\infty_{\min}(\tilde{F})\|\tilde{\alpha}'^{T'}\|_\infty \geq \frac{1}{2}k^{-c_{17}k^2} \cdot \tau_{big}. \tag{7.22}$$

Now suppose instead that there do exist coefficients $\alpha^{T'}$ for which (7.21) holds. We claim that ε_{err} will not exceed the lower bound computed in (7.22). Indeed, note that

$$\|\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J\cup T')}\alpha^{T'} - \mathbf{C}|^{T'}_{\mathcal{R}'(J\cup T')}\|_\infty = \|\sum_{\ell=k'+1}^{k} \pi^\ell \cdot \mathbf{M}|^\ell_{\mathcal{R}'(J\cup T')}\left((\mathbf{M}|^\ell_{\mathcal{B}})^\top \alpha^{T'} - (\mathbf{M}|^\ell_{T'})^\top\right)\|_\infty. \tag{7.23}$$

But for any $k'+1 \leq \ell \leq k$ and $T \subseteq J \cup T'$,

$$\|\pi^\ell \cdot \mathbf{M}|^\ell_{\mathcal{R}'(J\cup T')}(\mathbf{M}^\ell_T)^\top\|_\infty \leq \tau_{small}.$$

So by triangle inequality we can bound the right-hand side of (7.23) by

$$\sum_{\ell=k'+1}^{k}\left(\tau_{small} \cdot |\mathcal{B}| \cdot \|\alpha^{T'}\|_\infty + \tau_{small}\right) \leq 2k^2\tau_{small}\|\alpha^{T'}\|_\infty.$$

So we have that

$$\begin{aligned}
\varepsilon_{err} &\leq \|\mathbf{C}|^{\mathcal{B}}_{\mathcal{R}'(J\cup T')}\alpha^{T'} - \mathbf{C}|^{T'}_{\mathcal{R}'(J\cup T')}\|_\infty + \|\Delta|^{\mathcal{B}}\alpha^{T'}\|_\infty + \|\Delta^{T'}\|_\infty \\
&\leq 2k^2\tau_{small}\|\alpha^{T'}\|_\infty + k\varepsilon_{\mathrm{samp}}\|\alpha^{T'}\|_\infty + \varepsilon_{\mathrm{samp}} \\
&\leq (2k^2\tau_{small} + k\varepsilon_{\mathrm{samp}}) \cdot k^{c_{21}k^2}
\end{aligned}$$

for some $c_{21} > 0$, where in the last step we have bounded $\|\alpha^{T'}\|_\infty$ using Lemma 7.3.10:

$$\|\alpha^{T'}\|_\infty \leq \|(\mathbf{M}'|_{\mathcal{B}})^\top\alpha^{T'}\|_\infty / \sigma^\infty_{\min}((\mathbf{M}'|_{\mathcal{B}})^\top) \leq \|\mathbf{M}'_{T'}\|_\infty k^{c_{13}k^2} \leq k^{c_{13}k^2}.$$

We conclude that by picking $\rho = k^{-c_{16}k^2}$ and $\varepsilon_{\text{samp}} = k^{-c_{19}k^2}\tau_{big}$ small enough, then we will have $\varepsilon_{err} < k^{-c_{20}k^2}\tau_{big}$ for some $c_{20} > c_{17}$. □

Now we have a way to adapt GrowByOne to handle sampling noise as summarized in Algorithm 25. We do not know *a priori* the window $[\tau_{small}, \tau_{big}]$ used in the above analysis, so we include this as part of the input in GrowByOne and InSpan.

```
Algorithm 25: InSpan($\mathcal{D}, \mathcal{B}, T', \tau_{small}, \tau_{big}$)
  Input: Mixture of subcubes $\mathcal{D}$, certified full rank $\mathcal{B}$, $T' \subseteq [n]$, $[\tau_{small}, \tau_{big}]$
  Output: If rank($\mathbf{M}'|_{\mathcal{R}'(J \cup T')}$) $= k'$ for some realization of $\mathcal{D}$, the output is True if
          $\mathbf{M}'_{T'}$ lies in the row span of $\mathbf{M}'|_{\mathcal{B}}$, and False otherwise.
  Construct matrix $\tilde{E}$ with entries consisting of $\varepsilon_{\text{samp}}$-close empirical estimates of the
    entries of $E \triangleq \mathbf{C}|_{\mathcal{R}'(J \cup T')}$.
  Solve (7.20) and denote the corresponding $\varepsilon(\tilde{E}, T', \mathcal{B})$ by $\varepsilon_{err}$.
  if $\varepsilon_{err} \geq \frac{1}{2}k^{-c_{17}k^2}\tau_{big}$ then
      return False.
  else
      return True.
```

To put this in the context of the discussion in Section 7.2.3 and Section 7.3.2 note that the second statement in Lemma 7.7.1 — just like Lemma 7.2.9 — does not require $\text{rank}(\mathbf{M}'|_{\mathcal{R}'(J \cup \{i\})}) = k'$. Thus if we use ε_{err} to decide whether to add to $\mathcal{B}$, we will only ever add sets corresponding to rows of $\mathbf{M}'$ that are linearly independent. This is the sampling noise-robust analogue of being certified full rank. Furthermore, when we implement InSpan as above, the condition for termination in Step 10 of GrowByOne together with the condition for not returning FAIL in Step 13 constitute the sampling noise-robust analogue of being locally maximal.

Definition 7.7.2. *Given a collection $\mathcal{B} = \{T_1, T_2, \cdots, T_r\}$ of subsets we say that $\mathcal{B}$ is* robustly certified full rank *if* InSpan$(\mathcal{D}, \{T_1, ..., T_i\}, T_{i+1})$ *returns* True *for all $i = 1, ..., r - 1$.*

Definition 7.7.3. *Let $\mathcal{B} = \{T_1, T_2, \cdots, T_r\}$ be robustly certified full column rank. Let $J = \cup_i T_i$. Suppose there is no*

(1) $T' \subseteq J$ or

(2) $T' = T_i \cup \{j\}$ for $j \notin J$

for which InSpan$(\mathcal{D}, \mathcal{B}, T')$ *returns* False. *Then we say that* $\mathcal{B}$ *is* robustly locally maximal.

If GrowByOne with the above implementation of InSpan outputs some $\mathcal{B}^*$ with $J^* = \cup_{T \in \mathcal{B}^*} T$, then Lemma 7.7.1 implies that as long as $\text{rank}(\mathbf{M}'|_{\mathcal{R}'(J^* \cup \{i\})}) = k'$ for all $i \notin J^*$, $\mathcal{B}^*$ is both certified full rank and robustly certified full rank, as well as locally maximal and robustly locally maximal. Roughly this says that in non-degenerate mixtures, the robust and non-robust definitions coincide.

However, when $\text{rank}(\mathbf{M}'|_{\mathcal{R}'(J^* \cup T')}) < k'$ for some $T' \subseteq [n]$ and InSpan$(\mathcal{D}, \mathcal{B}^*, T')$ returns True, Lemma 7.7.1 tells us nothing about whether $\mathbf{C}|^{T_m \cup \{i\}}_{\mathcal{R}'(J^* \cup T')}$ lies inside the column span of $\mathbf{C}|^{\mathcal{B}^*}_{\mathcal{R}'(J^* \cup T')}$. So the output of GrowByOne under the above implementation of InSpan will not necessarily be certified full rank and locally maximal in the sense of Section 7.3. Still, it is not hard to modify the proofs of Lemmas 7.3.9 and 7.3.6 to obtain the following sampling noise-robust analogues.

Lemma 7.7.4 (Robust version of Lemma 7.3.9)**.** *Suppose* GrowByOne *has* InSpan *implemented as Algorithm 25 and has access to* ε_{samp}*-close estimates of any moment of* $\mathcal{D}$ *for* $\varepsilon_{samp} < k^{-c_{19}k^2}\tau_{big}$ *and* $\rho = k^{-c_{16}k^2}$. *If* GrowByOne *outputs* FAIL *and some set* J^*, *then* $\text{rank}(\mathbf{M}'|_{\mathcal{R}'(J^*)}) < k'$ *for some rank-*k *realization of* $\mathcal{D}$. *Otherwise,* GrowByOne *outputs* $\mathcal{B}^* = \{T_1, ..., T_r\}$, *and* $\mathcal{B}^*$ *is robustly certified full rank and robustly locally maximal.*

Proof. The proof of the lemma follows many of the steps in Lemma 7.3.9 but uses InSpan. Set J^* either to be the output of GrowByOne if it outputs FAIL, or if it outputs $\mathcal{B}^*$ then set $J^* = \cup_i T_i$. Now fix any rank-k realization of $\mathcal{D}$ and let $\mathbf{M}'$ be the corresponding moment matrix. Whenever the algorithm reaches Step 5 for $i \in J^*$, $\mathcal{B} = \{T_1, \cdots, T_r\}$, there are two possibilities. If $\text{rank}(\mathbf{M}'|_{\mathcal{R}'(J \cup \{i\})}) < k$, then $\text{rank}(\mathbf{M}'|_{J^*}) < k$ because J^* obviously contains $J \cup \{i\}$. Otherwise, inductively we know that by Lemma 7.7.1 that $\mathbf{M}'|_{\mathcal{B}}$ is a row basis for $\mathbf{M}'|_{2^J}$. So rows

$$T_1, \cdots, T_r, T_1 \cup \{i\}, \cdots, T_r \cup \{i\}$$

of $\mathbf{M}'$ span the rows of $\mathbf{M}'|_{2^{J \cup \{i\}}}$. If $\mathcal{B}'$ indexes a basis among these rows

$$T_1, \cdots, T_r, T_1 \cup \{i\}, \cdots, T_r \cup \{i\}$$

then by the second part of Lemma 7.7.1, $\textsc{InSpan}(\mathcal{D}, \mathcal{B}', T')$ outputs True for every $T' \subseteq J \cup \{i\}$. Step 8 of $\textsc{GrowByOne}$ simply finds such a $\mathcal{B}'$.

Therefore, when we exit the loop, either (a) the $\mathcal{B}^*$ we end up with at the end of $\textsc{GrowByOne}$ is such that $\textsc{InSpan}(\mathcal{D}, \mathcal{B}^*, T')$ outputs True for every $T' \subseteq J^*$ or (b) at some iteration of Step 3 J satisfies $\text{rank}(\mathbf{M}'|_{\mathcal{R}'(J)}) < k$ and thusrank$(\mathbf{M}'|_{\mathcal{R}'(J^*)}) < k$.

If (a) holds, $\textsc{GrowByOne}$ will reach Step 16 and output $\mathcal{B}^*$ which is by definition robustly certified full rank and robustly locally maximal. On the other hand, if $\textsc{GrowByOne}$ ever terminates at Step 13, we know that (b) holds, so it successfully outputs FAIL together with J^* satisfying $\text{rank}(\mathbf{M}'|_{\mathcal{R}(J^*)}) < k$. □

Lemma 7.7.5 (Robust version of Lemma 7.3.6)**.** *Fix a full rank realization of $\mathcal{D}$ and suppose* $\text{rank}(\mathbf{M}') = k'$. *Let $\mathcal{B} = \{T_1, T_2, \cdots, T_r\}$ be robustly certified full rank and robustly locally maximal. Let $J = \cup_i T_i$ and*

$$K = \left\{ i \middle| i \notin J \text{ and } \text{rank}(\mathbf{M}'|_{\mathcal{R}'(J \cup \{i\})}) = k' \right\}$$

If $K \neq \emptyset$ then the rows of $\mathbf{M}|_{\mathcal{B}}$ are a basis for the rows of $\mathbf{M}|_{2^{J \cup K}}$.

Proof. Our strategy is to apply Lemma 7.3.5 to $\mathbf{M}'$ and the set $J \cup K$ which will give the desired conclusion. We need to verify that the two conditions of Lemma 7.3.5 are met. The first condition of robust local maximality implies that there is no $T' \subseteq J$ for which $\textsc{InSpan}(\mathcal{D}, \mathcal{B}, T')$ returns False. Now we can invoke the first part of Lemma 7.7.1 which implies that $\mathbf{M}'_{T'}$ is in the span of $\mathbf{M}'|_{\mathcal{B}}$. This and the fact that $\mathcal{B}$ is robustly certified full rank imply that the rows of $\mathbf{M}'|_{\mathcal{B}}$ are indeed a basis for the rows of $\mathbf{M}'|_{2^J}$, which is the first condition we needed to check in Lemma 7.3.5.

For the second condition, the chain of reasoning is similar. Consider any $i \in K$ and any $T_{i'} \in \mathcal{B}$. Set $T' = T_{i'} \cup \{i\}$ and $J' = J \cup \{i\}$. Then $\text{rank}(\mathbf{M}'|_{\mathcal{R}'(J')}) = k$. Now the second condition of robust local maximality implies that $\textsc{InSpan}(\mathcal{D}, \mathcal{B}, T')$ returns True. We can once again invoke the first part of Lemma 7.7.1 to conclude that $\mathbf{M}_{T'}$ is in the span of $\mathbf{M}|_{\mathcal{B}}$, which is the second condition we needed to verify. This completes the proof. □

7.7.2 Robustly Tracking Down an Impostor

The bulk of adapting N-List to be sampling noise-robust rests on adapting Step 10 and proving a sampling noise-robust analogue of Lemma 7.3.8. Let $\mathcal{B} = T_1, ..., T_r$ be the output of Algorithm 25. Instead of solving (7.1), we can solve the regression problem

$$\tilde{\pi} \triangleq \underset{\pi \in [0,1]^r}{\operatorname{argmin}} \|\overline{\mathbf{M}}|_{\mathcal{B}} \cdot \pi^\top - \tilde{\mathbf{C}}|^{\emptyset}_{\mathcal{B}}\|_\infty. \tag{7.24}$$

We could then try solving an analogous regression problem for (7.2). The issue is that $\tilde{\pi}$ could have arbitrarily small entries (e.g. if $r < k$, in which case the assumption that $\mathcal{D}$ is $[\tau_{small}, \tau_{big}]$-avoiding tells us nothing). We handle this in the same way that we handle the possibility of π having small entries: sort the entries of $\tilde{\pi}$ as $\tilde{\pi}_1 \geq \tilde{\pi}_2 \geq \cdots \tilde{\pi}_r$, pick out the smallest $1 \leq r' < r$ for which

$$\tilde{\pi}^{r'}/\tilde{\pi}^{r'+1} > 2^{c_{22}k^2} \text{ and } \tilde{\pi}^{r'+1} < \upsilon$$

for sufficiently large $c_{22} > 0$ and υ to be chosen later — if no such r' exists, then set $r' = r$ — and show it is possible to at least learn the first r' columns of $\mathbf{m}$. Note that

$$\tilde{\pi}^{r'} \geq 2^{-c_{22}k^3}\upsilon. \tag{7.25}$$

For every $i \notin J$, we can solve the regression problem

$$\tilde{\mathbf{m}}_i \triangleq \underset{x \in [0,1]^{r'}}{\operatorname{argmin}} \|\overline{\mathbf{M}}|^{[r']}_{\mathcal{B}} \cdot \operatorname{diag}(\tilde{\pi}^{[r']}) \cdot x - \tilde{\mathbf{C}}|^{\{i\}}_{\mathcal{B}}\|_\infty. \tag{7.26}$$

We will show that for non-impostors i, these $\tilde{\mathbf{m}}_i$ can be rounded to the true values $\mathbf{m}'^{[r']}_i$.

Lemma 7.7.6 (Robust version of Lemma 7.3.8)**.** *There exist constants $c_{22}, c_{24}, c_{23} > 0$ for which the following is true. Let $\upsilon \leq \varepsilon \cdot k^{-c_{15}k-1}/18$, $\varepsilon_{samp} \leq \min(2^{-c_{24}k^3}\upsilon, k^{-c_{19}k^2}\tau_{big})$, $\tau_{small} \leq \min(2^{-c_{23}k^3}\upsilon, \rho\tau_{big})$. Suppose* GrowByOne *has access to ε_{samp}-close estimates of any moment of $\mathcal{D}$. Let $\mathcal{B} = \{T_1, ..., T_r\}$ be the output of* GrowByOne*, and let $K \subseteq [n]$ be the corresponding set of non-impostors, and suppose* $\operatorname{rank}(\mathbf{M}'|_{\mathcal{R}'(J)}) = k'$.

If $K \neq \emptyset$, then there exists a guess $\overline{\mathbf{m}}|_J \in \{0, 1/2, 1\}^{|J| \times r}$ for which the following holds:

Let $\tilde{\pi} \in \mathbb{R}^r$ and $\tilde{\mathbf{m}}_i \in \mathbb{R}^{r'}$ for $i \in K$ be solutions to (7.24) and (7.26). Assume without loss of generality that the entries of $\tilde{\pi}$ are sorted in nondecreasing order. For each $i \in K$, round $\tilde{\mathbf{m}}_i$ entrywise to the nearest $\overline{\mathbf{m}}_i \in \{0, 1/2, 1\}^{r'}$, and define $\overline{\pi} \in \Delta^{r'}$ to be the normalization of $\tilde{\pi}^{[r']}$. Define $\overline{\mathbf{m}} \in \{0, 1/2, 1\}^{|J \cup K| \times r'}$ to be the concatenation of $\overline{\mathbf{m}}_J^{[r']}$ and $\overline{\mathbf{m}}_i$ for all $i \in K$. Then the mixture $\overline{\mathcal{D}}$ of subcubes in $\{0,1\}^{|J \cup K|}$ with mixing weights $\overline{\pi}$ and marginals matrix $\overline{\mathbf{m}}$ satisfies

$$|\mathbb{E}_{\mathcal{D}}[x_S] - \mathbb{E}_{\overline{\mathcal{D}}}[x_S]| < \frac{1}{2}\varepsilon \cdot k^{-c_{15}k}$$

for all $S \subseteq J \cup K$ of size at most $2\log(2k)$.

Note that Lemma 7.7.6 is obviously true when $\mathbf{M}'$ has a single column: GrowByOne outputs the empty set, $\overline{\pi}$ has a single entry, 1, and $\overline{\mathbf{m}}$ is the column of marginals of the single product distribution corresponding to the single column of $\mathbf{M}'$.

In general, we will show Lemma 7.7.6 holds when $\overline{\mathbf{m}}|_J = \mathbf{m}'|_J$. Because $\mathrm{rank}(\mathbf{M}'|_{\mathcal{R}'(J \cup \{i\})}) = k'$ for all $i \in K$, Lemma 7.7.5 tells us that $\mathbf{M}'|_{\mathcal{B}}$ is a row basis for $\mathbf{M}'|_{2^{J \cup K}}$. In particular, $\mathrm{rank}(\mathbf{M}'|_{2^{J \cup K}}) = r$, so by Lemma 7.2.7, there exists r columns $\mathbf{m}^\dagger$ of $\mathbf{m}_{J \cup K}$ and $\pi^\dagger \in [0,1]^r$ for which $\mathbf{M}^\dagger \cdot \pi^\dagger = \mathbf{M}'|_{2^{J \cup K}} \cdot \pi'$.

Here is a simple perturbation bound.

Fact 7.7.7. *Pick any $S \subseteq \mathbb{R}^n$. Let $A \in \mathbb{R}^{m \times n}$, $x \in S$, and $b \in R^m$. If*

$$x^* \triangleq \underset{y \in S}{argmin} \|Ay - b\|_\infty,$$

then $\|x^* - x\|_\infty \leq 2\|Ax - b\|_\infty / \sigma^\infty_{\min}(A)$.

Proof. We have that

$$\|Ax^* - b\|_\infty \leq \|Ax - b\|_\infty,$$

so by the triangle inequality $\|A(x^* - x)\|_\infty \leq 2\|Ax - b\|_\infty$, from which the result follows. □

Corollary 7.7.8. $\|\tilde{\pi} - \pi^\dagger\|_\infty \leq 2\varepsilon_{samp} \cdot 2^{c_{14}k^2}$.

Proof. We know that

$$\|\mathbf{M}^\dagger|_{\mathcal{B}} \cdot (\pi^\dagger)^\top - \widetilde{\mathbb{E}}[x_S]\|_\infty \leq \|\mathbf{M}|_{\mathcal{B}} \cdot \pi^\top - \widetilde{\mathbb{E}}[x_S]\|_\infty + k\tau_{small} \leq \varepsilon_{\mathrm{samp}} + k\tau_{small}, \tag{7.27}$$

and $\sigma^{\infty}_{\min}(\mathbf{M}^{\dagger}|_{\mathcal{B}}) \geq 2^{-c_{14}k^2}$, so we can apply Fact 7.7.7 to get the desired bound on $\|\tilde{\pi} - \pi^{\dagger}\|_{\infty}$. □

To show Lemma 7.7.6, we will bound the objective value of (7.26) when x is chosen to be $\mathbf{m}^{[r']}_i$. Fact 7.7.7 will then let us conclude that the solution to (7.26) cannot be entrywise $1/4$-far from $\mathbf{m}^{[r']}_i$.

Lemma 7.7.9. *Let $i \notin J$ be a non-impostor. Then*

$$\|\mathbf{M}^{\dagger}|^{[r']}_{\mathcal{B}} \cdot \mathit{diag}(\tilde{\pi}^{[r']}) \cdot \mathbf{m}^{\dagger}|^{[r']}_i - \tilde{\mathbf{C}}|^{\{i\}}_{\mathcal{B}}\|_{\infty} \leq (k+1)(\varepsilon_{samp} + k\tau_{small}) + k \cdot \tilde{\pi}^{r'+1},$$

where we take $\tilde{\pi}^{r'+1}$ to be zero if $r' = r$.

Proof. We have that

$$\begin{aligned}\|\mathbf{M}^{\dagger}|_{\mathcal{B}} \cdot \mathrm{diag}(\tilde{\pi}) \cdot \mathbf{m}^{\dagger}_i - \tilde{\mathbf{C}}|^{\{i\}}_{\mathcal{B}}\|_{\infty} &\leq \|\mathbf{M}^{\dagger}|_{\mathcal{B}} \cdot \mathrm{diag}(\pi^{\dagger}) \cdot \mathbf{m}^{\dagger}_i - \tilde{\mathbf{C}}|^{\{i\}}_{\mathcal{B}}\|_{\infty} + k(\varepsilon_{\mathrm{samp}} + k\tau_{small}) \\ &\leq (k+1)(\varepsilon_{\mathrm{samp}} + k\tau_{small}),\end{aligned}$$

where in the second step we used the fact that

$$\mathbf{M}^{\dagger}|_{\mathcal{B}} \cdot \mathrm{diag}(\pi^{\dagger}) \cdot \mathbf{m}^{\dagger}_i = \mathbf{M}^{\dagger}|_{\{S \cup \{i\} : S \in \mathcal{B}\}} \cdot \mathrm{diag}(\pi^{\dagger}) = \mathbf{M}'|_{\{S \cup \{i\} : S \in \mathcal{B}\}} \cdot \mathrm{diag}(\pi').$$

For $r' < r$,

$$\|\mathbf{M}^{\dagger}|^{[r']}_{\mathcal{B}} \cdot \mathrm{diag}(\tilde{\pi}^{[r']}) \cdot \mathbf{m}^{\dagger}|^{[r']}_i - \mathbf{M}^{\dagger}|_{\mathcal{B}} \cdot \mathrm{diag}(\tilde{\pi}) \cdot \mathbf{m}^{\dagger}_i\|_{\infty} \leq k \cdot \tilde{\pi}^{r'+1},$$

so by the triangle inequality the claim follows. □

Corollary 7.7.10. *There exists some $c_{23} > 0$ for which the following holds. Let $i \notin J$ be a non-impostor. If $\varepsilon_{samp}, \tau_{small} < 2^{-c_{23}k^3}\upsilon$, then $\|\tilde{\mathbf{m}}_i - \mathbf{m}^{\dagger}|^{[r']}_i\|_{\infty} < 1/4$.*

Proof. By Fact 7.7.7,

$$|\tilde{\mathbf{m}}_i - \mathbf{m}^{\dagger}|^{[r']}_i| \leq \frac{2(k+1)(\varepsilon_{\mathrm{samp}} + k\tau_{small}) + 2\tilde{\pi}^{r'+1}}{\sigma^{\infty}_{\min}(\mathbf{M}^{\dagger}|^{[r']}_{\mathcal{B}}) \cdot \tilde{\pi}^{r'}}$$

$$< 2^{c_{14}k^2} \cdot \left(\frac{2(k+1)(\varepsilon_{\text{samp}} + k\tau_{small})}{2^{-c_{22}k^3}\upsilon} + 2^{-c_{22}k^2+1} \right).$$

where the second step follows from Lemma 7.3.10 and (7.25) We conclude that as long as $\varepsilon_{\text{samp}}, \tau_{small} \le 2^{-c_{23}k^3}\upsilon$ for sufficiently large $c_{23} > 0$, and c_{22} is large enough relative to c_{14}, we have that $|\tilde{\mathbf{m}}_i - \mathbf{m}^\dagger|_i^{[r']}| < 1/4$. □

In other words, Corollary 7.7.10 tells us that for every non-impostor i, if we round each entry of $\tilde{\mathbf{m}}_i$ to the nearest element of $\{0, 1/2, 1\}$, we will recover $\mathbf{m}^\dagger|_i^{[r']}$. We can now finish the proof of Lemma 7.7.6.

Proof of Lemma 7.7.6. We have already shown that $\overline{\mathbf{m}}$ defined in the statement of the lemma is equal to $\mathbf{m}^\dagger|_{J\cup K}^{[r']}$. By Corollary 7.7.8, $\|\tilde{\pi} - \pi^\dagger\|_\infty \le 2\varepsilon_{\text{samp}} \cdot 2^{c_{14}k^2}$, so

$$\begin{aligned}
\|\overline{\mathbf{M}} \cdot (\tilde{\pi}^{[r']})^\top - \tilde{\mathbf{C}}^{\emptyset}_{2^{J\cup K}}\|_\infty &\le \|\mathbf{M}^\dagger \cdot (\pi^\dagger)^\top - \tilde{\mathbf{C}}^{\emptyset}_{2^{J\cup K}}\| + k \cdot (2\varepsilon_{\text{samp}} \cdot 2^{c_{14}k^2} + \upsilon) \\
&= \|\mathbf{M}'_{2^{J\cup K}} \cdot (\pi')^\top - \tilde{\mathbf{C}}^{\emptyset}_{2^{J\cup K}}\|_\infty + k \cdot (2\varepsilon_{\text{samp}} \cdot 2^{c_{14}k^2} + \upsilon) \\
&= \varepsilon_{\text{samp}} + k \cdot \tau_{small} + k \cdot (2\varepsilon_{\text{samp}} \cdot 2^{c_{14}k^2} + \upsilon) \qquad (7.28)
\end{aligned}$$

It remains to show that we don't lose much if we take $\bar{\pi}$ to be the normalization of $\tilde{\pi}^{[r']}$. First note that $\sum_{i=1}^r \mathbf{pi}_i^\dagger = \mathbf{M}^\dagger_\emptyset \cdot (\pi^\dagger)^\top$. But $\emptyset \in \mathcal{B}$ and

$$\|\mathbf{M}^\dagger|_{\mathcal{B}} \cdot (\pi^\dagger)^\top - \widetilde{\mathbb{E}}[x_S]\|_\infty \le \|\mathbf{M}^\dagger|_{\mathcal{B}} \cdot (\tilde{\pi})^\top - \widetilde{\mathbb{E}}[x_S]\|_\infty \le \varepsilon_{\text{samp}} + k\tau_{small}$$

by (7.27) and the definition of $\tilde{\pi}$. So we get that

$$\sum_{i=1}^r \tilde{\pi}^i \ge \sum_{i=1}^r \pi_i^\dagger - \varepsilon_{\text{samp}} - k\tau_{small} = \sum_{i=1}^r \pi'^i - \varepsilon_{\text{samp}} - k\tau_{small} \ge 1 - \varepsilon_{\text{samp}} - 2k\tau_{small},$$

where the equality follows from Lemma 7.2.7. We conclude that

$$1/Z \triangleq \left(\sum_{i=1}^{r'} \tilde{\pi}^i \right)^{-1} \le (1 - \varepsilon_{\text{samp}} - 2k\tau_{small} - k\upsilon)^{-1} \le 1 + 2\varepsilon_{\text{samp}} + 4k\tau_{small} + 2k\upsilon$$

for $\varepsilon_{\text{samp}}, \tau_{small}, \upsilon$ small enough. So

$$\|\overline{\mathbf{M}} \cdot (\tilde{\pi}^{[r']})^\top - \frac{1}{Z}\overline{\mathbf{M}} \cdot (\tilde{\pi}^{[r']})^\top\|_\infty \le 2\varepsilon_{\text{samp}} + 4k\tau_{small} + 2k\upsilon.$$

This together with (7.28) give us the lemma provided the bounds on $\tau_{small}, \varepsilon_{\text{samp}}$ from the statement of Corollary 7.7.10 hold and provided $\tau_{small} \le \varepsilon \cdot k^{-c_{15}k-1}/30$, $\upsilon \le \varepsilon \cdot k^{-c_{15}k-1}/18$, and $\varepsilon_{\text{samp}} \le 2^{-c_{24}k^3} \cdot \varepsilon$ for sufficiently large $c_{24} > 0$. □

All of this gives us the subroutine NonDegenerateLearn specified by Algorithm 26. Given $\mathcal{D}, S, s$ for which $\Pr_{y\sim\mathcal{D}}[y_S = s]$ is sufficiently large and enough samples from $\mathcal{D}$, NonDegenerateLearn either successfully learns $(\mathcal{D}|x_S = s)$ if there are no impostors or outputs a list of subsets of size $2\log(2k)$, of which at least one must contain an impostor, and such that the size of the list does not depend on n.

We don't *a priori* know the interval $[\tau_{small}, \tau_{big}]$, so we instead consider $k+1$ windows

$$[\tau\rho, \tau], [\tau\rho^2, \tau\rho], ..., [\tau\rho^{k+1}, \tau\rho^k].$$

The mixing weights of our $[\tau_{small}, \tau_{big}]$-avoiding rank-$k$ realization of $\mathcal{D}$ avoid at least one of these windows, so NonDegenerateLearn will simply try each of them.

As we note in the fact below, the purpose of Step 4 is to ignore S, s for which $\Pr_{y\sim\mathcal{D}}[y_S = s]$ is too small for one to reliably simulate samples from $(\mathcal{D}|x_S = s)$.

Fact 7.7.11. *The following holds for any $\delta > 0$. Let $R = \tau_{trunc}^{-1} \cdot \ln(1/\delta)$, and let $\mathcal{D}$ be a mixture of k subcubes, $S \subseteq [n]$, and $s \in \{0,1\}^{|S|}$. If $Pr_{y\sim\mathcal{D}}[y_S = s] \le \tau_{trunc}\delta/\ln(1/\delta)$, then with probability at least $1-\delta$, NonDegenerateLearn terminates at Step 4. If $Pr_{y\sim\mathcal{D}}[y_S = s] \ge \tau_{trunc}$, then with probability at most δ, NonDegenerateLearn terminates at Step 4.*

Below, we summarize the guarantees for NonDegenerateLearn, which simply follow from Lemma 7.2.8 and the contrapositive of Lemma 7.7.6 applied to $(\mathcal{D}|x_S = s)$ instead of $\mathcal{D}$.

Lemma 7.7.12. *There exist $\varepsilon_{samp} = k^{-O(k^3)}\varepsilon$, $\tau = k^{-O(k^3)}\varepsilon$, and $\rho = k^{-O(k^2)}$ for which the following holds. Let $\mathcal{D}$ be a mixture of k subcubes, $S \subseteq [n]$, and $s \in \{0,1\}^{|S|}$. Suppose*

Algorithm 26: NonDegenerateLearn($(\mathcal{D}|x_S = s), k$) — for mixtures of subcubes

```
   Input: Mixture of subcubes $(\mathcal{D}|x_S = s)$, counter $k$
   Output: Either a mixture of subcubes with mixing weights $\bar{\pi}$ and marginals matrix
           $\overline{\mathbf{m}}$ realizing $\mathcal{D}'$ for which $d_{\mathrm{TV}}(\mathcal{D}', \mathcal{D}|x_S = s) \le \varepsilon$, or a set $\mathcal{U}$ of at most $3^{k^2}$
           subsets $W$, each of size at most $k + 2\log(2k)$ and for which
           $\mathrm{rank}(\mathbf{M}'|_{\mathcal{R}'(T)}) < k'$ for at least one $T$
1  $\varepsilon_{\mathrm{samp}} \leftarrow k^{-O(k^3)}\varepsilon$.
2  $\tau \leftarrow 2^{-O(k^3)}\varepsilon$.
3  $\rho \leftarrow k^{-O(k^2)}$.
4  Take $R$ samples $y$ from $\mathcal{D}$. If none of them are such that $y_S = s$, then return the
     distribution supported solely on $1^n$.
5  $\mathcal{U} \leftarrow \emptyset$.
6  for $[\tau_{small}, \tau_{big}] \in \{[\rho\tau, \tau], [\rho^2\tau, \rho\tau], ..., [\rho^{k+1}\tau, \rho^k\tau]\}$ do
7      Run GrowByOne to obtain $\mathcal{B} = \{T_1, ..., T_r\}$.
8      $J \leftarrow T_1 \cup \cdots \cup T_r$.
9      for $\overline{\mathbf{m}}|_J \subseteq \{0, 1/2, 1\}^{|J| \times r}$ do
10         Form estimates $\tilde{\mathbf{C}}|^{\emptyset}_{\mathcal{B}}$ and solve (7.24) for $\tilde{\pi} \in \Delta^r$. Sort the entries of $\tilde{\pi}$ so that
             $\tilde{\pi}^1 \ge \cdots \ge \tilde{\pi}^r$.
11         Pick the largest $1 \le r' < r$ for which $\tilde{\pi}^{r'}/\tilde{\pi}^{r'+1} \ge 2^{c_{22}k^2}$ and define $\tilde{\pi}^{[r']}$ to be
             the first $r'$ entries of $\tilde{\pi}$ and $\overline{\mathbf{M}}^{[r']}$ to be the first $r'$ columns of $\overline{\mathbf{M}}$. If no such
             $r'$ exists, pick $r' = r$.
12         For each $i \notin J$, form estimates $\tilde{\mathbf{C}}|^{\{i\}}_{\mathcal{B}}$, solve (7.26), and round entrywise to the
             nearest $\overline{\mathbf{m}}_i \in \{0, 1/2, 1\}^r$.
13         Normalize $\tilde{\pi}^{[r']}$ to $\bar{\pi} \in \Delta^{r'}$.
14         Define $\overline{\mathbf{m}} \in \{0, 1/2, 1\}^{n \times r'}$ to be the concatenation of $\overline{\mathbf{m}}^{[r']}$ and $\overline{\mathbf{m}}_i$ for all
             $i \notin J$.
15         for $T \subseteq [n]$ of size at most $2\log(2k)$ do
16             Compute estimate $\widetilde{\mathbb{E}}_{\mathcal{D}}[x_T]$ of $\mathbb{E}_{\mathcal{D}}[x_T]$ to within $\frac{1}{2}\varepsilon \cdot k^{-c_{15}k}$.
17             if $|\overline{\mathbf{M}}_T \cdot \bar{\pi}^\top - \widetilde{\mathbb{E}}_{\mathcal{D}}[x_T]| > \frac{1}{2}\varepsilon \cdot k^{-c_{15}k}$ for some $|T| \le 2\log(2k)$ then
18                 Add $J \cup T$ to $\mathcal{U}$ and return to line 6.
19             else
20                 return mixture of subcubes with mixing weights $\bar{\pi}$ and marginals
                     matrix $\overline{\mathbf{m}}$.
21 if $k = 1$ then
22     return FAIL.
23 else
24     return $\mathcal{U}$.
```

$(\mathcal{D}|x_S = s)$ has a rank-r realization. An ε_{samp}-sample-rich invocation of NonDegenerateLearn *on $\mathcal{D}|x_S = s$ that does not terminate at Step 4 outputs either a mixture of subcubes $\overline{\mathcal{D}}$ for which $d_{TV}(\mathcal{D}, \overline{\mathcal{D}}) \le \varepsilon$, or a collection $\mathcal{U}$ of at most 3^{k^2} subsets $W \subseteq [n]\backslash S$ containing some W for which $(\mathcal{D}|x_{S\cup W} = s \circ t)$ has a rank-$(r-1)$ realization for every $t \in \{0,1\}^{|W|}$.*

7.7.3 Correctness of N-List

We complete the proof of Theorem 7.1.1 by verifying that the conditions of Lemma 7.7.14 are satisfied by the output of a $(\varepsilon_{\text{samp}}(\cdot), \delta_{edge}, \tau_{trunc})$-sample-rich run of N-List on $\mathcal{D}$ and counter k.

Theorem 7.7.13. *There exists $\varepsilon_{samp} = \varepsilon \cdot k^{-O(k^3)}$ and absolute constant $c_{25} > 0$ such that the following holds for any $\delta > 0$. Let $\mathcal{D}$ be a mixture of k subcubes. If a run of* N-List *is $(\varepsilon_{samp}, \delta_{edge}, \tau_{trunc})$-sample-rich on input $\mathcal{D}$ and counter k, then with probability $1-3^{c_{25}k^3} \cdot k^k \cdot \delta$, the output is a sampling tree such that all leaves $v_{T,t}$ for which $Pr_{y\sim\mathcal{D}}[x_T = t] \ge \tau_{trunc}$ correspond to distributions ε-close to $(\mathcal{D}|x_T = t)$.*

Proof. By the proof of Lemma 7.6.7 with $S = k + 2\log(2k)$ and $U = k \cdot 3^{k^2}$, the total number of invocations of NonDegenerateLearn is at most $2^{Sk}U^k \le 3^{c_{25}k^3} \cdot k^k$. By Fact 7.7.11 and a union bound over these invocations, with probability at least $1 - 3^{c_{25}k^3}k^k\delta$ every invocation of NonDegenerateLearn on $(\mathcal{D}|x_S = s)$ for which $\Pr_{y\sim\mathcal{D}}[y_S = s] < \tau_{trunc}\delta/\ln(1/\delta)$ (resp. $\Pr_{y\sim\mathcal{D}}[y_S = s] \ge \tau_{trunc}$) does (resp. does not) terminate on Step 4. Henceforth suppose this is the case.

We call a sampling tree *good* if its leaves $v_{T,t}$ all satisfy that either $\Pr_{y\sim\mathcal{D}}[y_T = t] < \tau_{trunc}$ or they are ε-close to $(\mathcal{D}|x_T = t)$.

It suffices to show by induction on r that if $(\mathcal{D}|x_S = s)$ has a rank-r realization then N-List$(\mathcal{D}, S, s, r)$ returns a good sampling tree. This is certainly true for $r = 1$, in which case N-List returns the sampling tree given by a single node with distribution that's actually equal to $(\mathcal{D}|x_T = t)$.

Consider $r > 1$. There are three possibilities:

1. If $\Pr_{y\sim\mathcal{D}}[y_S = s] < \tau_{trunc}\delta/\ln(1/\delta)$, then NonDegenerateLearn terminates on

Step 4 instead of potentially returning FAIL, and the inductive step is vaculously complete.

2. If $\Pr_{y\sim\mathcal{D}}[y_S = s] \geq \tau_{trunc}$, then NonDegenerateLearn does not terminate at Step 4.

3. If $\tau_{trunc}\delta/\ln(1/\delta) \leq \Pr_{y\sim\mathcal{D}}[y_S = s] < \tau_{trunc}$, then either NonDegenerateLearn terminates on Step 4 and the inductive step is vacuously complete, or NonDegenerateLearn does not terminate at Step 4.

In cases 2) and 3) above where NonDegenerateLearn does not terminate, the invocation of NonDegenerateLearn is $\varepsilon_{\mathrm{samp}}$-sample-rich because N-List is $(\varepsilon_{\mathrm{samp}}, \delta_{edge}, \tau_{trunc})$-sample-rich by assumption, so by Lemma 7.7.12, either NonDegenerateLearn outputs a mixture with mixing weights π and marginals matrix $\mathbf{m}$ which is ε-close to $(\mathcal{D}|x_S = s)$ and we're done, or it outputs some collection $\mathcal{U}$.

We claim it's enough to show there is *some* guess $W \in \mathcal{U}$ for which N-List$(\mathcal{D}, S\cup W, s\circ t, r-1)$ does not return FAIL. Suppose we instead get some sampling tree $\mathcal{T}$. For any leaf node $v_{T,t}$ of $\mathcal{T}$ for which $\Pr_{y\sim\mathcal{D}}[y_T = t] \geq \tau_{trunc}$, the corresponding distribution is ε-close to $(\mathcal{D}|x_T = t)$ by Lemma 7.7.12. So any such $\mathcal{T}$ would be good, and Select would simply pick one of these.

Finally, to show the existence of such a guess W, we appeal once more to Lemma 7.7.12, which implies that $\mathcal{U}$ must contain some W for which $(\mathcal{D}|x_{S\cup W} = s\circ t)$ has a rank-$(r-1)$ realization for every $t \in \{0,1\}^{|W|}$, and we're done by induction. □

To complete the proof of Theorem 7.1.1, we use the following simple fact about sampling trees, which says that in a sampling tree $\mathcal{T}$, if all internal transition probabilities out of nodes $v_{S,s}$ for which $\Pr_{y\sim\mathcal{D}}[y_S = s]$ is sufficiently large are accurate, and if all distributions associated to leaves $v_{S,s}$ for which $\Pr_{y\sim\mathcal{D}}[y_S = s]$ is sufficiently large are accurate, then the distribution associated to $\mathcal{T}$ is close to $\mathcal{D}$.

Lemma 7.7.14. *Let $\mathcal{T}$ be a sampling tree with depth k, maximal fan-out d, and $M \triangleq d^{\Theta(k)}$ nodes corresponding to a distribution $\mathcal{D}^*$. Denote by V_{trunc} the set of S, s indexing nodes $v_{S,s}$ of $\mathcal{T}$ for which $Pr_{y\sim\mathcal{D}}[y_S = s] \leq \tau_{trunc} \triangleq \varepsilon/M$. Suppose $|w_{S,W,s,t} - Pr_{y\sim\mathcal{D}}[y_W = t|y_S = s]| \leq$*

$\eta \triangleq \frac{\varepsilon}{2^k M}$ for all $S, s \notin V_{trunc}$, and suppose that $d_{TV}(\mathcal{D}_{T,t}, \mathcal{D}|x_T = t) \le \varepsilon$ for any leaf $v_{S,s}$ with $S, s \in V_{trunc}$. Then $d_{TV}(\mathcal{D}_{\emptyset,\emptyset}, \mathcal{D}) \le O(\varepsilon)$.

Proof. Denote by $\mathcal{U}^{S,s}_{trunc}$ the set of all $x \in \{0,1\}^{n-|S|}$ for which there exist $W \subseteq [n], t \in \{0,1\}^{|W|}$ such that $x_W = t$, $v_{S\oplus W, s\oplus t}$ is a node of $\mathcal{T}$ (not necessarily the direct descendent of $v_{S,s}$) and $\Pr_{\mathcal{D}}[x_{S\cup W} = s \oplus t] \le \tau_{trunc}$. In other words, $\mathcal{U}^{S,s}_{trunc}$ corresponds to strings t over the coordinates W such that further conditioning on $x_W = t$ leads to a vertex of $\mathcal{T}$ which occurs rarely enough that it doesn't matter how well we learn the posterior distribution $\mathcal{D}|x_{S\cup W} = s \oplus t$.

For any node $v_{S,s}$ associated to distribution $\mathcal{D}_{S,s}$, define

$$\mathrm{err}_{trunc}(v_{S,s}) \triangleq \sum_{y \notin \mathcal{U}^{S,s}_{trunc}} \left| \Pr_{\mathcal{D}_{S,s}}[y] - \Pr_{\mathcal{D}|x_S = s}[y] \right|.$$

In particular, if $\mathcal{U}^{S,s}_{trunc}$ were empty, $\mathrm{err}_{trunc}(v_{S,s})$ would just be $2d_{\mathrm{TV}}(\mathcal{D}_{S,s}, \mathcal{D}|x_S = s)$.

First observe that it is enough to show that

$$\mathrm{err}_{trunc}(v_{\emptyset,\emptyset}) \le O(\varepsilon). \tag{7.29}$$

To show this, first note that $\sum_{x\in\mathcal{U}^{\emptyset,\emptyset}_{trunc}} \Pr_{\mathcal{D}}[x] \le M\tau_{trunc} = \varepsilon$. Furthermore, for any $y \in \mathcal{U}^{\emptyset,\emptyset}_{trunc}$, if $v_{W,t}$ is the closest node to the root for which $y_W = t$ and $\Pr_{\mathcal{D}}[x_W = t] \le \tau_{trunc}$, then because the weights on the edges of $\mathcal{T}$ are additively η-close to the true values and $v_{W,t}$ is distance at most k from the root,

$$|\Pr_{\mathcal{D}^*}[x_W = t] - \Pr_{\mathcal{D}}[x_W = t]| \le 2^k \eta.$$

So

$$\sum_{x\in\mathcal{U}^{\emptyset,\emptyset}_{trunc}} \Pr_{\mathcal{D}^*}[x] \le \sum_{x\in\mathcal{U}^{\emptyset,\emptyset}_{trunc}} \Pr_{\mathcal{D}}[x] + 2^k \cdot M \cdot \eta \le M\tau_{trunc} + 2^k \cdot M \cdot \eta.$$

By triangle inequality we would then be able to conclude that

$$2d_{\mathrm{TV}}(\mathcal{D}^*, \mathcal{D}) \le O(\varepsilon) + \sum_{x\in\mathcal{U}^{\emptyset,\emptyset}_{trunc}} (\Pr_{\mathcal{D}}[x] + \Pr_{\mathcal{D}^*}[x]) \le O(\varepsilon) + 2M\tau_{trunc} + 2^k \cdot M \cdot \eta,$$

and by picking $\tau_{trunc} \leq \varepsilon/M$ and $\eta \leq \varepsilon/2^k M$, this would tell us that $d_{\mathrm{TV}}(\mathcal{D}^*, \mathcal{D}) \leq O(\varepsilon)$.

To show (7.29), we show by induction that $\mathrm{err}_{trunc}(v_{S,s}) \leq O(\varepsilon)$ $\forall$ vertices $v_{S,s}$ of $\mathcal{T}$. This is vacuously true if $\Pr_{\mathcal{D}}[x_S = s] \leq \tau_{trunc}$, so suppose otherwise.

If $v_{S,s}$ is a leaf, then we're done by assumption. Otherwise, by induction we know

$$\mathrm{err}_{trunc}(v_{S\cup W, s\oplus t}) \leq \varepsilon' \tag{7.30}$$

for some $\varepsilon' > 0$ for all immediate descendants of $v_{S,s}$. Decompose $[n]\backslash S$ as $W \cup W'$. The true probability of drawing some string $t \oplus u \in \{0,1\}^{n-|S|}$ from $(\mathcal{D}|x_S = s)$ can be written as

$$\mathbb{P}_{\mathcal{D}|x_S=s}[t \oplus u] = \mathbb{P}_{y\sim\mathcal{D}|x_S=s}[y_W = t] \cdot \mathbb{P}_{\mathcal{D}|x_{S\cup W}=s\oplus t}[u] \triangleq w_t \cdot p_u.$$

Let $\Pr_{\mathcal{D}_{S\cup W,s\oplus t}}[u] = p_u + \delta_u$ for some $\delta_u > 0$ for all $u \notin \mathcal{U}^{S\cup W,s\oplus t}_{trunc}$. By inductive assumption (7.30),

$$\sum_{u\notin\mathcal{U}^{S\cup W,s\oplus t}_{trunc}} |\delta_u| = \mathrm{err}_{trunc}(v_{S\cup W,s\oplus t}) \leq \varepsilon'$$

for all $t \in \{0,1\}^{|W|}$. Then

$$\begin{aligned}
\mathrm{err}_{trunc}(v_{S\cup W,s\oplus t}) &\leq \sum_{t\in\{0,1\}^{|W|}} \sum_{u\notin\mathcal{U}^{S\cup W,s\oplus t}_{trunc}} \left|\mathbb{P}_{\mathcal{D}_{S,s}}[t\oplus u] - \mathbb{P}_{\mathcal{D}|x_S=s}[t\oplus u]\right| \\
&\leq \sum_{t\in\{0,1\}^{|W|}} \sum_{u\notin\mathcal{U}^{S\cup W,s\oplus t}_{trunc}} w_t \cdot |\delta_u| + \sum_{t\in\{0,1\}^{|W|}} (\eta + \eta\varepsilon') \\
&\leq (2^{|W|}\eta + 1)\varepsilon' + 2^{|W|}\eta \leq (d\eta + 1)\varepsilon' + d\eta.
\end{aligned}$$

Unrolling the resulting recurrence tells us that

$$\mathrm{err}_{trunc}(v_{\emptyset,\emptyset}) \leq (d\eta+1)^k\varepsilon + (d\eta+1)^{k+1} - 1,$$

so as long as $\eta \leq \frac{\varepsilon}{dk^2}$, we have $\mathrm{err}_{trunc}(v_{\emptyset,\emptyset}) \leq 3\varepsilon$. Because we are assuming $\eta \leq \varepsilon/2^k M$ and $M = d^{\Theta(k)}$, we certainly have that $\eta \leq \frac{epsilon}{dk^2}$, completing the proof of (7.29). □

Proof of Theorem 7.1.1. Let $\alpha = \delta/(3^{c_{25}k^3}k^k)$. Apply Lemma 7.6.7 with $\tau_{trunc} = \varepsilon/2^{k^2}$, $Z = n^{O(\log k)}$, $M = 1$, $U = 3^{k^2}$, $S = k + 2\log(2k)$, $T(r) = n^{O(\log k)} + \tau_{trunc}^{-1}\ln(1/\alpha)$, $\varepsilon_{\mathrm{samp}} = k^{-O(k^3)}\varepsilon$, and $\varepsilon_{select} = O(\varepsilon)$ to get that achieving a $(\varepsilon_{\text{samp}}, \delta_{edge}, \tau_{trunc}\alpha/\ln(1/\alpha))$-sample-rich run of N-List on $\mathcal{D}$ with counter k with probability $1-\delta$ requires $O(k^{O(k^3)}\varepsilon^{-3}\ln(1/\delta)n^{O(\log k)})$ time and $O(k^{O(k^3)}\varepsilon^{-2}\log(n)\log(1/\delta))$ samples. By taking $\delta_{edge} = \varepsilon/2^{k+k^2}$, we conclude by Theorem 7.7.13 and Lemma 7.7.14 with $d = 2^k$ that the output of N-List is 2ε-close to $\mathcal{D}$. □

7.8 Appendix: Learning Mixtures of Product Distributions Over $\{0,1\}^n$

In Section 7.5 we described our algorithm for learning general mixtures of product distributions over $\{0,1\}^n$ under the assumption that we had exact access to the accessible entries of $\mathbf{C}$, when in reality we only have access to them up to some sampling noise $\varepsilon_{\text{samp}} > 0$. In this section, we show how to remove the assumption of zero sampling noise and thereby give a complete description of the algorithm for learning mixtures of product distributions.

It will be convenient to define the following:

Definition 7.8.1. *Two distributions $\mathcal{D}$ and $\mathcal{D}'$ over $\{0,1\}^n$ are (ε, d)*-moment-close *if $|\mathbb{E}_{\mathcal{D}'}[x_S] - \mathbb{E}_{\mathcal{D}}[x_S]| \le \varepsilon$ for all $S \subseteq [n]$ such that $|S| \le d$. We say mixing weights π and marginals matrix $\mathbf{m}$ constitute an (ε, d)-moment-close realization of $\mathcal{D}$ if the distribution they realize is (ε, d)-moment close to $\mathcal{D}$.*

Let $\mathcal{D}$ be a mixture of k product distributions over $\{0,1\}^n$. As in Section 7.7, we will use $\widetilde{\mathbb{E}}[x_S]$ to denote any $\varepsilon_{\text{samp}}$-close estimate of $\mathbb{E}x_S]$ and $\tilde{\mathbf{C}}$ to denote a matrix consisting of $\varepsilon_{\text{samp}}$-close estimates of the accessible entries of $\mathbf{C}$.

7.8.1 NonDegenerateLearn and Its Guarantees

Let $s(k) = 2k + 1 + (1 + 2 + \cdots + (k-1))$. We will define $\mathcal{R}_k^\dagger(J)$ to be all subsets of $[n]\backslash J$ of size at most $s(k-1)$. The main properties we need about s are that $s(0) = 1$ and that

$$s(k) = k + 1 + s(k-1)m \tag{7.31}$$

Note that $s(k) = \Theta(k^2)$ even though we showed in Lemma 7.5.6 that degree-$O(k)$ moments are enough to robustly identify any mixture of k product distributions. Roughly, the reason for doing so is that whereas we can always perfectly collapse matrices that are not full rank as in Lemma 7.2.7 for learning mixtures of subcubes, collapsing matrices that are merely ill-conditioned as in Lemma 7.5.10 for learning mixtures of product distributions necessarily incurs some loss every time. We must ensure after collapsing ill-conditioned matrices k' times from recursively conditioning $\mathcal{D}$ k' times for any $k' \le k$ that these losses do not compound so that the resulting moment matrix of the conditional distribution is still close to a mixture of at most $k - k'$ product distributions. In particular, (7.31) will prove crucial in the proof of Lemma 7.8.4 in the next subsection.

We now recall the algorithm outlined in Section 7.5: 1) exhaustively search for a barycentric spanner $J \subseteq [n]$ for the rows of $\mathbf{m}$ which may be any size $r \le k$, 2) express the remaining rows of $\mathbf{m}$ as linear combinations of rows J by solving

$$\tilde{\alpha}_i \triangleq \operatorname*{argmin}_{\alpha \in [-1,1]^r} \|\tilde{\mathbf{C}}|^{\{i_1\},\dots,\{i_r\}}_{\mathcal{R}^\dagger_r(J\cup\{i\})}\alpha - \tilde{\mathbf{C}}|^{\{i\}}_{\mathcal{R}^\dagger_r(J\cup\{i\})}\|_\infty. \tag{7.32}$$

for each $i \notin J$, and 3) grid the mixing weights and entries of rows J. The details of this are given in Algorithm 27 below.

The main technical lemma of this section, Lemma 7.8.2 below, tells us that as long as the gridding in Step 7 of Algorithm 27 below is done with $\mathrm{poly}(\varepsilon, 1/k, 1/n)$ granularity and $\mathcal{D}$ obeys a suitable non-degeneracy condition, the above algorithm will produce a list of mixtures containing a mixture which is close in parameter distance to $\mathcal{D}$. In fact it says more: for *any* k, if $\mathcal{D}$ has *any* moment-close rank-k realization by mixing weights π and marginals matrix $\mathbf{m}$, the output list of NonDegenerateLearn with $\mathcal{D}$ and counter k as inputs will contain a mixture with mixing weights close to π and marginals matrix close to $\mathbf{m}$.

By Lemma 7.5.6, moment-closeness implies closeness in total variation distance, and by Lemmas 7.5.1, 7.5.2, and 7.5.3, parameter closeness also implies closeness in total variation distance. The upshot of all of this is that for *any* k for which $\mathcal{D}$ has a moment-close rank-k realization, applying hypothesis selection to the output list of NonDegenerateLearn with $\mathcal{D}$ and counter k as inputs will yield a distribution close to $\mathcal{D}$. This will allow us to leverage our insights from Section 7.5.5 about collapsing ill-conditioned moment matrices to give a full proof of correctness of N-List in later subsections.

Algorithm 27: NonDegenerateLearn($\mathcal{D}, k$) — for mixtures of general product distributions

```
Input: Mixture of product distributions $\mathcal{D}$, counter $k$
Output: List of mixtures of product distributions containing one that is $\varepsilon$-close to
        $\mathcal{D}$, and/or the set $\mathcal{U}$ of all subsets $W$ of size at most $k+1$
$\sigma_{cond}(k) \leftarrow \left(\frac{\varepsilon^2}{c_{27}nk^2 2^k}\right)^k$.
$\varepsilon_{\mathrm{samp}}(k) \leftarrow \sigma_{cond}(k) \cdot \frac{c_{26}\varepsilon^2}{k^3 n}$.
$\alpha \leftarrow 2\varepsilon/3k^2$.
$\delta \leftarrow \frac{\varepsilon}{8k^2 n}$.
$\mathcal{M} \leftarrow \emptyset$.
for all guesses of coordinates $J = i_1, ..., i_r \subseteq [n]$ where $r \le k$ and all guesses of
  mixture weights $\overline{\pi}^1, ..., \overline{\pi}^{k-1} \in \{0, \alpha, 2\alpha, ..., \lfloor 1/\alpha \rfloor \alpha\}^k$ do
    $\overline{\pi}_{k-1} \leftarrow 1 - \overline{\pi}^1 - \cdots - \overline{\pi}^{k-1}$.
    for $i \notin J$ do
        Compute an entrywise $\varepsilon_{\mathrm{samp}}(k)$-close estimate $\tilde{E}$ for the entries of $\mathbf{C}|^{\{i_1\},...,\{i_r\}}_{\mathcal{R}^\dagger_r(J\cup\{i\})}$.
        Compute an entrywise $\varepsilon_{\mathrm{samp}}(k)$-close estimate $\tilde{b}$ for the entries of $\mathbf{C}|^{\{i\}}_{\mathcal{R}^\dagger_r(J\cup\{i\})}$.
        Solve for $\tilde{\alpha}_i$ in (7.32).
    for $\overline{\mathbf{m}}|_J \subseteq \{0, \delta, 2\delta, ..., 1\}^r$ do
        For every $i \notin J$ define $\overline{\mathbf{m}}_i = \overline{\mathbf{m}}|_J \cdot \alpha_i$.
        Append the mixture with mixing weights $\overline{\pi}$ and marginals matrix $\overline{\mathbf{m}}$ to $\mathcal{M}$.
return $\mathcal{M}$. If $k > 1$, also output the set of all $W \subseteq [n]$ of size at most $k+1$.
```

Lemma 7.8.2. *For some small absolute constant $0 < c_{26} < 1$, the following holds for any $\sigma_{cond}(k) > 0$. Suppose $\varepsilon_{samp}(k) = \sigma_{cond}(k) \cdot \frac{c_{26}\varepsilon^2}{k^3 n}$. Let mixing weights π and marginals matrix $\mathbf{m}$ constitute any $(\varepsilon_{samp}(k), s(k))$-moment-close realization of $\mathcal{D}$. If $J = \{i_1, ..., i_r\} \subseteq [n]$ is a barycentric spanner for the rows of $\mathbf{m}$ for some $r \le k$ and $\sigma^\infty_{\min}(\mathbf{M}|_{\mathcal{R}^\dagger_k(J\cup\{i\})}) \ge \sigma_{cond}(k)$ for all $i \notin J$, then for any $\overline{\mathbf{m}}|_J \in [0,1]^{r\times k}$ for which $\overline{\mathbf{m}}|_J$ and $\mathbf{m}$ are entrywise δ-close for $\delta = \frac{\varepsilon}{8k^2 n}$, we have that $|\tilde{\alpha}_i^\top \cdot \overline{\mathbf{m}}|^j_J - \mathbf{m}^j_i| \le \varepsilon/4kn$ for all $i \notin J$ and j for which $\pi^j \ge \varepsilon/6k$, where $\tilde{\alpha}_i \in [-1,1]^r$ is defined in (7.32).*

Proof. Let $\mathbf{C}$ be the expectations matrix of the distribution realized by mixing weights π and marginals matrix $\mathbf{m}$, and let $\tilde{\mathbf{C}}$ be the empirical expectations matrix of $\mathcal{D}$ which approximates the expectations matrix of $\mathcal{D}$ to entrywise error $\varepsilon_{\text{samp}}(k)$. Denote $\mathbf{C}|^{\{i_1\},\dots,\{i_r\}}_{\mathcal{R}^\dagger_k(J\cup\{i\})}$ and $\tilde{\mathbf{C}}|^{\{i_1\},\dots,\{i_r\}}_{\mathcal{R}^\dagger_k(J\cup\{i\})}$ by $E, \tilde{E} \in [0,1]^{n^{O(k)}\times r}$ respectively. Because the entries of E and $\tilde{E}$ correspond to moments of degree at most $s(k-1)+1 \le s(k)$, by triangle inequality and moment-closeness of $\mathcal{D}$ to the mixture given by π and $\mathbf{m}$, we have that $\tilde{E} = E + \Delta_E$ for $\|\Delta_E\|_{\max} \le 2\varepsilon_{\text{samp}}(k)$. Likewise, denote $\tilde{\mathbf{C}}|^{\{i\}}_{\mathcal{R}^\dagger_k(J\cup\{i\})}$ and $\tilde{\mathbf{C}}|^{\{i\}}_{\mathcal{R}^\dagger_k(J\cup\{i\})}$ by $b, \tilde{b} \in [0,1]^{n^{O(k)}}$ respectively so that $\tilde{b} = b + \Delta_b$ for $\|\Delta_b\|_\infty \le 2\varepsilon_{\text{samp}}(k)$. Also define $D = \text{diag}(\pi^1, ..., \pi^k)$ and $P = \mathbf{M}|_{\mathcal{R}^\dagger_k(J\cup\{i\})}$. As in the proof of Lemma 7.3.14, we have the decompositions

$$E = PD(\mathbf{m}|_J)^\top, \; b = PD(\mathbf{m}_i)^\top \tag{7.33}$$

Because J is a barycentric spanner for the rows of $\mathbf{m}$, there exists $\alpha_i \in [-1,1]^r$ for which $(\mathbf{m}|_J)^\top\alpha_i - (\mathbf{m}_i)^\top = \mathbf{0}$. We conclude that for $\tilde{\alpha}_i$ defined by (7.32),

$$\|\tilde{E}\tilde{\alpha}_i - \tilde{b}\|_\infty \le \|\tilde{E}\alpha_i - \tilde{b}\|_\infty \le \|E\alpha_i - b\| + \|\Delta_E\alpha_i\| + \|\Delta_b\| \le 2(r+1)\varepsilon_{\text{samp}}(k). \tag{7.34}$$

By (7.33) we can express

$$\tilde{E}\tilde{\alpha}_i - \tilde{b} = PD(\tilde{\alpha}_i^\top\mathbf{m}|_J - \mathbf{m}_i)^\top + \Delta_E\tilde{\alpha}_i - \Delta_b\tilde{\alpha}_i.$$

Because $\tilde{\alpha}_i \in [-1,1]^r$, $\|\Delta_E\tilde{\alpha}_i - \Delta_b\|_\infty \le 2(r+1)\varepsilon_{\text{samp}}(k)$ as in (7.34). It follows that

$$\|PD(\tilde{\alpha}_i^\top\mathbf{m}|_J - \mathbf{m}_i)^\top\|_\infty \le 4(r+1)\varepsilon_{\text{samp}}(k).$$

Because $\sigma^\infty_{\min}(P) \ge \sigma_{cond}(k)$, we get that

$$|\tilde{\alpha}_i^\top\mathbf{m}|^j_J - \mathbf{m}^j_i)| \le \frac{4(r+1)\varepsilon_{\text{samp}}(k)}{\sigma_{cond}(k)\cdot\pi^j}$$

for all $j \in [k]$. Lastly, because $\tilde{\alpha}_i \in [-1,1]^r$, it follows that $\|\tilde{\alpha}_i^\top(\mathbf{m}|_J - \overline{\mathbf{m}}|_J)\|_\infty \le r\delta$ for any $\overline{\mathbf{m}}_J$ which is entrywise δ-close to $\mathbf{m}|_J$, and we conclude that

$$\|\tilde{\alpha}_i^\top\overline{\mathbf{m}}|_J - \mathbf{m}_i\|_\infty \le \frac{4(r+1)\varepsilon_{\text{samp}}(k)}{\sigma_{cond}(k)\cdot\pi^j} + r\delta.$$

Obviously $r \leq k$, so by picking $\varepsilon_{\mathrm{samp}}(k) = \sigma_{cond}(k) \cdot \frac{\varepsilon}{192(k+1)^2 kn}$, and $\delta = \frac{\varepsilon}{8k^2 n}$, we obtain the desired bound of $|\tilde{\alpha}_i^\top \overline{\mathbf{m}}|_J^j - \mathbf{m}_i^j| \leq \varepsilon/4kn$ for all j such that $\pi^j \geq \varepsilon/6k$. □

We conclude this subsection by deducing that when $\mathcal{D}$ obeys the non-degeneracy condition in the statement of Lemma 7.8.2, the mixture that is output by NonDegenerateLearn is not just close in parameter distance to a moment-close realization of $\mathcal{D}$, but close in total variation distance to $\mathcal{D}$ itself. This is the only place in our analysis where we need the robust low-degree identifiability machinery developed in Section 7.5.4, but as it will form the base case for our inductive proof of the correctness of N-List in later subsections, this corollary is essential.

Corollary 7.8.3 (Corollary of Lemma 7.8.2)**.** *The following holds for any $\sigma_{cond}(k) > 0$. Suppose that $\varepsilon_{samp}(k) = \sigma_{cond}(k) \cdot \frac{c_{26}\varepsilon^2}{k^3 n}$ and that $\varepsilon_{samp}(k) \leq \eta(n, 2k, \varepsilon)$. If there exists an $(\varepsilon_{samp}(k), s(k))$-moment-close rank-$k$ realization of $\mathcal{D}$ by mixing weights π and marginals matrix $\mathbf{m}$ and $\sigma^\infty_{\min}(\mathbf{M}|_{\mathcal{R}^\dagger_k(W)} \geq \sigma_{cond}(k)$ for all $W \subseteq [n]$ of size at most $k+1$, then an ε_{samp}-sample-rich run of* NonDegenerateLearn *on input $\mathcal{D}$ outputs a list of mixtures among which is a mixture $\overline{\mathcal{D}}$ for which $d_{TV}(\mathcal{D}, \overline{\mathcal{D}}) \leq 2\varepsilon$.*

Proof. Lemma 7.8.3 certifies that under these assumptions, there is at least one mixture close to $\mathcal{D}$ among the candidate mixtures compiled by NonDegenerateLearn. Specifically, consider the following marginals matrix $\overline{\mathbf{m}}$. Let $J \subseteq [n]$ be a barycentric spanner for the rows of $\mathbf{m}$. Round the entries of $\mathbf{m}|_J$ to the nearest multiples of $\frac{\varepsilon}{8k^2 n}$ to get $\overline{\mathbf{m}}_J \in [0,1]^{r \times k}$ satisfying the assumptions of Lemma 7.8.2. By Lemma 7.8.3 then implies that if we define $\overline{\mathbf{m}}_i = \overline{\mathbf{m}}|_J \cdot \alpha_i$ for $i \notin J$ as in NonDegenerateLearn, then $|\overline{\mathbf{m}}_i^j - \mathbf{m}_i^j| \leq \varepsilon/4kn$ for all $i \notin J$ and j for which $\pi^j \geq \varepsilon/6k$. By restricting to those entries of π and normalizing to obtain some $\tilde{\pi}$, and restricting $\overline{\mathbf{m}}$ to the corresponding columns, we get by Lemmas 7.5.1 and 7.5.3 that the mixture $(\tilde{\pi}, \overline{\mathbf{m}})$ is $2\varepsilon/3$-close to the mixture $(\pi, \mathbf{m})$. We can round every entry of $\tilde{\pi}$ except the last one to the nearest multiple of $2\varepsilon/3k^2$ and replace the last entry by 1 minus these rounded entries. The resulting vector $\overline{\pi}$ is entrywise $2\varepsilon/3k$-close to $\tilde{\pi}$, so by Lemma 7.5.2, $(\overline{\pi}, \overline{\mathbf{m}})$ is $2\varepsilon/3 + \varepsilon/3 = \varepsilon$-close to $(\pi, \mathbf{m})$, which is ε-close to $\mathcal{D}$ by Lemma 7.5.6. □

7.8.2 Making Progress When $\mathbf{M}|_{\mathcal{R}_k^\dagger(J\cup\{i\})}$ is Ill-Conditioned

NonDegenerateLearn will successfully output a mixture close to $\mathcal{D}$ provided $\varepsilon_{\text{samp}}$ is sufficiently small relative to $\sigma^\infty_{\min}(\mathbf{M}|_{\mathcal{R}_k^\dagger(J\cup\{i\})})$, i.e. provided $\mathbf{M}|_{\mathcal{R}_k^\dagger(J\cup\{i\})}$ is sufficiently well-conditioned. In this subsection, we argue that when this is not the case and NonDegenerateLearn fails, we can condition on some set of coordinates and recursively learn the resulting conditional distributions.

Specifically, we show that Lemma 7.5.10 and the contrapositive of Lemma 7.8.2 imply that if NonDegenerateLearn fails to output a mixture of r product distributions close to $\mathcal{D}$, one of the subsets W that NonDegenerateLearn outputs satisfies that $(\mathcal{D}|x_W = s)$ for all $s \in \{0,1\}^{|W|}$ is moment-close to some mixture $\mathcal{D}'$ of fewer than r product distributions.

Lemma 7.8.4. *The following holds for any $\tau_{trunc} > 0$ for which $\varepsilon_{samp}(r) \leq \tau_{trunc}/2$ holds for all $r > 1$. If there exists an $(\varepsilon_{samp}(r), s(r))$-moment-close rank-$r$ realization of $\mathcal{D}$ by mixing weights π and marginals matrix $\mathbf{m}$ but none of the mixtures $\overline{\mathcal{D}}$ output by an ε_{samp}-sample-rich run of* NonDegenerateLearn *on input $\mathcal{D}$ satisfies $d_{TV}(\mathcal{D},\overline{\mathcal{D}}) \leq 2\varepsilon$, then in the set of subsets $\mathcal{U}$ in the output there exists some W such that for all $s \in \{0,1\}^{|W|}$, either $Pr_{y\sim\mathcal{D}}[y_W = s] \leq \tau_{trunc}$ or there is a $(\delta, s(r-1))$-moment-close rank-r' realization $\mathcal{D}'$ of $(\mathcal{D}|x_W = s)$, where $\delta \triangleq 3\sigma_{cond}(r)k^2/\sqrt{2} + 2^{r+3}\varepsilon_{samp}(r)/\tau_{trunc}$ and $r' < r$.*

Proof. Let $\tilde{\mathcal{D}}$ denote the mixture realized by mixing weights π and marginals matrix $\mathbf{m}$. Because NonDegenerateLearn fails to output a mixture of r product distributions, by the contrapositive of Lemma 7.8.2 we know that $\sigma^\infty_{\min}(\mathbf{M}|_{\mathcal{R}_k^\dagger(J\cup\{i\})}) \leq \sigma_{cond}(r)$. Let $W = J \cup \{i\}$. By Lemma 7.5.10, for any $s \in \{0,1\}^{|W|}$ there exists a mixture of at most $r-1$ product distributions $\mathcal{D}'$ such that $(\tilde{\mathcal{D}}|x_W = s)$ and $\mathcal{D}'$ are $(\sigma_{cond}(r)\cdot 3k^2/\sqrt{2}, s(r-1))$-moment-close. And because $|W| \leq r+1$, $W \in \mathcal{U}$.

It remains to show that $(\tilde{\mathcal{D}}|x_W = s)$ and $(\mathcal{D}|x_W = s)$ are moment-close. Take any $T \subseteq [n]\backslash W$ of size at most $s(r-1)$. By Bayes' we have

$$\left|\mathbb{E}_{\mathcal{D}|x_W=s}[x_T] - \mathbb{E}_{\tilde{\mathcal{D}}|x_W=s}[x_T]\right| = \left|\frac{\Pr_{y\sim\mathcal{D}}[y_W = s \wedge y_T = 1^{|T|}]}{\Pr_{y\sim\mathcal{D}}[y_W = s]} - \frac{\Pr_{y\sim\tilde{\mathcal{D}}}[y_W = s \wedge y_T = 1^{|T|}]}{\Pr_{y\sim\tilde{\mathcal{D}}}[y_W = s]}\right|$$
$$\leq \frac{\varepsilon_{\text{samp}}(r)\cdot 2^{|W|}\left(\Pr_{y\sim\mathcal{D}}[y_W = s \wedge y_T = 1^{|T|}] + \Pr_{y\sim\tilde{\mathcal{D}}}[y_W = s \wedge y_T = 1^{|T|}]\right)}{\Pr_{y\sim\mathcal{D}}[y_W = s]\cdot(\Pr_{y\sim\mathcal{D}}[y_W = s] - \varepsilon_{\text{samp}}(r))}$$

$$\leq 2^{|W|+2} \cdot \varepsilon_{\mathrm{samp}}(r)/\tau_{trunc}^2 \leq 2^{r+3} \cdot \varepsilon_{\mathrm{samp}}(r)/\tau_{trunc}^2,$$

The first inequality follows from 1) the fact that the probability that $y_W = s$ and $y_T = 1^{|T|}$ may be written as a linear combination (with ±1 coefficients) of at most $2^{|W|}$ moments of degree at most

$$|W| + |T| \leq r + 1 + s(r-1) \leq s(r),$$

where the last inequality follows from (7.31), and 2) the fact that $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are $(\varepsilon_{\mathrm{samp}}(r), s(r))$-close. The second inequality follows from the fact that the probabilities in the numerator are both bounded above by 1, while the probabilities in the denominator are bounded below by τ_{trunc} and $\tau_{trunc} - \varepsilon_{\mathrm{samp}}(r) \geq \tau_{trunc}/2$ respectively.

By the triangle inequality, we conclude that $\mathcal{D}'$ and $(\mathcal{D}|x_W = s)$ are $(\delta, s(r-1))$-moment-close, where δ is as defined above. □

7.8.3 Correctness of N-List

Finally, we are ready to prove Theorem 7.1.6. We will prove the following stronger statement which is more amenable to induction.

Theorem 7.8.5. *There is an absolute constant $c_{27} > 0$ for which the following holds. Let*

$$\sigma_{cond}(r) = \left(\frac{c_{27}\min(\tau_{trunc}, \varepsilon^2)}{2^k n}\right)^r, \qquad \varepsilon_{samp}(r) = \sigma_{cond}(r) \cdot \frac{c_{26}\varepsilon^2}{k^3 n}, \qquad \tau_{trunc}, \delta_{edge} \leq \frac{2\varepsilon}{2^{r+1}\cdot 5}.$$

If there is an $(\varepsilon_{samp}(r), s(r))$-moment-close rank-$r$ realization of $\mathcal{D}$ by mixing weights π and marginals matrix $\mathbf{m}$, then an $(\varepsilon_{samp}(r), \delta_{edge}, \tau_{trunc}^r)$-sample-rich run of N-List *on $\mathcal{D}$ will output a distribution $\overline{\mathcal{D}}$ for which $d_{TV}(\mathcal{D}, \overline{\mathcal{D}}) \leq 10^r\varepsilon$.*

To prove this, we first record a simple fact about sampling trees, similar in spirit to Lemma 7.7.14.

Lemma 7.8.6. *Let $\mathcal{T}$ be a sampling tree such that for each of the* immediate *descendants $v_{W,s}$ of the root, either $d_{TV}(\mathcal{D}_{W,s}, \mathcal{D}|x_W = s) \leq \varepsilon'$, or $Pr_{y\sim\mathcal{D}}[y_W = s] \leq \tau$ for $\tau = \frac{2\varepsilon}{2^{|W|}\cdot 5}$. Suppose further that all weights $w_{\emptyset,W,\emptyset,s}$ satisfy $|w_{\emptyset,W,\emptyset,s} - Pr_{y\sim\mathcal{D}}[y_W = s]| \leq \delta_{edge}$ for $\delta_{edge} = \frac{2\varepsilon}{2^{|W|}\cdot 5}$. Then $d_{TV}(\mathcal{D}^*, \mathcal{D}) \leq \varepsilon' + \varepsilon$, where $\mathcal{D}^*$ is the distribution associated to $\mathcal{T}$.*

Proof. We wish to bound $\sum_{x\in\{0,1\}^n} |\Pr_{\mathcal{D}^*}[x] - \Pr_{\mathcal{D}}[x]| = 2d_{\mathrm{TV}}(\mathcal{D}^*, \mathcal{D})$. Denote by $\mathcal{U}_{trunc}$ the set of all $x \in \{0,1\}^n$ for which $x_W = s$ and $\Pr_{y\sim\mathcal{D}}[y_W = s] \le \tau$. Also, let M be the number of $s \in \{0,1\}^{|W|}$ for which $\Pr_{y\sim\mathcal{D}}[y_W = s] \le \tau$. Then

$$\sum_{x\in\mathcal{U}_{trunc}} \Pr_{\mathcal{D}^*}[x] \le \sum_{x\in\mathcal{U}_{trunc}} \Pr_{\mathcal{D}}[x] + M\delta_{edge} \le M(\tau + \delta_{edge}),$$

so by triangle inequality we have that $\sum_{x\in\mathcal{U}_{trunc}} |\Pr_{\mathcal{D}^*}[x] - \Pr_{\mathcal{D}}[x]| \le 2^{|W|}(2\tau + \delta_{edge})$. For $x \notin \mathcal{U}_{trunc}$, decompose x as $s \circ t$ for $s \in \{0,1\}^{|W|}$ and $t \in \{0,1\}^{n-|W|}$. By Bayes' we have $\Pr_{\mathcal{D}^*}[x] = w_{\emptyset,W,\emptyset,s} \cdot \Pr_{\mathcal{D}_{W,s}}[t]$ and $\Pr_{\mathcal{D}}[x] = \Pr_{y\sim\mathcal{D}}[y_W = s] \cdot \Pr_{\mathcal{D}|x_W=s}[t] \triangleq w_s \cdot p_t$. For all $t \in \{0,1\}^{n-|W|}$, let $\Pr_{\mathcal{D}_{W,s}}[t] = p_t + \delta_t$ for some $\delta_t > 0$. Because $d_{\mathrm{TV}}(\mathcal{D}_{W,s}, \mathcal{D}|x_W = s) \le \varepsilon'$, we have that $\sum_t |\delta_t| \le 2\varepsilon'$ for all immediate descendants $v_{W,s}$ of the root of $\mathcal{T}$. Moreover, by assumption we have that $|w_s - w_{\emptyset,W,\emptyset,s}| \le \delta_{edge}$. We conclude that

$$\begin{aligned}\sum_{x\notin\mathcal{U}_{trunc}} |\Pr_{\mathcal{D}^*}[x] - \Pr_{\mathcal{D}}[x]| &= \sum_{\substack{s\in\{0,1\}^{|W|}:\\ d_{\mathrm{TV}}(\mathcal{D}_{W,s},\mathcal{D}|x_W=s)\le\varepsilon'}} w_s \cdot \left(\delta_{edge} + 2\delta_{edge}\varepsilon' + \sum_{t\in\{0,1\}^{n-|W|}} |\delta_t|\right) \\ &\le 2\varepsilon' + (2^{|W|} - M)(\delta_{edge}\varepsilon' + \delta_{edge}).\end{aligned}$$

When $\delta_{edge}, \tau \le \frac{2\varepsilon}{2^{|W|}\cdot 5}$, we conclude that $d_{\mathrm{TV}}(\mathcal{D}^*, \mathcal{D}) \le \varepsilon' + \varepsilon$. □

Remark 7.8.7. *Lemma 7.8.6 is weaker than Lemma 7.7.14 in that it can be used to give an inductive proof of Lemma 7.7.14 with far worse guarantees. Specifically, we would need τ_{trunc} in the statement of Lemma 7.7.14 to be $O(\varepsilon^d)$ instead of $O(\varepsilon)$, where $d \le k$ is the depth of the sampling tree.*

However, using Lemma 7.8.6 instead of Lemma 7.7.14 here greatly simplifies our inductive analysis of N-List*. And the need to grid the entries of* **m** *in* NonDegenerateLearn *already makes our algorithm run in time $(n/\varepsilon)^{\Omega(k^2)}$ to begin with, so we can afford the cost of this simplification.*

Proof of Theorem 7.8.5. We induct on r. If $r = 1$ so that π and **m** realize a single product distribution, then for any $W \subseteq [n]$, $\mathbf{M}|_{\mathcal{R}_r^\dagger(W)}$ is a single column whose entries contain 1 (corresponding to the empty set), so $\sigma_{\min}^{\infty}(\mathbf{M}|_{\mathcal{R}_r^\dagger(W)}) \ge 1 > \sigma_{cond}(1)$. The base case then follows by Corollary 7.8.3.

Suppose $r > 1$ and let $\mathcal{M}$, $\mathcal{U}$ be the output of NonDegenerateLearn on $\mathcal{D}$ and counter r. For each $W \in \mathcal{U}$, we are recursively calling N-List on $(\mathcal{D}|x_W = s)$ for each $s \in \{0,1\}^{|W|}$ and connecting the resulting sampling trees to $v_{\emptyset,\emptyset}$ to obtain some sampling tree rooted at $v_{\emptyset,\emptyset}$. Call this collection of sampling trees $\mathcal{M}'$. We are done by Lemma 7.6.2 if we can show that $\mathcal{S} = \mathcal{M} \cup \mathcal{M}'$ contains a distribution close to $\mathcal{D}$.

Suppose $\mathcal{M}$ contains no distribution $\overline{\mathcal{D}}$ for which $d_{\mathrm{TV}}(\mathcal{D}, \overline{\mathcal{D}}) \le 2\varepsilon$. Then by Lemma 7.8.4, there is some $W \in \mathcal{U}$ such that $(\mathcal{D}|x_W = s)$ is $(\delta, s(r-1))$-moment-close to a mixture of at most $r-1$ product distributions $\mathcal{D}'$ for every s for which $\Pr_{y\sim\mathcal{D}}[y_W = s] > \tau_{trunc}$, where $\delta = 3\sigma_{cond}k^2/\sqrt{2} + 2^{r+3}\varepsilon_{\mathrm{samp}}(r)/\tau_{trunc}$. One can check that for the above choice of $\varepsilon_{\mathrm{samp}}(\cdot)$ and $\sigma_{cond}(\cdot)$, $\delta < \varepsilon_{\mathrm{samp}}(r-1)$. By induction on r, the distribution output by N-List on input $(\mathcal{D}|x_W = s)$ is $10^{r-1}\varepsilon$-close to $(\mathcal{D}|x_W = s)$ for each s such that $\Pr_{y\sim\mathcal{D}}[y_W = s] > \tau_{trunc}$. By Lemma 7.8.6 we conclude that in $\mathcal{M}'$, there is some distribution $\overline{\mathcal{D}}$ for which $d_{\mathrm{TV}}(\mathcal{D}, \overline{\mathcal{D}}) \le 10^{r-1} + \varepsilon$, so by Lemma 7.6.2, Select$(\mathcal{S}, \mathcal{D})$ outputs a distribution at most $9.1(10^{r-1}\varepsilon + \varepsilon) \le 10^r\varepsilon$-close to $\mathcal{D}$. □

Proof of Theorem 7.1.6. Apply Lemma 7.6.7 with $\tau_{trunc} = \frac{2\varepsilon}{2^{k+1.5}}$, $Z = n^{O(k^2)}$, $M = n^{O(k)} \cdot 2^{k^2}$, $U = n^{O(k)}$, $S = k+1$, $T(r) = (nk^2/\varepsilon)^{O(k^2)}$, $\varepsilon_{\mathrm{samp}}(\cdot)$ as defined in Theorem 7.8.5, and $\varepsilon_{select} = O(\varepsilon)$ to get that achieving a $(\varepsilon_{\mathrm{samp}}, \delta_{edge}, \tau_{trunc})$-sample-rich run of N-List on $\mathcal{D}$ with counter k with probability $1-\delta$ requires $\mathrm{poly}(n,k,1/\varepsilon)^{k^2}\ln(1/\delta)$ time and $n^{O(k^2)}\varepsilon^{O(k)}\ln(1/\delta)$ samples. By taking $\delta_{edge} = \frac{2\varepsilon}{2^{k+1.5}}$, we conclude by Theorem 7.8.5 that the output of N-List is $10^k\varepsilon$-close to $\mathcal{D}$. Replace ε by $\varepsilon/10^k$ and the result follows. □

7.9 Appendix: Application to Learning Stochastic Decision Trees

In this section, we prove Theorem 7.1.3. We begin with a warmup:

Example 7.9.1 (Parity and juntas)**.** *The uniform distribution $\mathcal{D}$ over the positive examples of a k-junta $f : \{0,1\}^n \to \{0,1\}$ is a mixture of at most 2^k subcubes in $\{0,1\}^n$. Let $I \subseteq [n]$ be the k coordinates that f depends. Every $s \in \{0,1\}^{|I|}$ for which $f(x) = 1$ for all x satisfying*

$x_I = s$ corresponds to a subcube with mixture weight $1/N$, where $N \le 2^k$ is the number of such s (e.g. when f is a parity, $N = 2^{k-1}$). In the same way we can show that the uniform distribution over the negative examples is also a mixture of at most 2^k subcubes.

So given access to examples $(x, f(x))$ where x is uniformly distributed over $\{0,1\}^n$, we can learn f as follows. With high probability, we can determine $b^* \in \{0,1\}$ for which f outputs b^* on at least $1/3$ of the inputs. As we have shown in this chapter, our algorithm can then learn some $\mathcal{D}'$ that is ε-close to the uniform distribution over $\{x : f(x) = b^*\}$. We then output the hypothesis g given by $g(x) = b^*$ if $\mathcal{D}'(x) \le 1/2^{n+1}$ and $g(x) = 1 - b^*$ otherwise. It is easy to see that g is ε-accurate.

This approach can handle mild random classification noise γ: if we take the distribution over examples (x, b) where x is drawn from the uniform distribution over $\{0,1\}^n$ and b is labeled by $f(x)$ with probability $1-\gamma$ and $1 - f(x)$ with probability γ, and we condition on $b = 1$, the resulting distribution is still a mixture of subcubes: every s for which $f(x) = 1$ for all $x_I = s$ corresponds to a subcube of weight $(1-\gamma)/N$, and every other s corresponds to a subcube of weight γ/N. This mixture is $O(\gamma)$-far from the uniform distribution over $\{x : f(x) = 1\}$, so in the above analysis, our algorithm would give an $(\varepsilon + O(\gamma))$-accurate hypothesis.

Finally, note that if mixing weights π and marginals matrix $\mathbf{m}$ realize $\mathcal{D}$, then $\mathbf{m}_i \in \{0,1\}^k$ if f depends on coordinate k, and $\mathbf{m}_i = (1/2, ..., 1/2)$ otherwise, meaning the rows of $\mathbf{M}$ are spanned by all entrywise products of degree less than $\log_2(N) \le k$, rather than $2\log(N)$ as is required in general by N-List. *So the algorithm we described above has the same performance as the brute-force algorithm.*

The above example serves simply to suggest the naturality of the problem of learning mixtures of subcubes, but because there are strong SQ lower bounds against learning sparse noisy parity [BFJ+94], it's inevitable that our algorithm gives no new improvements over such problems. We now describe an application of N-List which does achieve a new result on a classical learning theory problem. First recall the definition of stochastic decision trees from Definition 1.2.20.

Lemma 7.9.2. *For any k-leaf stochastic decision tree T on n bits, the distribution of $(x, b) \sim D_T$ conditioned on $b = 1$ is a mixture of k subcubes.*

Proof. Consider any path p in T from the root to a leaf labeled with 1. If along this path there are m decision nodes corresponding to some variables $i_1, ..., i_m \in [n]$ and with outgoing edges $b_1, ..., b_m$, then any $x \in \{0,1\}^n$ from the subcube corresponding to the conjunction $(x_{i_1} = b_1) \wedge \cdots \wedge (x_{i_m} = b_m)$ evaluates to 1 along this path with probability equal to the product μ_p of the edge weights along this path which emanate from stochastic nodes. So the distribution of $(x,b) \sim D_T$ conditioned on $b = 1$ is a mixture of k such subcubes, where the p-th subcube has mixture weight proportional to $\mu_p/2^{d_p}$, where d_p is the number of decision nodes along path p. □

The following immediately implies Theorem 7.1.3.

Lemma 7.9.3. *Let T be any k-leaf stochastic decision tree corresponding to a joint probability distribution D_T on $\{0,1\}^n \times \{0,1\}$. Given access to samples from D_T, D_T can be learned to within total variation distance ε with probability at least $1-\delta$ in time $O_{k,s}(n^{O(s+\log k)}(1/\varepsilon)^{O(1)} \log 1/\delta)$ and with sample complexity $O_{k,s}((\log n/\varepsilon)^{O(1)} \log 1/\delta)$*

Proof. Denote by A our algorithm for learning mixtures of subcubes, given by Theorem 7.1.1. To learn D_T, we can first estimate $\pi(b) \triangleq \Pr_{(x,b')\sim D_T}[b' = b] \geq 1/3$ for each $b \in \{0,1\}$ to within accuracy ε and confidence $1 - \alpha/3$ by drawing $O((1/\varepsilon)^2 \log(1/\alpha))$ samples, by Fact 7.6.6. We pick $b^* \in \{0,1\}$ for which $\Pr_{(x,b)\sim D_T}[b = b^*] \geq 1/3$ and denote our estimate for $\pi(b^*)$ by $\pi'(b^*)$.

By Lemma 7.9.2, $\mathcal{D}$ is a mixture of k subcubes, so we can run A with error parameter $\varepsilon/2$ and confidence parameter $\alpha/3$ on $\mathcal{D}$ and get a distribution $\mathcal{D}'$ for which $d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}') \leq \varepsilon/4$. Our algorithm outputs the distribution D' given by $D'(x, b^*) = \pi'(b^*) \cdot \mathcal{D}(x)$ and $D'(x, 1-b^*) = 1 - \pi'(b^*) \cdot \mathcal{D}(x)$.

Now because $D_T(x, b^*) = \pi_{b^*} \cdot \mathcal{D}(x)$, we have that

$$\sum_{x\in\{0,1\}^n} |D_T(x,b^*) - D'(x,b^*)| \leq \frac{\varepsilon}{2} \cdot \sum_{x\in\{0,1\}^n} \mathcal{D}(x) + \pi'(b^*) \cdot \sum_{x\in\{0,1\}^n} |\mathcal{D}(x) - \mathcal{D}'(x)| \leq \frac{\varepsilon}{2} + 2 \cdot \frac{\varepsilon}{4} = \varepsilon.$$

We thus also get that $\sum_{x\in\{0,1\}^n} |D_T(x, 1-b^*) - D'(x, 1-b^*)| = \sum_{x\in\{0,1\}^n} |D_T(x,b^*) - D'(x,b^*)| \leq \varepsilon$, so $d_{\mathrm{TV}}(D_T, D') \leq \varepsilon$ as desired. □

Chapter 8

Mixed Linear Regression

8.1 Introduction

The second mixture model we study in this thesis, mixtures of linear regressions (or MLRs for short), is a popular generative model that has been studied extensively in machine learning and theoretical computer science. We recall the setup from Definition 1.2.22 in slightly different notation: we have k unknown *mixing weights* $p_1, \ldots, p_k$ which are non-negative and sum to 1, k unknown *regressors* $w_1, \ldots, w_k \in \mathbb{R}^d$, and a *noise rate* $\varsigma \geq 0$. A sample from the MLR is drawn as follows: we first select $i \in [k]$ with probability p_i, then we receive (x, y) where $x \in \mathbb{R}^d$ is distributed as $\mathcal{N}(0, \mathrm{Id})$ and

$$y = \langle w_i, x \rangle + \eta,$$

where $\eta \sim \mathcal{N}(0, \varsigma^2)$. As mentioned in Section 1.2.3, this model has applications to problems ranging from trajectory clustering [GS99] to phase retrieval [BCE06, CSV13, NJS13] and is also widely studied as a natural non-linear generative model for supervised data [FS10, CYC13, CL13, YCS14, YCS16, ZJD16, SJA16, KYB17, BWY17, KQC+18, LL18, KC19].

The basic learning question for MLRs is as follows: given i.i.d. samples $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ from an unknown MLR, can we learn the parameters of the underlying MLR? To ensure that the parameters are identifiable, it is also typically assumed that the regressors are separated in some way, e.g. there is some $\Delta > 0$ so that $\|w_i - w_j\|_2 \geq \Delta$ for all $i \neq j$.

Despite the apparent simplicity of the problem, efficiently learning MLRs given samples has proven to be a surprisingly challenging task. Even in the special case where $\varsigma = 0$, that is, we assume that there is no noise on the samples, the fastest algorithms for this problem run in time depending on $k^{\Omega(k)}$ [LL18, ZJD16]. It turns out that there are good reasons for this barrier.

Previous algorithms for this problem with end-to-end provable guarantees—and indeed, the vast majority of statistical learning algorithms in general—build in some form or another on the *method of moments* paradigm. At a high level, these methods require that there exists some statistic which depends only on low degree moments of the unknown distribution, so that a sufficiently good estimate of this statistic will uniquely identify the parameters of the distribution. This includes widely-used techniques based on tensor decomposition [CL13, YCS16, ZJD16, SJA16], and SDP hierarchies such as the Sum-of-Squares meta-algorithm [KKK19, RY19]. If degree t moments are necessary to devise such a statistic, then these methods require $\exp(\Omega(t))$ sample and computational complexity.

Unfortunately, for MLRs, it is not hard to demonstrate pairs of mixtures some of whose parameters are far apart from each other, where all moments of degree at most $2k-1$ of the two mixtures agree exactly (see Appendix 8.10 for more details). As a result, any moment-based estimator would need to use moments of degree at least $\Omega(k)$, and hence require a runtime of $\exp(\Omega(k))$. This imposes a natural bottleneck: any algorithm that hopes to achieve sub-exponential time must somehow incorporate additional information about the geometry of the underlying learning problem.

A related problem, which shares a similar bottleneck, is the problem of learning mixtures of Gaussians under the assumption of *angular separation.* A concrete instantiation of this problem is a model we call *learning mixtures of hyperplanes.* A mixture of hyperplanes is parameterized by mixing weights $p_1, \ldots, p_k$, a separation parameter $\Delta > 0$, and k unit vectors $v_1, \ldots, v_k$ satisfying $\|v_i \pm v_j\|_2 \geq \Delta$ for all $i \neq j$ (note that the reason for the $\pm$ is that the directions of a mixture of hyperplanes are only identifiable up to sign). To draw a sample, we first draw $i \in [k]$ with probability p_i, and then draw a sample from $\mathcal{N}(0, \mathrm{Id} - v_i v_i^\top)$.

As before, the corresponding learning question is the following: given samples from an unknown mixture of hyperplanes, can one recover the underlying parameters? This problem can be thought of as a particularly hard case of the well-studied problem of *subspace recovery*, where current techniques would require time which is exponential in k.

In this paper, we give algorithms which are able to achieve strong recovery guarantees for the problems of learning MLRs and learning mixtures of hyperplanes, and which run in time which is sub-exponential in k. To the best of our knowledge, this is the first algorithm for the basic problem of learning MLRs which achieves sub-exponential runtime without placing strong additional assumptions on the model. At a high level, our key insight is that while low degree moments of the MLR are unable to robustly identify the instance, low degree moments of suitable projections of the *Fourier transform* of the MLR can be utilized to extract nontrivial information about the regressors. We then give efficient algorithms for computing such "Fourier moments" by leveraging algorithms for univariate density estimation [CDSS14b, ADLS17]. This allows us to dramatically improve the runtime and sample complexity of the *moment descent* algorithm of [LL18], and allows us to obtain our desired sub-exponential runtime. We believe that this sort of algorithmic application of the continuous Fourier transform and of univariate density estimation to a high dimensional learning problem is novel, and may be of independent interest.

8.1.1 Our Contributions

Here, we describe our contributions in more detail. For simplicity of exposition, in this section we will assume that the mixing weights are uniform, i.e. $p_i = 1/k$ for all $i \in [k]$, although as we show, our algorithms can handle non-uniform mixing weights.

Our main results for learning MLRs are twofold. Throughout the paper we let $\widetilde{O}(f) = O(f \log^c(f))$ for some universal constant c. First, in the well-studied case where there is no regression noise, we show:

Theorem 8.1.1 (Informal, see Theorem 8.6.2)**.** *Assume that the noise rate $\varsigma = 0$. Let $w_1, \ldots, w_k \in \mathbb{R}^d$ be the parameters of an unknown MLR $\mathcal{D}$ with separation Δ. Then, there is an algorithm which takes $N = \widetilde{O}(d) \cdot \exp(\widetilde{O}(\sqrt{k}))$ samples from $\mathcal{D}$, runs in time $\widetilde{O}(N \cdot d)$, and outputs $\widetilde{w}_1, \ldots, \widetilde{w}_k \in \mathbb{R}^d$ so that with high probability, there exists some permutation*

$\pi : [k] \to [k]$ satisfying

$$\left\| w_i - \widetilde{w}_{\pi(i)} \right\|_2 \leq \frac{\Delta}{k^{100}} \ , \forall i \in [k].$$

By combining this "warm start" with the boosting result of [LL18], we can also obtain arbitrarily good accuracy with minimal overhead in both the sample complexity and runtime. See Section 8.6 for more details.

Secondly, in the case when the noise rate ς is large, we can also obtain a similar result, though with an additional exponential dependence on Δ:

Theorem 8.1.2 (Informal, see Theorem 8.7.1)**.** *Let $w_1, \ldots, w_k \in \mathbb{R}^d$ be the parameters of an unknown MLR $\mathcal{D}$ with separation Δ, and noise rate $\varsigma > 0$. Then, there is an algorithm which takes $N = \widetilde{O}(d) \cdot \exp(\widetilde{O}(\sqrt{k}/\Delta^2))$ samples from $\mathcal{D}$, runs in time $\widetilde{O}(N \cdot d)$, and outputs $\widehat{w}_1, \ldots, \widehat{w}_k \in \mathbb{R}^d$ so that with high probability, there exists some permutation $\pi : [k] \to [k]$ satisfying*

$$\left\| w_i - \widehat{w}_{\pi(i)} \right\|_2 \leq \frac{\Delta}{k^{100}} + O(\varsigma) \ , \forall i \in [k].$$

In particular, if $\Delta = \Omega(1)$, we again attain runtime which is sub-exponential in k. In the special case when the mixing weights are all known, and assuming that $\varsigma = O(\frac{\Delta}{k^2 \text{polylog}(k)})$, by combining this result with the local convergence result of [KC19], we can again attain arbitrarily good accuracy by slightly increasing the runtime; see Section 8.7.5 and Theorem 8.7.33 for details.

Finally, for the problem of learning mixtures of hyperplanes, we are able to obtain qualitatively similar results. Again, for simplicity of exposition, we assume the mixing weights are uniform just in the current section. We obtain:

Theorem 8.1.3 (Informal, see Theorem 8.8.1)**.** *Let $\varepsilon > 0$, and let $v_1, \ldots, v_k \in \mathbb{R}^d$ be the parameters of a mixture of hyperplanes $\mathcal{D}$ with separation $\Delta > 0$. Then, there is an algorithm which takes $N = \widetilde{O}(d) \cdot \exp(\widetilde{O}(k^{0.6}))$ samples from $\mathcal{D}$, runs in time $\widetilde{O}(N \cdot d)$, and which outputs $\widehat{v}_1, \ldots, \widehat{v}_k \in \mathbb{R}^d$ so that with high probability, there is a permutation $\pi : [k] \to [k]$ so that*

$$\left\| v_i - \widehat{v}_{\pi(i)} \right\|_2 \leq \frac{\Delta}{k^{100}} \ , \forall i \in [k].$$

8.1.2 Related Work

Mixtures of linear regressions were introduced in [DV89], and later by [JJ94], under the name of *hierarchical mixtures of experts*, and have been studied extensively in the theory and ML communities ever since. Previous work on the problem with provable guarantees can roughly speaking be divided into three groups. Some of the previous work focuses on special cases of the problem, in particular, when the number of components is small [CYC13, KYB17,BWY17,KQC+18]. In contrast, we focus on the setting where k is quite large, which is the setting which is typically true in applications, but is also much more algorithmically complicated.

Another line of work has focused on demonstrating local convergence guarantees for non-convex methods such as expectation maximization or alternating minimization [FS10, YCS14, YCS16, ZJD16, KYB17, BWY17, KQC+18, LL18, KC19]. These papers demonstrate that given a sufficiently good warm start, non-convex methods are able to boost this warm start to arbitrarily good accuracy. These results should be viewed as largely complementary to our results, as our main result is a method which is able to provably achieve a good warm start. That said, we also demonstrate new algorithms for learning given a warm start that work under a weaker initialization and can tolerate more regression noise than was previously known in the literature.

The final class of results use moment-based methods to learn MLRs. Here, the literature has focused largely on the case of $\varsigma = 0$ and spherical covariates, that is, covariates all drawn from $\mathcal{N}(0, \mathrm{Id})$.[1] A line of work has studied tensor decomposition-based methods [CL13, YCS16, ZJD16, SJA16]. However, these require additional non-degeneracy conditions on the MLR instance beyond separation. Indeed, as we argued in the Introduction, and more formally in Appendix 8.10, moment based methods cannot obtain runtime which is sub-exponential in k. The work that is closest to ours, and that we build off of, is that of [LL18], which demonstrates an algorithm which runs in $2^{\widetilde{O}(k)}$ for learning a MLR under separation conditions. However, as their warm start algorithm is ultimately moment based,

[1] To the best of our knowledge, the primary exception to this is [LL18], which considered noise-less MLRs whose components' covariates are drawn from arbitrary unknown Gaussians satisfying some condition number bounds and obtained a $d \cdot \exp(k^2)$ algorithm in this setting.

since it interacts through the samples through the moment-based univariate GMM learning algorithm of [MV10], it cannot achieve runtime sub-exponential in k.

List-Decodable Regression A related problem to—indeed, a generalization of—the problem of learning MLRs is that of *list-decodable regression* [CSV17, KKK19, RY19]. Here, we assume that we are given a set of data points $(x_1, y_1), \dots, (x_n, y_n)$, where an α-fraction of them come from an unknown linear regression $y_i = \langle w, x_i \rangle + \eta$, where x_i is Gaussian, η is Gaussian noise, and $\alpha < 1/2$. The goal is then to recover a list of $O(1/\alpha)$ possible $w_1, \dots, w_{O(1/\alpha)} \in \mathbb{R}^d$ so that $\|w_i - w\|_2$ is small for at least one element in the list.

It is not hard to see that given a uniform MLR instance, if we feed it into an algorithm for list-decodable regression, the list must contain something which is close to each of the regressors in the MLR instance, as each mixture component is an equally valid solution to the list-decodable regression problem. Thus one could hope that these algorithms for list-decodable regression could yield improved algorithms for learning MLRs as well.

Unfortunately, all known techniques, including the state of the art [KKK19, RY19], either are too weak to be applied to our setting, or use the Sum-of-Squares SDP hierarchy and again interact through the data via estimating high-degree moments of the distribution. As a result, these latter algorithms still suffer runtimes which are exponential in k.

Subspace Clustering The mixtures of hyperplanes problem we consider in this paper can be thought of as a special case of the *subspace clustering* or *hyperplane clustering* problem, where data is thought of as being drawn from a union of linear subspaces. In our problem, we additionally assume that the data is Gaussian within each subspace. The literature on subspace clusterings is vast and we cannot do it justice here; see [PHL04, Vid11, EV13] and references therein for a more complete treatment. On the one hand, the mixture of hyperplanes problem arises naturally in practical contexts of projective motion segmentation [VH07] and hybrid system identification [Bak11]. On the other, it also corresponds to a challenging setting of the problem due to the low codimensionality of the subspaces. Indeed, essentially all algorithms for subspace clustering with provable guarantees either run in time exponential in the dimension of the subspaces (e.g. RANSAC [FB81], algebraic subspace clustering [VMS05], spectral curvature clustering [LLY$^+$12]) or require the codimension to be at least some small but constant *fraction* of the ambient dimension [EV13, CSV13, LMZ$^+$12, TV15]. To our knowledge the only work which addresses the codimension 1 case is [TV17], though their setting and guarantees are quite different from ours.

The Fourier Transform in Distribution Learning One of our main algorithmic tools will be the univariate (continuous) Fourier transform, as a way to estimate Fourier moments of our distribution. In recent years, the question of learning the Fourier transform of a function has attracted a considerable amount of interest in theoretical computer science [HIKP12, IK14, Moi15, PS15, Kap16, CKPS16, Kap17, NSW19]. Our application is somewhat different in that we have explicit access to the function we will take the Fourier transform of.

In the context of distribution learning, the discrete Fourier transform has been used to learn families of distributions such as sums of independent integer random variables [DKS16b], Poisson Binomial distributions [DKS16c], and Poisson multinomial distributions [DKS16b, DKS16a]. These algorithms typically work by exploiting Fourier sparsity of the underlying distribution. However, the way we use the Fourier transform is quite different: we only use it to compute different statistics of the data, namely, the Fourier moments of our distribution.

Univariate Density Estimation Another important algorithmic primitive we use is univariate density estimation and specifically, the piecewise polynomial-based estimators given in [CDSS14b, ADLS17]. Univariate density estimation has a long history in statistics, ML, and theoretical computer science, and a full literature review of the field is out of the scope of this paper; see e.g. [Dia16] for a more comprehensive overview of the literature. However, to the best of our knowledge, there are few previous cases where univariate density estimation has been used as a key tool for a high dimensional learning task.

8.2 Preliminaries

In this section, we give some basic technical preliminaries.

8.2.1 Probabilistic Models

In this section, we formally define the models we consider throughout this paper, namely, mixtures of linear regression and hyperplanes, and some important parameters for these models:

Mixtures of Linear Regressions We start by restating the setup for MLRs, originally introduced in Definition 1.2.22, in slightly different notation:

Definition 8.2.1 (Mixtures of Linear Regressions). *Given mixing weights $p \in [0,1]^k$ with $\sum_{i=1}^k p_i = 1$, regressors $w_1, \cdots, w_k \in \mathbb{R}^d$, and noise rate $\varsigma \geq 0$, the corresponding* mixture of spherical linear regressions *(or simply mixture of linear regressions) is the distribution over pairs $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ where $x \sim \mathcal{N}(0, Id)$ and $y = \langle w_i, x \rangle + g$ for w_i sampled with probability p_i and $g \sim \mathcal{N}(0, \varsigma^2)$.*

When $\varsigma = 0$, we say that the MLR is *noiseless.*

In this paper, we will study the *parameter learning* problem for mixtures of linear regressions, that is, we wish to recover the parameters of the mixture. To this end, we will need some assumptions to ensure that the regressors are uniquely identifiable. These assumptions are standard throughout the literature. Given a mixture of linear regressions $\mathcal{D}$, let $\Delta = \min_{i \neq j} \|w_i - w_j\|_2$ be the minimum L_2 separation among all w_i in the mixture, and we let $p_{\min} = \min_{i \in [k]} p_i$. To normalize the instance, we will also assume that $\|w_i\|_2 \leq 1$ for all $i = 1, \ldots, k$. However, more generally, our algorithms will have a mild polynomial dependence on the maximum L_2 norm of any w_i. We omit this case for simplicity of exposition.

Mixtures of Hyperplanes We now turn our attention to mixtures of hyperplanes. Formally:

Definition 8.2.2. *Given mixing weights $p \in [0,1]^k$ with $\sum_{i=1}^k p_i = 1$ and unit vectors $v_1, \cdots, v_k \in \mathbb{S}^{d-1}$, the corresponding* mixture of hyperplanes *is the distribution with law given by $\sum_{i=1}^k p_i \mathcal{N}(0, \Pi_i)$ where $\Pi_i \triangleq Id - v_i v_i^\top \in \mathbb{R}^{d \times d}$.*

As before, we need some assumptions on the parameters to ensure identifiability. Like before, let $p_{\min} = \min_{i \in [k]} p_i$. Because v_i are now only identifiable up to sign, we define Δ to be the minimum quantity such that or all $i \neq j$ and $\varepsilon_i, \varepsilon_j \in \{\pm 1\}$, $\|\varepsilon_i v_i - \varepsilon_j v_j\|_2 \geq \Delta$.

8.2.2 Miscellaneous Notation

We collect some notation and terminology specific to this chapter. For real-valued functions $\mathcal{F} : \mathbb{R} \to \mathbb{R}$ and $p \in \mathbb{N}$, we will use $\mathcal{M}_p(\mathcal{F})$ to denote $\int_{-\infty}^{\infty} x^p \cdot \mathcal{F}(x)\,\mathrm{d}x$.

Given $v \in \mathbb{R}^{d+1}$, define $\Sigma_\eta(v) \triangleq \begin{pmatrix} \mathrm{Id}_d & v \\ v^\top & \|v\|^2 + \eta^2 \end{pmatrix}$. Note that this is the covariance matrix of a single spherical linear regression with noise variance η^2. When $\eta = 0$, we will denote this matrix by $\Sigma(v)$.

We will sometimes refer to a univariate mixture $\mathcal{F}$ of zero-mean Gaussians with mixing weights $p \in \Delta^k$ and variances $\sigma_1^2, ..., \sigma_k^2$ as a mixture of k univariate zero-mean Gaussians "with parameters $(\{p_i\}_{i\in[k]}, \{\sigma_i\}_{i\in[k]})$." We will define $\sigma_{\min}(\mathcal{F}) \triangleq \min_{i\in[k]} \sigma_i$ and $\sigma_{\max}(\mathcal{F}) \triangleq \max_{i\in[k]} \sigma_i$ and refer to $\sigma_{\min}(\mathcal{F})^2$ and $\sigma_{\max}(\mathcal{F})^2$ as the *minimum* and *maximum variance* of $\mathcal{F}$, respectively.

Finally, we will need the following monotonicity property of moments of Gaussians, restricted to the tails of the Gaussian:

Fact 8.2.3. *Let $\sigma^* > 0$ and $\tau > \sigma^*$, and let $p \in \mathbb{N}$ be even. Then*

$$\int_{[-\tau,\tau]^c} \mathcal{N}(0, \sigma^2; x) \cdot x^p \,\mathrm{d}x < \int_{[-\tau,\tau]^c} \mathcal{N}(0, (\sigma^*)^2; x) \cdot x^p \,\mathrm{d}x$$

for all $0 < \sigma < \sigma^*$.

We defer the proof of this fact to Section 8.12.2.

8.3 Overview of Techniques

In this section, we give a high-level overview of how our algorithms work. For clarity of exposition, in this subsection we will assume $p_{\min} = 1/k$ and $\Delta = \Theta(1)$.

8.3.1 Fourier Moment Descent

We first describe our techniques that achieve Theorem 8.1.1, before describing how to adapt these techniques to achieve Theorems 8.1.2 and 8.1.3.

We begin by briefly recapping the *moment descent* algorithm of [LL18]. Moment descent is an iterative algorithm which attempts to find the parameters of one component at a time as follows. Let $w_1, \ldots, w_k \in \mathbb{R}^d$ be the parameters of a MLR $\mathcal{D}$ with separation $\Delta > 0$ and noise rate $\varsigma = 0$, and again for simplicity let us assume that the mixing weights are uniform. To learn a single regressor, the idea is to maintain a guess $a_t \in \mathbb{R}^d$ for one of the regressors at each time step t, and iteratively refine it by making random steps and checking progress. The measure of progress they consider is simply

$$\sigma_t^2 \triangleq \min_{i \in [k]} \|w_i - a_t\|_2^2 \; .$$

Concretely, the algorithm proceeds as follows. First, by a straightforward PCA step, we can essentially assume that $d \leq k$. Then, given a guess a_t, the moment descent procedure updates by sampling a random unit vector $z \in \mathbb{S}^{d-1}$, defining $a'_{t+1} \triangleq a_t - \eta_t \cdot z \in \mathbb{R}^d$ for some learning rate η_t, and letting

$$(\sigma'_t)^2 \triangleq \min_{i \in [k]} \|w_i - a_t\|_2^2 \; .$$

In general, for $a \in \mathbb{R}^d$, the univariate distribution of the residual $y - \langle a, x \rangle$, where $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ is sampled from $\mathcal{D}$, is distributed as

$$\frac{1}{k} \sum_{i=1}^{k} \mathcal{N}(\mu, \|w_i - a\|_2^2) \; ,$$

that is, it is distributed as a mixture of univariate Gaussians with mixing weights which are the same as those of $\mathcal{D}$, and variances which are equal to the squared L_2 distances between the regressors and a. In particular, to estimate $\min_{i \in [k]} \|w_i - a_t\|_2^2$, one can learn the univariate mixture sufficiently well via [MV10], and simply read off the minimum variance. By doing so, they can check if $\sigma'_t < \sigma_t$. If so, they set $a_{t+1} = a'_{t+1} \in \mathbb{R}^d$, and repeat.

The main bottleneck in this routine is the univariate learning step. Specifically, the algorithm of [MV10] relies on the method of moments to learn the parameters of the univariate mixture of Gaussians, and as a result, takes $k^{O(k)}$ samples and time. In fact, this is inherent: [MV10] demonstrates that $k^{\Omega(k)}$ samples are necessary to learn the parameters of a mixture of Gaussians, precisely by leveraging moment matching instances.

However, all we need is an estimate of the minimum variance of the mixture of Gaussians. One can first observe that it is possible to estimate the *maximum* variance of a component in a univariate mixture of Gaussians based on a sufficiently high degree moment. This is because the p-th moment of a uniform mixture of Gaussians $\mathcal{F}$ with variances $s_1^2, \ldots, s_k^2$ has the following form, for p even:

$$\mathbb{E}_{Z\sim\mathcal{F}}[Z^p] = \sum_{i=1}^k \frac{1}{k} \cdot s_i^p \cdot (p-1)!! = [c/k, 1] \cdot \sigma_{\max}(\mathcal{F})^p \cdot p^{p/2} \ , \tag{8.1}$$

for $\sigma_{\max}(\mathcal{F}) \triangleq \max_{i\in[k]} s_i$ and some universal constant $c > 0$. Therefore, for any $\kappa > 0$, if we set $p = \Theta(\log k/\log(1+\kappa)) = \Theta(\kappa^{-1}\log k)$, we have that

$$p^{-1/2}\mathbb{E}_{Z\sim\mathcal{F}}[Z^p]^{1/p} \in [1 - \Theta(\kappa), 1] \cdot \sigma_{\max}(\mathcal{F})$$

which yields a $(1+\kappa)$ approximation to the maximum variance approximation to the largest variance of $\mathcal{F}$. Moreover, we can estimate the left-hand side in $p^{O(p)} = 2^{\widetilde{O}(\kappa^{-1}\log k)}$ samples:

Lemma 8.3.1 (Concentration of empirical moments)**.** *Let $p \in \mathbb{N}$ be even and $t \in \mathbb{N}$, and let $\delta, \beta > 0$. Then for*

$$N = k^{-2} \cdot \beta^{-2} \cdot p^{\Theta(p)} \cdot \ln(1/\delta)^{\Theta(p)} \cdot \sigma_{\max}(\mathcal{F})^{\Theta(p)},$$

we have that

$$\Pr_{Z_1,\cdots,Z_N\sim\mathcal{F}}\left[\frac{1}{N}\sum_{i=1}^N Z_i^p = (1\pm\beta)\cdot\mathbb{E}_{Z\sim\mathcal{F}}[Z^p]\right] \geq 1-\delta.$$

For clarity of exposition, we defer the proof of this lemma to Section 8.12.1.

Unfortunately, *a priori* this argument says nothing about estimating the minimum variance. It is easy to see that two mixtures of univariate zero-mean Gaussians D_1, D_2 can have very similar p-th moments but wildly different minimum variances (e.g. take D_1 to be a single Gaussian $\mathcal{N}(0, k^{100})$, and take D_2 to be a uniform mixture of $\mathcal{N}(0, k^{100})$ and $\mathcal{N}(0, 1)$).

The key insight is that while higher-degree moments of a mixture D of zero-mean Gaussians tell us nothing about the minimum variance, those of its *Fourier transform* do. The reason is because of the following observation:

Observation 8.3.2. If the components of D have variances $s_1^2, ..., s_k^2$ and mixing weights $p_1, ..., p_k$, the *Fourier transform* of the density of D is a new (unnormalized) mixture of Gaussians with variances $\Theta(s_1^{-2}), \cdots, \Theta(s_k^{-2})$ and mixing weights proportional to $p_1/s_1, ..., p_k/s_1$ (see Fact 1.3.12).

In particular, if we have a sufficiently good estimate of the maximum variance of any component in the Fourier transform of D, then by inverting this estimate, we can estimate the minimum variance of any component in D. So if we had access to the Fourier transform of D, we could then use the moments of this distribution to estimate the maximum variance of any component of the Fourier transform, which would allow us to learn the minimum variance of D.

What remains is to estimate moments of the Fourier transform of D using solely samples from D. Here we use existing primitives for univariate density estimation [CDSS14b, ADLS17] to obtain an explicit approximation $\widetilde{D}$ to the density of D, after which we can explicitly compute moments of the Fourier transform of $\widetilde{D}$. We defer the technical details of how to argue that its moments are close to those of the Fourier transform of D to Section 8.5.1, as they are rather involved. In short, this allows us to achieve the same sorts of guarantees for estimating min-variance as for estimating max-variance: for any $\kappa > 0$, we can learn the minimum variance of the mixture to multiplicative error $1 + O(\kappa)$ with $2^{\widetilde{O}(\kappa^{-1} \log k)}$ samples and time.

However, there is an important subtlety here. Namely, the quality of the approximation to the minimum variance we can obtain strongly depends on the degree of the Fourier moment we use. The degree in turn dictates the sample complexity of the algorithm: the higher the Fourier degree we need, the better the univariate density estimate must be, and therefore, the more samples we need in order to adequately perform the density estimate. Therefore, if the difference between σ_t and σ'_t is too small, we cannot reliably check if we've made progress without taking too many samples. Unfortunately, the difference between these two quantities is typically quite small. Since a'_{t+1} is a random perturbation of a_t, we have that with probability $1/\text{poly}(k)$, it holds that

$$\sigma'_t \leq \left(1 - \Omega\left(\frac{1}{k}\right)\right) \sigma_t \ ,$$

and moreover, this is tight. In particular, this says that we would need to take $\kappa = \Omega(1/k)$ in the discussion above, which would result in a $2^{\Theta(k)}$ runtime, which we wish to avoid.

However, we show that with subexponentially large probability, the difference is sufficiently large so that we can detect this difference using subexponentially many samples. In particular, observe that for any constant $0 < c < 1/2$, if z is a random unit vector in the span of $\{w_i - a_t\}$, then with probability $\exp(-k^{1-2c})$ we have that

$$\left\langle z, \frac{w_i - a_t}{\|w_i - a_t\|_2} \right\rangle \geq \Omega(k^{-c}), \tag{8.2}$$

in which case if we define $a'_{(t+1)}$ as previously, we get that with probability $\exp(-k^{1-2c})$,

$$\sigma'_t \leq \left(1 - \Omega\Big(\frac{1}{k^{2c}}\Big)\right) \sigma_t \ .$$

So by trying super-polynomially many random directions z at every step, we ensure with high probability that one of those directions will make $1 - \Omega(k^{-2c})$ progress, for some $c < 1$. By combining this with our certification procedure as described above, we show that we can make non-trivial progress in the algorithm after only sub-exponentially many samples.

By iteratively applying this update, we are able to obtain an a_T so that

$$\|a_T - w_i\|_2 \leq \frac{\Delta}{k^{100}} \ ,$$

in subexponential time. We call this subroutine *Fourier moment descent*, and it allows us to learn a single regressor to good accuracy. For technical reasons, the complexity of this approach grows as we get closer to w_i, however, it allows us to obtain a very good “warm start”. In the noiseless case, this can be combined with the boosting procedure from [LL18] to obtain arbitrarily high accuracy.

This technology now allows us to learn a single regressor to very high accuracy. In the noiseless setting, this allows us to “peel off” the samples from this component almost completely, and we can now repeat this process on the sub-mixture with this component removed to learn another component, and iterate to eventually learn all of the regressors.

That said, as we shall see, this is much trickier in the presence of noise.

8.3.2 Learning With Regression Noise

What changes when we assume that there is a significant amount of noise ς? From the perspective of our Fourier moment descent algorithm, it turns out not much does, at least to a certain extent: in fact, essentially the same argument goes through and allows us to learn a single component to error at most

$$\|a_T - w_i\|_2 \leq \frac{\Delta}{k^{100}} + O(\varsigma) \ ,$$

where ς is the standard deviation of the white noise.

However, learning *all* components becomes substantially more difficult. In particular, the peeling process no longer works: the fact that there is regression noise does not allow us to perfectly remove the influence of a component that we have learned from the rest of the mixture. As a result, it is no longer clear how to go from an algorithm that can learn a single component to one that can learn all components.

To avoid this, we circumvent the need for peeling altogether. By a delicate analysis, we will show that with decent probability, we can control the dynamics of the Fourier moment descent algorithm, so that it will converge to the regressor that it was initially closest to. This is the key technical ingredient behind getting Fourier moment descent to handle regression noise.

We sketch its proof below. Let a_t be the current iterate of Fourier moment descent, and suppose that $i^* = \mathrm{argmin}_{i \in [k]} \|w_i - a_t\|_2$. Then, one can show that if $a'_{t+1} \in \mathbb{R}^d$ is a random perturbation of a_t, then with probability at least $\exp(-\Omega(1/\Delta^2))/\mathrm{poly}(k)$, we have that i^* is still the closest component to a'_{t+1}. Moreover, this bound is tight up to polynomial factors, if all we assume about the w_i is that they are Δ-separated. One could hope that by using this sort of argument, we could argue that with at least subexponentially large probability, we always stay closest to w_{i^*}. However, this argument runs into a couple of difficulties, which we address one at a time.

The first difficulty is that while this happens with decent probability for any individual $a'_{t+1} \in \mathbb{R}^d$, our algorithm will typically need to try sub-exponentially many perturbations before we find one that makes progress. Thus the naive union bound over all sub-exponentially many $a'_{t+1} \in \mathbb{R}^d$ would be far too loose to say anything here. To get around this, we instead demonstrate that conditioning on the event that $\sigma'_t < \sigma_t$, we remain closest to i^* with non-trivial probability.

However, even this is insufficient, as we only have a multiplicative estimate of σ_t and σ'_t. To get around this, we make a stronger assumption: we assume that not only is our iterate the closest to $w_{i^*} \in \mathbb{R}^d$, but we also assume that its distance to the other $w_i \in \mathbb{R}^d$ for $i \neq i^*$ is at least some multiplicative factor larger than its distance to $w_{i^*} \in \mathbb{R}^d$. Specifically, we assume that

$$\|w_i - a_t\|_2 \geq \left(1 + c\frac{\Delta^2}{\sqrt{k}}\right) \|w_{i^*} - a_t\|_2 \ , \tag{8.3}$$

for all $i \neq i^*$, and some constant $c > 0$. By making this stronger assumption, we are able to demonstrate that with probability at least $1/\mathrm{poly}(k)$, this gap is maintained for the a'_{t+1} which Fourier moment descent chooses as the next iterate.

Our overall algorithm then will be to demonstrate an initialization scheme for a_0 so that for any $i^* \in [k]$, (8.3) holds for a_0 and that i^* with probability at least $\exp(-\widetilde{O}(\sqrt{k}/\Delta^2))$. If we can do so, then if we run Fourier moment descent starting from these random initializations enough times, then eventually with high probability we will output multiple estimates for each w_i for $i = 1, \ldots, k$, and then we can simply run a basic clustering algorithm to recover all of the regressors. Furthermore, because each run of moment descent only needs to go on for $\widetilde{O}(\sqrt{k}/\Delta^2)$ iterations, and at each iteration we stay closest to the same component that the previous iterate was closest to with $1/\mathrm{poly}(k)$ probability, it suffices to try $\exp(\widetilde{O}(\sqrt{k}/\Delta^2))$ random initializations for all this to work.

Lastly, it turns out this initialization scheme is also quite delicate. For instance, if we simply chose random initializations over the unit ball, then these will, with overwhelming probability, favor being close to w_i with small norm. If we have regressors with different norms, we will thus likely never be close to a $w_i \in \mathbb{R}^d$ with large norm in our initializations, and as a result, cannot argue that our Fourier moment descent algorithm will ever recover this regressor. To get around this, we demonstrate a gridding scheme, where we initialize randomly over spheres of up to radius r, for r on a fine grid between 0 and $O(k^{1/4})$.

We emphasize that this is quite counterintuitive: when we start at radius $O(k^{1/4})$, we are exceedingly far away from every $w_i \in \mathbb{R}^d$, as they are assumed to have norm at most 1. However, our analysis of Fourier moment descent works fine in this setting, and by having such large radius, we are able to ensure that for each $i^* \in [k]$, (8.3) holds for a_0 and i^* with at least the desired probability. The details of this are quite technically involved, and we defer them to Section 8.7.

More Noise-Robust Boosting As mentioned previously, in the noiseless setting, the boosting algorithm of [LL18] allows us to bootstrap our warm start, obtained via Fourier moment descent, to arbitrarily high accuracy. It turns out that in the noisy setting, their boosting algorithm also allows one to go slightly below $O(\varsigma)$. Interestingly, motivated again by the connection to Fourier analysis, we demonstrate an improved boosting algorithm that is able to tolerate substantially more noise as well as a much weaker warm start.

The boosting algorithm of [LL18] is based on stochastic gradient descent on a regularized form of gravitational potential, which was notably used in [HPZ18]. While this objective is *concave*, they demonstrate that in a small neighborhood around the true regressors, SGD updates based on this objective contract in expectation, and hence they make progress.

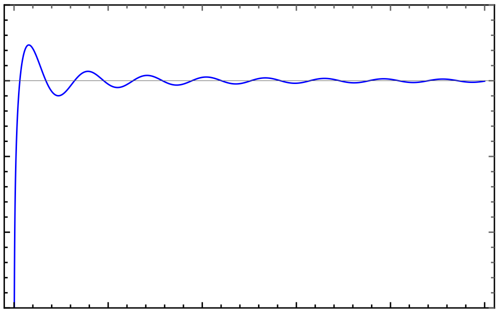


Figure 8-1: Cosine integral function $\mathrm{Ci}(x) = -\int_x^\infty \frac{\cos(t)}{t}\,\mathrm{d}t$

In contrast, we propose an update based on a regularized form of the *cosine integral objective* (see Figure 8-1 for a plot of the cosine integral — the objective function we use is a regularized version of the objective $g(v) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathrm{Ci}(|\langle v, x\rangle - y|))$. This objective looks much worse behaved: it is neither convex nor concave—indeed, it is not even monotone! However, the key technical fact which makes this objective more noise tolerant is precisely Observation 8.3.2. This allows us to argue that the contribution to the gradient of the objective from the "good" component, which we are close to, dominates the contribution of the gradient from the other "bad" components. At a high level, ignoring many technical issues for the time being, this is because if v is the current iterate, a main part of the contribution to the gradient from component i is $\mathbb{E}_{g\sim\mathcal{N}(0,\beta_i^2)}[\cos(g)]$, where β_i is a monotone function of $\|w_i - v\|_2$. This is exactly the real part of the Fourier transform of $\mathcal{N}(0,\beta_i^2)$, and the fact that this decreases as $\exp(O(\beta_i^{-2}))$ follows precisely from Observation 8.3.2. This allows us to have much finer control over the contribution from the "bad" components, which allows us to tolerate significantly more noise and a substantially weaker warm start. For the quantitative details, see Section 8.9.

8.3.3 Learning Mixtures of Hyperplanes

As we will see, mixtures of hyperplanes share enough qualitative features with MLRs that an appropriate instantiation of our techniques also suffices for this problem.

It is not hard to show that vanilla moment descent can be modified as follows to get a $\widetilde{O}(d\cdot\exp(\widetilde{O}(k)))$-time algorithm for mixtures of hyperplanes. First it is not hard to see that as with spherical MLRs, here we can still effectively reduce the dimension of the problem from d to k. At time t we still maintain an estimate a_t for one of the components, and in lieu of the usual progress measure $\min_{i\in[k]}\|w_i - a_t\|_2$ to which we no longer have access, we can use the modified progress measure $\sigma_t \triangleq \min_{i\in[k]}\|\Pi_i a_t\|_2$. (See Definition 8.2.2 for definition of $\Pi_i \in \mathbb{R}^{d\times d}$) We can estimate σ_t in $\exp(\widetilde{O}(k))$ time by simply projecting in the direction of a_t itself to get a mixture of univariate Gaussians with variances $\{\|\Pi_i a_t\|_2^2\}_{i\in[k]}$ and learning that mixture via [MV10]. This leads to the following straightforward modification of moment descent: repeatedly update $a_t \in \mathbb{R}^d$ by sampling many random steps $\{\eta_t \cdot z_j\}$ for step size η_t and $z_j \in \mathbb{S}^{d-1}$ and taking one of them to get to $a_{t+1}\in\mathbb{R}^d$ if it contracts the progress measure.

One immediate issue with this approach, even for achieving $\widetilde{O}(d)\cdot\exp(\widetilde{O}(k))$ runtime, is that if we don't control $\|a_t\|_2$, then for all we know σ_t^2 contracts simply because our random steps are taking us closer and closer to 0. The natural workaround is to insist $a_t \in \mathbb{S}^{d-1}$ for all t by projecting onto $\mathbb{S}^{d-1}$ after every step. That is, in every iteration of moment descent, given a random collection of candidate steps $\{\eta\cdot z_j\}_j$, choose one for which, if we define

$$a_{t+1} = \frac{a_t + \eta\cdot z_j}{\|a_t + \eta\cdot z_j\|_2},$$

then the progress measure contracts. One can show this already suffices to get a $\widetilde{O}(d)\,\cdot$

$\exp(\widetilde{O}(k))$-time algorithm for learning mixtures of hyperplanes.

The extra projection onto $\mathbb{S}^{d-1}$ at every step introduces a variety of technical challenges for trying to implement the strategy of Section 8.3.1 to achieve sub-exponential runtime. Specifically, in order to carry out the balancing of parameters from Section 8.3.1, one needs to be careful in choosing the right events, analogous to the event in (8.2), such that 1) conditioned on those events σ_t contracts by at least a factor of $1-\Omega(k^{-a})$ for some $a>0$, and 2) these events all occur with probability $\exp(-k^{1-a'})$ for some $a'>0$.

If for instance a_t is positively correlated with some $v_{i^*}\in\mathbb{R}^d$, it turns out the right events to choose are that the random step is both $k^{-1/5}$-positively correlated with $\frac{\Pi_{i^*}a_t}{\|\Pi_{i^*}a_t\|_2}$ and at least $k^{-1/5}$-negatively correlated with $v_{i^*}\in\mathbb{R}^d$, and because these directions are orthogonal, if z_j is a random vector in the span of $v_1,...,v_k$ then we could lower bound the probability of both events occurring by the product of the probabilities they individually occur, which is $\exp(-k^{3/5})$, and get 2) for $a'=2/5$. The analysis for showing 1) (for $a=3/5$) is involved, so we defer it to Section 8.8.1. These together yield a $\widetilde{O}(d)\cdot\exp(k^{3/5})$-time algorithm to learning one component of a mixture of hyperplanes.

To learn all components, we would like to implement some kind of boosting procedure. Our approach here is to regard $\mathcal{D}$ in a certain way as a non-spherical MLR with well-conditioned covariances, at which point we can invoke, e.g., the boosting algorithm of [LL18]. We defer these details to Section 8.8.4. Once we are able to refine an estimate for a direction of $\mathcal{D}$ to arbitrary precision, we can carry out the "peeling" procedure outlined at the end of Section 8.3.1 to learn all components.

8.4 Roadmap

Here we give a brief overview of the organization of the rest of the paper. In Section 8.5 we present our Fourier moment descent algorithm for learning a single component. In Section 8.6 we show how to use this to learn all the components, when there is no regression noise. In Section 8.7 we demonstrate a modification of our algorithm to learn all the components in the presence of regression noise. In Section 8.8 we demonstrate our subexponential time algorithm for learning a mixture of hyperplanes. Finally, in Section 8.9 we demonstrate our improved boosting algorithm based on the cosine integral objective. Deferred proofs appear in the Appendix, as well as our moment-matching example.

8.5 Warm Start via Fourier Moment Descent

Here we propose a technique for moment descent based on approximating the minimum variance of a component in a mixture of univariate zero-mean Gaussians. The main result of this section is an algorithm, which we call FourierMomentDescent, for learning a single component of a mixture of k linear regressions in time and sample complexity subexponential in k:

Theorem 8.5.1 (Fourier moment descent). *Given $\delta, \varepsilon > 0$ and a mixture of spherical linear regressions $\mathcal{D}$ with separation Δ and noise rate $\varsigma = O(\varepsilon)$, there is an algorithm (*FourierMomentDesc ent*$(\mathcal{D}, \delta, \varepsilon)$ inAlgorithm 31) that outputs a vector $v \in \mathbb{R}^d$ such that with probability $1 - \delta$, we have $\|w_i - v\|_2 \leq O(\varepsilon)$ for some $i \in [k]$. Furthermore,* FourierMomentDescent *requires sample complexity*

$$N = \widetilde{O}\left(d\varepsilon^{-2}\ln(1/\delta)p_{\min}^{-4} \cdot \mathrm{poly}\left(k, \ln(1/p_{\min}), \ln(1/\varepsilon)\right)^{O(\sqrt{k}\ln(1/p_{\min}))}\right)$$

and time complexity $Nd \cdot \mathrm{poly}\log(k, d, 1/\Delta, 1/p_{\min}, 1/\varepsilon)$.

In Section 8.5.1 we give an algorithm for estimating the minimum variance of a mixture of univariate, zero-mean Gaussians via its Fourier transform. In Section 8.5.2 we show how to leverage this technology to obtain our algorithm FourierMomentDescent and then give a proof of Theorem 8.5.1.

8.5.1 Estimating Minimum Variance

Here we give the key primitive underlying all of the algorithmic results of this chapter: an algorithm for estimating the minimum variance of a mixture of zero-mean Gaussians. This requires some setup regarding existing technology for density estimation.

Density Estimation in L_2

Our main density estimation tool will be to use piecewise polynomials. We favor them because there are clean algorithms for density estimation via piecewise polynomials, and moreover, the form of the estimator will be useful for us later on. Formally:

Definition 8.5.2. *An s-piecewise degree-d polynomial $p : \mathbb{R} \to \mathbb{R}$ is specified by a collection of intervals $I_1 = [-\infty, a_1], I_2 = [a_1, a_2], I_3 = [a_2, a_3], ..., I_{s-1} = [a_{s-2}, a_{s-1}], I_s = [a_{s-1}, +\infty]$ and s degree-d polynomials $p_1, ..., p_s$ such that for any $i \in [s]$ and $x \in I_s$, $p(x) = p_i(x)$. We refer to $a_1, ..., a_{s-1}$ as the* nodes *of p.*

We will use the following algorithm as a black box:

Theorem 8.5.3 (Theorem 43 in [ADLS17]). *For any $\eta > 0$, there is an algorithm that, given sample access to a mixture $\mathcal{F}$ of k univariate Gaussians, outputs a $O(k)$-piecewise degree-$O(\log 1/\eta)$ polynomial hypothesis distribution $\mathcal{F}'$ for which $d_{TV}(\mathcal{F}, \mathcal{F}') \le \eta$, using $N = O\left((k/\eta^2)\ln(1/\eta)\ln(1/\delta)\right)$ samples and running in time $\widetilde{O}(N)$.*

```
Algorithm 28: L2Estimate($\mathcal{F}, \underline{\sigma}, \eta, \delta$)
  Input: Sample access to univariate $k$-GMM $\mathcal{F}$, a number $\underline{\sigma}$ for which $\underline{\sigma} \le \sigma_{\min}(\mathcal{F})$,
         precision parameter $\eta > 0$, failure probability $\delta > 0$
  Output: Piecewise polynomial function $\mathcal{G}$ for which $\|\mathcal{F} - \mathcal{G}\|_2^2 \le \eta$
  $N \leftarrow \Theta\left(\frac{k}{2\pi\underline{\sigma}^2\eta^2}\log\left(\frac{1}{\sqrt{2\pi}\underline{\sigma}\eta}\right)\right)$.
  Draw $N$ samples from $\mathcal{F}$ and run the algorithm from Theorem 8.5.3 to obtain an
    estimate $\mathcal{F}'$ for which $d_{\mathrm{TV}}(\mathcal{F}, \mathcal{F}') \le \sqrt{2\pi}\underline{\sigma} \cdot \eta$.
  For each $x \in \mathbb{R}$, define $\mathcal{G}(x) = \min\left\{\max\{\mathcal{F}'(x), 0\}, \frac{1}{\sqrt{2\pi}\underline{\sigma}}\right\}$.
  return $\mathcal{G}$.
```

Corollary 8.5.4 (Guarantee for L2Estimate). *For any $0 < \eta, \delta < 1$, mixture of k univariate Gaussians $\mathcal{F}$, and $\underline{\sigma} > 0$ for which $\underline{\sigma} \le \sigma_{\min}(\mathcal{F})$, with probability at least $1 - \delta$* L2Estimate*$(\mathcal{F}, \underline{\sigma}, \eta, \delta)$ (Algorithm 28) outputs a $O(k\log(1/(\eta\underline{\sigma})))$-piece degree-$O(\log(1/(\eta\underline{\sigma}))$ polynomial hypothesis distribution $\mathcal{G} : \mathbb{R} \to \mathbb{R}_{\ge 0}$ for which $\|\mathcal{F} - \mathcal{G}\|_2^2 \le \eta$, using $N = O\left((k/\eta^2)\log(1/(\eta\underline{\sigma}))\log(1/\delta)\right)$ samples and running in time $\widetilde{O}(N)$.*

Proof. Because $\mathcal{F}$ has range in $\left[0, \frac{1}{\sqrt{2\pi}\sigma_{\min}(\mathcal{F})}\right]$, by construction we have that $d_{\mathrm{TV}}(\mathcal{F}, \mathcal{G}) \le d_{\mathrm{TV}}(\mathcal{F}, \mathcal{F}')$. By Holder's, it suffices to show $\|\mathcal{F} - \mathcal{G}\|_\infty \le \frac{1}{\sqrt{2\pi}\cdot\underline{\sigma}}$. But because $\mathcal{G}(x) \le \mathcal{F}(x)$

for all $x \in \mathbb{R}$, it is enough to show $\|\mathcal{F}\|_\infty \leq \frac{1}{\sqrt{2\pi}\cdot\underline{\sigma}}$, which just follows from the fact that $\mathcal{F}$ is a convex combination of Gaussians of variance at least $\sigma_{\min}(\mathcal{F}) \geq \underline{\sigma}$. Note that $\mathcal{G}$ is a piecewise polynomial because we can refine the intervals defining $\mathcal{F}'$ to incorporate the intersections of $\mathcal{F}'$ with the lines $y = \frac{1}{\sqrt{2\pi}\underline{\sigma}}$ and $y = 0$. Since each individual component can intersect with these lines at most $O(\log(1/(\eta\underline{\sigma})))$ times since they have degree at most $O(\log(1/(\eta\underline{\sigma})))$, this yields the desired bound on the number of pieces of the resulting piecewise polynomial estimate. □

Minimum Variance Via Fourier Transform Moments

We now show how to use an L_2-close estimator for the density of a mixture $\mathcal{F}$ of zero-mean univariate Gaussians to approximate $\sigma_{\min}(\mathcal{F})$. As a first step, we show how to use an L_2-close estimator to estimate high moments of $\mathcal{F}$:

Lemma 8.5.5. *For any even integer $p \in \mathbb{N}$ and $\xi > 0$ the following holds. Let $\mathcal{F} : \mathbb{R} \to \mathbb{R}$ be an mixture of k Gaussians given by*

$$\mathcal{F}(x) = \sum_{i=1}^{k} p_i \cdot \mathcal{N}(0, \sigma_i^2; x),$$

and define $L \triangleq \sum_{i=1}^{k} p_i$. Let $\overline{\sigma} > 0$ be any number for which $\overline{\sigma} \geq \max_{i\in[k]} \sigma_i$.

Let $\tau = 8\overline{\sigma}^2 \cdot \max(p, \ln(4L/\xi))$ and $\eta = \frac{\xi^2 p}{8\tau^{2p+1}}$. Then if function $\mathcal{G} : \mathbb{R} \to \mathbb{R}_{\geq 0}$ satisfies $\|\mathcal{F} - \mathcal{G}\|_2^2 \leq \eta$, then $\mathcal{G}'$ defined by $\mathcal{G}'(x) = \mathbb{1}[x \in [-\tau, \tau]] \cdot \mathcal{G}(x)$ for all $x \in \mathbb{R}$ satisfies

$$|\mathcal{M}_p(\mathcal{F}) - \mathcal{M}_p(\mathcal{G}')| \leq \xi. \tag{8.4}$$

Proof. For simplicity, we define

$$\sigma_{\max} = \sigma_{\max}(\mathcal{F}).$$

We would like to pick the truncation threshold τ so that

$$\int_{[-\tau,\tau]^c} x^p \cdot \mathcal{F}(x)\, \mathrm{d}x \leq \xi/2. \tag{8.5}$$

For this, it suffices to take τ for which

$$x^p \leq e^{x^2/(4\sigma_{\max}^2)} \cdot \frac{\xi}{4L}, \quad \forall x \notin [-\tau, \tau], \tag{8.6}$$

in which case

$$\begin{aligned}
\int_{[-\tau,\tau]^c} x^p \cdot \mathcal{F}(x) \,\mathrm{d}x = & \sum_{i=1}^{k} p_i \int_{[-\tau,\tau]^c} x^p \cdot \frac{1}{\sqrt{2\pi}\sigma_i} \cdot e^{-x^2/(2\sigma_i^2)} \,\mathrm{d}x \\
& \leq L \cdot \int_{[-\tau,\tau]^c} x^p \cdot \frac{1}{\sqrt{2\pi}\sigma_{\max}} \cdot e^{-x^2/(2\sigma_{\max}^2)} \,\mathrm{d}x \\
& \leq L \cdot \int_{[-\tau,\tau]^c} \frac{1}{\sqrt{2\pi}\sigma_{\max}} \cdot e^{-x^2/(4\sigma_{\max}^2)} \cdot \frac{\xi}{4L} \,\mathrm{d}x \\
& \leq \xi/2
\end{aligned}$$

where the first step follows from definition of $\mathcal{F}(x)$, the second step follows from Fact 8.2.3, and the third step follows from Eq. (8.6).

To reach (8.6), we want

$$p \ln x \leq \frac{x^2}{4\sigma_{\max}^2} - \ln(4L/\xi), \quad \forall x \notin [-\tau, \tau], \tag{8.7}$$

For $x \geq 8\sigma_{\max}^2 \ln(4L/\xi)$, we get that

$$\frac{x^2}{4\sigma_{\max}^2} - \ln(4L/\xi) \geq \frac{x^2}{8\sigma_{\max}^2},$$

and for $x \geq 8p\sigma_{\max}^2$, we have that

$$p \ln x \leq \frac{x^2}{8\sigma_{\max}^2}.$$

We conclude that for $\tau = 8\overline{\sigma}^2 \cdot \max\{p, \ln(4L/\xi)\}$, Eq. (8.7) holds.

We can now complete the proof of (8.4). We may write $|\mathcal{M}_p(\mathcal{F}) - \mathcal{M}_p(\mathcal{G}')|$ as

$$|\mathcal{M}_p(\mathcal{F}) - \mathcal{M}_p(\mathcal{G}')| = \left| \int_{-\infty}^{\infty} x^p \cdot \mathcal{F}(x) \,\mathrm{d}x - \int_{-\tau}^{\tau} x^p \cdot \mathcal{G}(x) \,\mathrm{d}x \right|$$

$$
\begin{aligned}
&\leq \left|\int_{[-\tau,\tau]^c} x^p \cdot \mathcal{F}(x)\,\mathrm{d}x\right| + \left|\int_{-\tau}^{\tau} x^p \cdot (\mathcal{F}-\mathcal{G})(x)\,\mathrm{d}x\right| \\
&\leq \xi/2 + \left(\int_{-\tau}^{\tau} x^{2p}\,\mathrm{d}x\right)^{1/2} \cdot \|\mathcal{F}-\mathcal{G}\|_2 \\
&= \xi/2 + \left(\frac{2\tau^{2p+1}}{2p+1}\eta\right)^{1/2} \leq \xi,
\end{aligned}
$$

where the second step follows from triangle inequality, the third step follows by (8.5) and the last step follows if we take $\eta = \frac{\xi^2 p}{8\tau^{2p+1}}$. □

This is useful as good estimates of high moments of the mixture allow us to approximate the maximum variance of any component well, as components with large variance contribute significantly more to the high moments than do the components with small variance. However, we wish to estimate the minimum variance of our mixture. We now show that we can do so by taking high moments of the *Fourier transform* of our L_2-close estimator. As an important subroutine, we show that it is efficient to compute the Fourier moments of our density estimate, by using the fact that it is piecewise polynomial. Specifically:

Lemma 8.5.6. *Given the description of a s-piece degree-d polynomial $p : \mathbb{R} \to \mathbb{R}$, and any $\tau > 0$ and nonnegative integer $\ell > 0$, there is an algorithm which runs in time $O(s\ell d^3)$ and which outputs*

$$
\int_{-\tau}^{\tau} \widehat{p}[\omega]\omega^{\ell}\mathrm{d}\omega \, .
$$

We defer the description of this algorithm as well as the proof of correctness to Appendix 8.11. With this primitive, we can now show:

```
Algorithm 29: EstimateMinVariance($\mathcal{F}, \overline{\sigma}, \underline{\sigma}, p, \delta$)
  Input: Sample access to mixture of $k$ univariate Gaussians $\mathcal{F}$, numbers $\overline{\sigma}, \underline{\sigma}$ for
         which $\overline{\sigma} \geq \sigma_{\max}(\mathcal{F})$ and $\underline{\sigma} \leq \sigma_{\min}(\mathcal{F})$, degree $p \in \mathbb{N}$, failure probability $\delta > 0$
  Output: Estimate $\sigma^*$ for which (8.8) holds
  $\xi \leftarrow (2\pi)^{-p-1/2} p^{p/2} p_{\min} \overline{\sigma}^{-p-1}$.
  $L \leftarrow \sqrt{2\pi} \cdot \underline{\sigma}^{-1}$ $\tau \leftarrow 8\overline{\sigma}^2 \cdot \max(p, \ln(2L\sqrt{2}/\xi))$ .
  $\eta \leftarrow \frac{\xi^2 p}{8\tau^{2p+1}}$.
  $\mathcal{G} \leftarrow$ L2Estimate($\mathcal{F}, \underline{\sigma}, \eta, \delta$).                // Algorithm 28, Corollary 8.5.4
  Explicitly compute $\mathcal{M}_p(\widehat{\mathcal{G}})$ using Lemma 8.5.6.
  return $\sigma^* \triangleq \left(\frac{\mathcal{M}_p(\widehat{\mathcal{G}})}{(2\pi)^{-p-1/2} p^{p/2}}\right)^{-1/(p-1)}$.
```

```
Algorithm 30: CompareMinVariances($\mathcal{F}_1, \mathcal{F}_2, \overline{\sigma}, \underline{\sigma}, \kappa_1, \kappa_2, \delta$)
  Input: Sample access to two mixtures of $k$ univariate Gaussians $\mathcal{F}_1, \mathcal{F}_2$, numbers
         $\overline{\sigma}, \underline{\sigma}$ for which $\overline{\sigma} \geq \max(\sigma_{\max}(\mathcal{F}_1), \sigma_{\max}(\mathcal{F}_2))$ and
         $\underline{\sigma} \leq \min(\sigma_{\min}(\mathcal{F}_1), \sigma_{\min}(\mathcal{F}_2))$, tolerance parameters $\kappa_1 < \kappa_2$, failure
         probability $\delta > 0$
  Output: If the output is True, then $\sigma_{\min}(\mathcal{F}_1) \geq (1+\kappa_1)\sigma_{\min}(\mathcal{F}_2)$, otherwise
          $\sigma_{\min}(\mathcal{F}_1) \leq (1+\kappa_2)\sigma_{\min}(\mathcal{F}_2)$
  $p \leftarrow \Omega\left(\frac{\ln(1/p_{\min})}{\kappa_2 - \kappa_1}\right)$.
  $\sigma^*_j \leftarrow$ EstimateMinVariance($\mathcal{F}_j, \overline{\sigma}, \underline{\sigma}, p, \delta$) for $j = 1, 2$.    // Algorithm 29,
    Lemma 8.5.7
  if $\frac{\mathcal{M}_p(\widehat{\mathcal{G}}_1)}{\mathcal{M}_p(\widehat{\mathcal{G}}_2)} > \frac{1}{2} p_{\min}(1+\kappa_2)^{p-1}$ then
      return True.
  else
      return False.
```

Lemma 8.5.7 (Guarantee for EstimateMinVariance)**.** *Let $\mathcal{F}$ be a mixture of k univariate zero-mean Gaussians with parameters $(\{p_i\}_{i\in[k]}, \{\sigma_i\}_{i\in[k]})$. Let $\overline{\sigma} \geq \sigma_{\max}(\mathcal{F})$, $\underline{\sigma} \leq \sigma_{\min}(\mathcal{F})$. Then with probability at least $1-\delta$,* EstimateMinVariance*$(p, \mathcal{F}, \overline{\sigma}, \underline{\sigma}, \delta)$ (Algorithm 29) takes*

$$N = p_{\min}^{-4} k \ln(1/\delta) \cdot \mathrm{poly}\left(\overline{\sigma}, p, \ln(1/p_{\min}), \ln(1/\underline{\sigma})\right)^{O(p)}$$

samples, runs in time $\widetilde{O}(N)$, and outputs a number σ^* for which

$$\left(\frac{3}{4}\right)^{1/(p-1)} \cdot \sigma_{\min}(\mathcal{F}) \leq \sigma^* \leq \left(\frac{3}{2p_{\min}}\right)^{1/(p-1)} \cdot \sigma_{\min}(\mathcal{F}). \tag{8.8}$$

Proof. Note that in the pseudocode our choice of η is given by

$$\eta \triangleq \frac{\xi^2 p}{8\left(8\overline{\sigma}^2 \max\left(p, \ln\left(\frac{4\sqrt{\pi}}{\underline{\sigma}\xi}\right)\right)\right)^{2p+1}}, \quad \xi \triangleq (2\pi)^{-p-1/2} p^{p/2} p_{\min} \overline{\sigma}^{-p-1}\,. \tag{8.9}$$

Consequently, the runtime and sample complexity bounds for general p just follow from the fact that these quantities are dominated by the cost of running L2Estimate$(\mathcal{F}, \underline{\sigma}, \eta, \delta)$ for η as defined in (8.9). By Corollary 8.5.4, if we run L2Estimate$(\mathcal{F}, \underline{\sigma}, \eta, \delta)$ and produce the piecewise polynomial $\mathcal{G}$, we know that $\|\mathcal{F} - \mathcal{G}\|_2^2 \leq \eta$. By Plancherel's, $\|\widehat{\mathcal{F}} - \widehat{\mathcal{G}}\|_2^2 \leq \eta$. To apply Lemma 8.5.5, first note that by Fact 1.3.12,

$$\widehat{\mathcal{F}}(\omega) = \sum_{i=1}^{k} p_i \frac{1}{\sqrt{2\pi}\sigma_i} \mathcal{N}\left(0, \frac{1}{4\pi^2\sigma_i^2}, \omega\right).$$

So $\widehat{\mathcal{F}}$ is an affine linear combination of Gaussian densities, and its coefficients sum to

$$\sum_{i=1}^{k} p_i \cdot \frac{1}{\sqrt{2\pi}\sigma_i} \leq \frac{1}{\sqrt{2\pi}\sigma_{\min}(\mathcal{F})} \leq L,$$

where the last step follows by our choice of L in EstimateMinVariance.

So by Lemma 8.5.5, if we define $\widehat{\mathcal{G}}'$ by

$$\widehat{\mathcal{G}}'(x) = \mathbb{1}[x \in [-\tau, \tau]] \cdot \widehat{\mathcal{G}}(x),$$

then we get that

$$|\mathcal{M}_p(\widehat{\mathcal{F}}) - \mathcal{M}_p(\widehat{\mathcal{G}}')| \leq \xi. \tag{8.10}$$

Furthermore, note that

$$\begin{aligned}\mathcal{M}_p(\widehat{F}) &\leq \sum_{i=1}^{k} p_i \cdot \frac{1}{\sqrt{2\pi}\sigma_i} \cdot p^{p/2} \cdot \left(\frac{1}{4\pi^2(\sigma_i)^2}\right)^{p/2} \\ &= (2\pi)^{-p-1/2} p^{p/2} \sum_{i=1}^{k} p_i \left(\sigma_i\right)^{-p-1}. \end{aligned} \tag{8.11}$$

If we had $\xi = (2\pi)^{-p-1/2}p^{p/2}\xi'$ for some $\xi' > 0$, then we get by (8.10) and (8.11) that

$$\mathcal{M}_p(\widehat{\mathcal{G}}) = (2\pi)^{-p-1/2} p^{p/2} \left[\sum_{i=1}^{k} p_i \left(\sigma_i\right)^{-p-1} \pm \xi'\right].$$

If we take $\xi' \triangleq \frac{1}{3} p_{\min} \overline{\sigma}^{-p-1}$, then observe that because

$$p_{\min} \cdot \sigma_{\min}(\mathcal{F})^{-p-1} \leq \sum_{i=1}^{k} p_i (\sigma_i)^{-p-1} \leq \sigma_{\min}(\mathcal{F})^{-p-1},$$

we have

$$\sigma_{\min}(\mathcal{F}) \leq \left(\sum_{i=1}^{k} p_i(\sigma_i)^{-p-1} \right)^{-1/(p-1)} \leq p_{\min}^{-1/(p-1)} \cdot \sigma_{\min}(\mathcal{F}),$$

so $\sigma^* \triangleq \left(\frac{\mathcal{M}_p(\widehat{\mathcal{G}})}{(2\pi)^{-p-1/2} p^{p/2}} \right)^{-1/(p-1)}$ satisfies (8.8).

For the last part of the lemma, take $p = 20 \ln \left(\frac{3}{2p_{\min}} \right) + 1 \geq 4$. It is straightforward to check the sample and time complexity bounds for this choice of p, and the bound on σ^* follows from the fact that $(3/4)^{1/(p-1)} \geq 0.9$ for $p \geq 4$ and

$$\left(\frac{3}{2p_{\min}} \right)^{\left(20 \ln\left(\frac{3}{2p_{\min}}\right)\right)^{-1}} = e^{1/20} \leq 1.1.$$

□

We now identify two specific parameter settings for this algorithm which will be useful later on. First, if we take the degree p to be relatively small, we are able to get a constant approximation to the minimum variance very efficiently:

Corollary 8.5.8. *Let $p = \Theta(\ln(1/p_{\min}))$. Then, the algorithm* EstimateMinVariance*$(p, \mathcal{F}, \overline{\sigma}, \underline{\sigma}, \delta)$ has sample and time complexity*

$$\widetilde{O}\left(p_{\min}^{-4} k \ln(1/\delta) \cdot \mathrm{poly}\left(\overline{\sigma}, p, \ln(1/p_{\min}), \ln(1/\underline{\sigma})\right)^{O(p)} \right),$$

and the output σ^* satisfies

$$0.9 \cdot \sigma_{\min}(\mathcal{F}) \leq \sigma^* \leq 1.1 \cdot \sigma_{\min}(\mathcal{F}).$$

We also have:

Corollary 8.5.9 (Guarantee for CompareMinVariances)**.** *Let $0 < \kappa_1 < \kappa_2 \leq 1$, and let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two mixtures of k univariate zero-mean Gaussians with parameters $(\{p_i\}, \{\sigma_i^{(1)}\})$ and $(\{p_i\}, \{\sigma_i^{(2)}\})$ respectively. Let $\overline{\sigma} \geq \max(\sigma_{\max}(\mathcal{F}_1), \sigma_{\max}(\mathcal{F}_2))$, $\underline{\sigma} \leq \min(\sigma_{\min}(\mathcal{F}_1), \sigma_{\min}(\mathcal{F}_2))$. Then, with probability $1 - \delta$, the algorithm* CompareMinVariances*$(\mathcal{F}_1, \mathcal{F}_2, \overline{\sigma}, \underline{\sigma}, \kappa_1, \kappa_2, \delta)$ (Algorithm 30) satisfies:*

- *If* $\sigma_{\min}(\mathcal{F}_1) \geq (1+\kappa_2)\,\sigma_{\min}(\mathcal{F}_2)$*, then it outputs* True*.*

- *If* $\sigma_{\min}(\mathcal{F}_1) \leq (1+\kappa_1)\sigma_{\min}(\mathcal{F}_2)$*, then it outputs* False*.*

Moreover, this algorithm takes

$$N = p_{\min}^{-4} k \ln(1/\delta) \cdot \mathrm{poly}\left(\overline{\sigma}, (\kappa_2-\kappa_1)^{-1}, \ln(1/p_{\min}), \ln(1/\underline{\sigma})\right)^{O((\kappa_2-\kappa_1)^{-1}\ln(1/p_{\min}))}$$

samples and runs in time $\widetilde{O}(N)$*.*

Proof. Let σ_j^* be the estimate produced by

$$\textsc{EstimateMinVariance}(p, \mathcal{F}_j, \overline{\sigma}, \underline{\sigma}, \delta), \quad \forall j = 1, 2.$$

By Lemma 8.5.7, we have that

$$\frac{1}{2} \cdot p_{\min} \cdot \left(\frac{\sigma_{\min}(\mathcal{F}_1)}{\sigma_{\min}(\mathcal{F}_2)}\right)^{p-1} \leq \left(\frac{\sigma_1^*}{\sigma_2^*}\right)^{p-1} \leq 2 \cdot p_{\min}^{-1} \cdot \left(\frac{\sigma_{\min}(\mathcal{F}_1)}{\sigma_{\min}(\mathcal{F}_2)}\right)^{p-1}. \tag{8.12}$$

Now for the first part of the lemma, by hypothesis $\frac{\sigma_{\min}(\mathcal{F}_1)}{\sigma_{\min}(\mathcal{F}_2)} \geq 1+\kappa_2$, so by the lower bound in (8.12) we conclude that

$$\left(\frac{\sigma_1^*}{\sigma_2^*}\right)^{p-1} \geq \frac{1}{2} p_{\min} \left(1+\kappa_2\right)^{p-1}.$$

For the second part of the lemma, by hypothesis $\frac{\sigma_{\min}(\mathcal{F}_1)}{\sigma_{\min}(\mathcal{F}_2)} \leq 1+\kappa_1$, so by the upper bound in (8.12) we conclude that

$$\left(\frac{\sigma_1^*}{\sigma_2^*}\right)^{p-1} \leq 2 p_{\min}^{-1} \left(1+\kappa_1\right)^{p-1}.$$

The content of this lemma is that the right-hand side of (6) is strictly greater than the right-hand side of (6). Indeed, we need to check that

$$\left(\frac{1+\kappa_2}{1+\kappa_1}\right)^{p-1} \geq 4 p_{\min}^{-2}.$$

But if we take $p-1=\frac{2\log_2(4/p_{\min}^2)}{\kappa_2-\kappa_1}$, then by the fact that $1+\kappa_1\le 2$ and $\frac{\kappa_2-\kappa_1}{1+\kappa_1}\le 1$, and by the elementary inequality $(1+1/x)^x\ge 2$ for $x\ge 1$, we conclude that

$$\left(\frac{1+\kappa_2}{1+\kappa_1}\right)^{p-1}=\left(1+\frac{\kappa_2-\kappa_1}{1+\kappa_1}\right)^{p-1}\ge 8p_{\min}^{-1}$$

as desired. □

8.5.2 Moment Descent

In this section we will show how to obtain a warm start using the CompareMinVariances subroutine of the previous section.

The first ingredient we need is a subroutine to estimate $\mathrm{span}(\{w_i-a\}_{i\in[k]})$, where a is our current guess for a direction. For any $x,y,a\in\mathbb{R}^d$, define the matrix

$$\mathbf{M}_a^{x,y}\triangleq\frac{1}{2}\left[(y-\langle a,x\rangle)^2xx^\top-(y-\langle a,x\rangle)^2\cdot\mathrm{Id}\right]\qquad\in\mathbb{R}^{d\times d}\tag{8.13}$$

and let

$$\widehat{\mathbf{M}}_a^{(N)}\triangleq\frac{1}{N}\sum_{i=1}^N\mathbf{M}_a^{x_i,y_i}\tag{8.14}$$

for $(x_1,y_1),...,(x_N,y_N)$ i.i.d. samples from $\mathcal{D}$. Notice there is a matrix-vector oracle for $\widehat{\mathbf{M}}_a^{(N)}$ which runs in time $O(Nd)$.

We then have:

Lemma 8.5.10 ($\widehat{\mathbf{M}}_a^{(N)}$ approximates $\mathrm{span}(\{w_i-a_t\})$). *Let $\mathcal{D}$ be a mixture of k spherical linear regressions. Then for any $a\in\mathbb{R}^d$, we have that*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{M}_a^{x,y}\right]=\sum_{i=1}^k p_i(w_i-a)(w_i-a)^\top.$$

Furthermore, for any $\beta,\delta>0$ and

$$N=\widetilde{\Omega}\left(\max_{i\in[k]}\|w_i-a\|_2^2\cdot p_{\min}^{-1}\cdot\beta^{-2}\cdot d\cdot\ln(k/\delta))\right),$$

we have that

$$Pr\left[\|\widehat{\mathbf{M}}_a^{(N)} - \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{M}_a^{x,y}\right]\|_2 \geq \beta\right] \leq \delta$$

We emphasize that $\mathbb{E}[\mathbf{M}_a^{x,y}]$ is the same regardless of η, but the value of η will slightly affect concentration, though not in the regimes in which we will apply Lemma 8.5.10. We defer the proof of this lemma to Appendix 8.12.5. Combining this with Fact 1.3.6 allows us to quantify the effectiveness of approximate k-SVD of an empirical estimate for $\mathbb{E}[\mathbf{M}_a^{x,y}]$ for capturing the span of the w_i:

Lemma 8.5.11 (Correlation of the top principal subspace)**.** *Let*

$$N = \widetilde{\Omega}\left(\frac{\max_{i\in[k]}\|w_i - a\|_2^2}{\min_{i\in[k]}\|w_i - a\|_2^2}\cdot p_{\min}^{-2}\cdot k^2\cdot d\cdot \ln(k/\delta)\right).$$

Then ApproxBlockSVD $(\widehat{\mathbf{M}}^{(N)}, 1/10, \delta/2)$ *runs in time* $\widetilde{O}\left(k\cdot N\cdot d\right)$ *and outputs a matrix* $\mathbf{U}$ *so that with probability at least* $1-\delta$*,*

$$\frac{1}{2} \leq \frac{\|\mathbf{U}^\top(w_i - a)\|_2}{\|w_i - a\|_2} \leq 1. \tag{8.15}$$

Proof. The upper bound is trivial. We now prove the lower bound. We first observe that Lemma 8.5.10 implies that $\mathsf{gap}_k(\widehat{\mathbf{M}}_a^{(N)})^{-1} \geq \mathrm{poly}(1/p_{\min}, 1/\min_{i\in[k]}\|w_i - a\|_2^2)$, which proves the runtime claim. The lower bound follows by taking β in Lemma 8.5.10 to be $\beta = \frac{1}{8k}\cdot p_{\min}^{1/2}\cdot \min_{i\in[k]}\|w_i - a\|_2$ and applying Fact 1.3.6. Finally, to demonstrate the runtime, observe that there is a matrix-vector oracle for $\widehat{\mathbf{M}}_a^{(N)}$ which runs in time $O(Nd)$. □

We are now ready to analyze the amount of progress each step of moment descent makes.

Lemma 8.5.12 (Progress of moment descent per step)**.** *For any* $\delta > \exp(-\sqrt{k})$*, the following holds. Let* $\sigma_t^2 \triangleq \min_{i\in[k]}\|w_i - a_t\|_2$*. Denote the minimizing index* i *by* i^**. For* $M \triangleq e^{\sqrt{k}}\ln(2/\delta)$ *and* $g_1, ..., g_M \sim \mathcal{N}(0, Id_k)$*, let* $v_j = \frac{\mathbf{U}g_i}{\|\mathbf{U}g_i\|_2} \in \mathbb{S}^{d-1}$ *for* $j \in [M]$*. Let* σ^* *be a number for which* $0.9\sigma_t \leq \sigma^* \leq 1.1\sigma_t$*. Let* $\eta = \frac{1}{2}k^{-1/4}\sigma^*$*.*

Then we have that with probability at least $1-\delta$*,*

1. *There exists at least one* $j \in [M]$ *for which* $\|w_{i^*} - a_t - \eta\cdot v_j\|_2^2 \leq \left(1 - \frac{1}{5\sqrt{k}}\right)\sigma_t^2$*.*

2. *For all* $j \in [M]$ *and* $i \in [k]$*,* $\|w_i - a_t - \eta\cdot v_j\|_2^2 \geq \left(1 - \frac{9}{\sqrt{k}}\right)\sigma_t^2$*.*

Proof. For any $i \in [k]$, we may write

$$\begin{aligned}\|w_i - a_t - \eta v_j\|_2^2 &= \|w_i - a_t\|_2^2 + \eta^2 \|v_j\|_2^2 - 2\eta \langle w_i - a_t, v_j \rangle \\ &= \|w_i - a_t\|_2^2 + \eta^2 - 2\eta \langle w_i - a_t, v_j \rangle. \qquad (8.16)\end{aligned}$$

Define $\widetilde{w}_i \triangleq \frac{w_i - a_t}{\|w_i - a_t\|_2}$. For every $j \in [M]$, let A_j be the event that $\langle v_j, \widetilde{w}_{i^*} \rangle \geq \frac{1}{2} k^{-1/4}$. For every $j \in [M]$ and $i \in [k]$, let $B_j[i]$ be the event that $\langle v_j, \widetilde{w}_i \rangle \leq 3k^{-1/4}$. We would like to condition on the event that $\mathcal{E} \triangleq \left(\bigvee_{j \in [M]} A_j \right) \wedge \left(\bigwedge_{i \in [k], j \in [M]} B_j[i] \right)$.

We first verify that conditioned on $\mathcal{E}$, 1) and 2) of the lemma hold. We get that there is at least one $j \in [M]$ for which

$$\begin{aligned}\|w_{i^*} - a_t - \eta v_j\|_2^2 &\leq \|w_{i^*} - a_t\|_2^2 + \eta^2 - \eta \sigma_t \cdot k^{-1/4} \\ &\leq \left(1 + \frac{1}{4} \left(\left(\frac{\sigma^*}{\sigma_t} \right)^2 - 2 \left(\frac{\sigma^*}{\sigma_t} \right) \right) k^{-1/2} \right) \sigma_t^2 \\ &\leq \left(1 - \frac{1 - 0.1^2}{4\sqrt{k}} \right) \sigma_t^2 \leq \left(1 - \frac{1}{5\sqrt{k}} \right) \sigma_t^2.\end{aligned}$$

where the first step follows from the fact that we have conditioned on A_j and also $\|w_{i^*} - a_t\|_2 = \sigma_t$ by definition, and the third step follows from $\sigma^* / \sigma_t \in [0.9, 1.1]$.

For every $i \in [k], j \in [M]$ we have that

$$\|w_i - a_t - \eta v_j\|_2^2 \geq \|w_i - a_t\|^2 + \eta^2 - 6\eta \sigma_t \cdot k^{-1/4} \geq \left(1 - \frac{9}{\sqrt{k}} \right) \sigma_t^2,$$

where the first step follows from the fact that we have conditioned on $B_j[i]$ and also $\|w_i - a_t\|_2 \leq \sigma_t$ by definition.

Finally, we show that $\Pr[\mathcal{E}] \geq 1 - \delta$. For any $j \in [M]$ and $i \in [k]$, by Corollary 1.3.19, with probability at least $e^{-\sqrt{k}}$ we have that

$$\left\langle \frac{g_j}{\|g_j\|_2}, \frac{\mathbf{U}^\top (w_i - a_t)}{\|\mathbf{U}^\top (w_i - a_t)\|_2} \right\rangle \geq k^{-1/4}. \qquad (8.17)$$

Because

$$\langle g_j, \mathbf{U}^\top (w_i - a_t) \rangle = \langle \mathbf{U} g_j, w_i - a_t \rangle,$$

and $\|\mathbf{U}g_j\|_2 = \|g_j\|_2$ by orthonormality of the columns of $\mathbf{U}$, we can rewrite the left-hand side of (8.17) as

$$\frac{\langle v_j, w_i - a_t\rangle}{\|\mathbf{U}^\top(w_i - a_t)\|_2} \le 2\langle v_j, \widetilde{w}_i\rangle,$$

where the inequality follows by the lower bound in (8.15). So by taking $i = i^*$, we conclude that $\Pr[A_j] \ge e^{-\sqrt{k}}$. The probability that $\bigvee_{j\in[M]} A_j$ does not occur is thus

$$\Pr\left[\bigwedge_{j\in[M]} \overline{A}_j\right] \le \left(1 - e^{-\sqrt{k}}\right)^M.$$

On the other hand, by the same analysis, this time invoking the *second* part of Corollary 1.3.19 and the *upper* bound in (8.15), we see that $\Pr[B_j[i]] \ge e^{-3\sqrt{k}}$, so the probability that $\bigwedge B_j[i]$ does not occur is

$$\Pr\left[\bigvee_{i\in[k],j\in[M]} \overline{B}_j[i]\right] \le kM\cdot e^{-3\sqrt{k}}.$$

So by taking $M = e^{\sqrt{k}}\ln(2/\delta)$ and noting that for this choice of M, $kMe^{-3\sqrt{k}} < \delta/2$ because $\delta > e^{-\sqrt{k}}$, we get that $\Pr[\mathcal{E}] \ge 1-\delta$ as claimed. □

Lemma 8.5.13. *There is an absolute constant $C > 0$ for which the following holds. Let $\mathcal{D}$ be a mixture of spherical linear regressions with mixing weights $\{p_i\}$, directions $\{w_i\}$, and noise rate ς. For any $\varepsilon, \delta > 0$ and $\varsigma^2 \le \varepsilon^2/10$, with probability at least $1-\delta$,* FourierMomentDescent*$(\mathcal{D},\delta,\varepsilon)$ (Algorithm 31) outputs direction $a_T \in \mathbb{R}^d$ for which $\min_{i\in[k]}\|w_i - a_T\|_2 \le \varepsilon$.*

Proof. Let $\sigma_t \triangleq \min_{i\in[k]}\|w_i - a_t\|_2$. Because $a_0 = 0$, we have that $\sigma_0 \le \max_{i\in[k]}\|w_i\|_2 \le 1$.

By a simple union bound, we first upper bound the probability that the steps of moment descent in the t-th iteration of the outer loop in FourierMomentDescent all succeed.

Claim 8.5.14. *Let $i \in [S]$. With probability at least $1-\delta$, the randomized components of the t-th iteration of the outer loop in* FourierMomentDescent *all succeed.*

Proof. Each t-th iteration of the outer loop in FourierMomentDescent (Algorithm 31) has the following randomized components: computing $\widehat{\mathbf{M}}_{a_t}^{(N_1)}$, running EstimateMinVariance (Algorithm 29), trying the Gaussian vectors g in the inner loop over $j \in [M]$, running ApproxBlockSVD, and running CompareMinVariances (Algorithm 30) in this inner loop.

Because the failure probability δ' for the first four of these tasks was chosen to be $\frac{\delta}{5T}$, and the failure probability δ'' for the last task was chosen to be $\frac{\delta}{5MT}$, we can bound the overall failure probability by δ. □

Call the event in Claim 8.5.14 $\mathcal{E}$. Next, we show that provided $\mathcal{E}$ occurs, σ_t can be naively bounded by a constant.

Claim 8.5.15. *Let $0 \le t < T$ and condition on $\mathcal{E}$. Then $\sigma_t \le 4$.*

Proof. At the start of the t-th step, our initial estimate a_{t-1} is at distance at most 1 from some w_{i^*} (this is a very loose bound). After the t-th step, the new estimate a_t satisfies $\|w_{i^{**}} - a_t\|_2 \le \|w_{i^*} - a_{t-1}\|_2 \le 1$ for some $i^{**} \in [k]$. So we have that $\|w_i - a_t\|_2 \le \|w_{i^{**}} - w_i\|_2 + \|w_{i^{**}} - a_t\|_2 \le 3$. Recalling that $\sigma_t^2 = \min_{i\in[k]}\|w_i - a_t\|_2^2 + \varsigma^2$ and noting that $\varsigma < 1$, we conclude that $\overline{\sigma} = 4$ is a valid upper bound on the standard deviation of any component of any univariate mixture of Gaussians $\mathcal{F}_t$ or $\mathcal{F}_t'^{(j)}$ encountered during the course of FourierMomentDescent. □

Next, we show that provided $\mathcal{E}$ occurs, then we can bound the extent to which every iteration of the outer loop in FourierMomentDescent contracts σ_t^2.

Claim 8.5.16. *Let $0 \le t < T$ and condition on $\mathcal{E}$. Suppose $\varsigma^2 \le \frac{1}{5}\|w_i - a_t\|_2^2$ for any $i \in [k]$. Then*

1. **(Completeness)** *Either $\|w_i - a_t\|_2 \le \varepsilon$ already, or there exists some $j \in [M]$ for which* CompareMinVariances*$(\mathcal{F}_t, \mathcal{F}_t'^{(j)}, \overline{\sigma}, \underline{\sigma}, \kappa, 2\kappa, \delta'')$ outputs* True *for $\kappa = \frac{1}{24\sqrt{k}}$.*

2. **(Soundness)** *For any such $j \in [M]$ for which* CompareMinVariances *outputs* True*,*

$$\left(1 - \frac{9}{\sqrt{k}}\right)\sigma_t^2 \le \sigma_{t+1}^2 \le \left(1 - \frac{1}{48\sqrt{k}}\right)\sigma_t^2. \tag{8.18}$$

Proof. We first show completeness. Suppose $\|w_i - a_t\|_2 \geq \varepsilon$ for all $i \in [k]$. By the first part of Lemma 8.5.12, there exists some $j \in [M]$ for which

$$\min_{i\in[k]} \|w_i - a_t'^{(j)}\|_2^2 \leq \left(1 - \frac{1}{5\sqrt{k}}\right) \min_{i\in[k]} \|w_i - a_t\|_2^2$$

and therefore

$$\begin{aligned} \varsigma^2 + \min_{i\in[k]} \|w_i - a_t'^{(j)}\|_2^2 &\leq \varsigma^2 + \left(1 - \frac{1}{5\sqrt{k}}\right) \min_{i\in[k]} \|w_i - a_t\|_2^2 \\ &\leq \left(1 - \frac{1}{6\sqrt{k}}\right) \left(\varsigma^2 + \min_{i\in[k]} \|w_i - a_t\|_2^2\right), \end{aligned}$$

where in the last step we used the assumption that $\varsigma^2 \leq \frac{1}{5}\|w_i - a_t\|_2^2$ for any $i \in [k]$.

We conclude that $\sigma_t^{(j)} \triangleq \min_{i\in[k]} \left\{\varsigma^2 + \|w_i - a_t'^{(j)}\|_2\right\}$ satisfies $(\sigma_t^{(j)})^2 \leq \left(1 - \frac{1}{6\sqrt{k}}\right)(\sigma_t)^2$, and because $1 - \frac{1}{6\sqrt{k}} \leq \left(\frac{1}{1+2\kappa}\right)^2$ for $\kappa = \frac{1}{24\sqrt{k}}$, CompareMinVariances$(\mathcal{F}_t, \mathcal{F}_t'^{(j)}, \overline{\sigma}, \underline{\sigma}, \kappa, 2\kappa, \delta'')$ would return True, completing the proof of completeness.

For soundness, if CompareMinVariances$(\mathcal{F}_t, \mathcal{F}_t'^{(j)}, \overline{\sigma}, \underline{\sigma}, \kappa, 2\kappa, \delta'')$ returns True for some $j \in [M]$, by Corollary 8.5.9 this means

$$(\sigma_t^{(j)})^2 \leq (1+\kappa)^{-2} \cdot \sigma_t^2 \leq (1 - \kappa/2) \cdot \sigma_t^2 \leq \left(1 - \frac{1}{48\sqrt{k}}\right) \cdot \sigma_t^2,$$

where the second step follows from $\kappa \in (0,1)$, which gives the upper bound in (8.18).

Finally, for the lower bound in (8.18), note that the second part of Lemma 8.5.12 tells us that

$$\min_{i\in[k]} \|w_i - a_t'^{(j)}\|_2^2 \geq \left(1 - \frac{9}{\sqrt{k}}\right) \|w_i - a_t\|_2^2$$

and therefore

$$\varsigma^2 + \min_{i\in[k]} \|w_i - a_t'^{(j)}\|_2^2 \geq \varsigma^2 + \left(1 - \frac{9}{\sqrt{k}}\right) \|w_i - a_t\|_2^2 \geq \left(1 - \frac{9}{\sqrt{k}}\right) \left(\varsigma^2 + \|w_i - a_t\|_2^2\right).$$

□

We are now ready to complete the proof of Lemma 8.5.13. Let $\rho = 1.1/0.9$ and condition on $\mathcal{E}$.

If there does not exist $0 \le t < T$ for which we have that

$$\min_{i\in[k]} \|w_i - a_t\|_2^2 \le \varepsilon^2/\rho^2 - \varsigma^2, \tag{8.19}$$

then because $\varsigma^2 \le \varepsilon^2/10$, we get that $\|w_i - a_t\|_2^2 \ge \varepsilon^2/2$. So $\varsigma^2 \le \frac{1}{5}\|w_i - a_t\|_2^2$, and by completeness and soundness in Claim 8.5.16, σ_t^2 has contracted by at least a factor of $(1 - 1/48\sqrt{k})$ and by at most a factor of $(1 - 9/\sqrt{k})$ at every step. So if we take $T = \Omega(\sqrt{k} \cdot \ln(1/\varepsilon))$, we are guaranteed that

$$\min_{i\in[k]} \|w_i - a_T\|_2 \le \sigma_T \le \varepsilon.$$

On the other hand, if (8.19) holds for some $0 \le t < T$, then

$$(\sigma_t^*)^2 \le 1.21\sigma_t^2 \le 1.21 \cdot \left(\min_{i\in[k]} \|w_i - a_t\|_2^2 + \varsigma^2\right) \le 0.99^2\varepsilon^2,$$

so FourierMomentDescent breaks out at Line 12 and correctly outputs a_t.

Conversely, if FourierMomentDescent breaks out at Line 12 because $\sigma_t^* \le 0.99\varepsilon$, this implies that

$$\min_{i\in[k]} \|w_i - a_t\|^2 + \varsigma^2 \le (\sigma_t^*)^2/0.99^2 \le \varepsilon^2,$$

so $\min_{i\in[k]} \|w_i - a_t\|_2 \le \varepsilon$.

The last thing to check is that $\underline{\sigma} = \varepsilon/3$ is always a valid lower bound for any σ_t. If (8.19) holds for some t, t is necessarily the first (and last) t in FourierMomentDescent for which (8.19) holds because of Line 12. So it must be that

$$\sigma_{t-1}^2 \ge \min_{i\in[k]} \|w_i - a_{t-1}\|_2^2 > \varepsilon^2/\rho^2 - \varsigma^2 \ge \varepsilon^2 \cdot \left((0.9/1.1)^2 - 1/5\right) \ge 0.4\varepsilon^2$$

and thus, by the fact that $\sigma_t \ge (1 - 9/\sqrt{k})\sigma_{t-1} \ge 0.99\sigma_{t-1}$, we conclude that $\sigma_t > \varepsilon/3$ as desired. □

Lastly, we calculate the runtime and sample complexity of FourierMomentDescent.

Algorithm 31: FourierMomentDescent$(\mathcal{D}, \delta, \varepsilon)$

```
Input: Sample access to mixture of linear regressions $\mathcal{D}$ with separation $\Delta$ and
       noise rate $\varsigma$, failure probability $\delta$, error $\varepsilon$
Output: $a_T \in \mathbb{R}^d$ satisfying $\min_{i\in[k]}\|w_i - a_T\|_2 \le \varepsilon$, with probability at least $1-\delta$
$a_0 \leftarrow 0$, $T \leftarrow \Omega(\sqrt{k}\cdot\ln(1/\varepsilon))$.
$\delta' \leftarrow \frac{\delta}{5T}$.
$M \leftarrow e^{\sqrt{k}}\ln(2/\delta')$.
$\delta'' \leftarrow \frac{\delta}{5MT}$.
$\overline{\sigma} \leftarrow 4$, $\underline{\sigma} \leftarrow \varepsilon/3$.
for $0 \le t < T$ do
    Let $\mathcal{F}_t$ be the univariate mixture of Gaussians which can be sampled from by
      drawing $(x,y) \sim \mathcal{D}$ and computing $y - \langle x, a_t\rangle$.
    $p \leftarrow 20\ln\left(\frac{3}{2p_{\min}}\right) + 1$.
    $\kappa \leftarrow \frac{1}{24\sqrt{k}}$.
    $\sigma^*_t \leftarrow$ EstimateMinVariance$(\mathcal{F}_t, \overline{\sigma}, \underline{\sigma}, p, \delta')$.        // Algorithm 29
    if $\sigma^*_t < 0.99\varepsilon$ then
        return $a_t$.
    $N_1 \leftarrow \widetilde{\Omega}\left(\frac{\overline{\sigma}^2}{(\sigma^*_t)^2}\cdot p_{\min}^{-2}\cdot k^2\cdot d\cdot\ln(k/\delta')\right)$.
    Draw $N_1$ i.i.d. samples $\{(x_i,y_i)\}_{i\in[N_1]}$ from $\mathcal{D}$ and form the matrix $\widehat{\mathbf{M}}^{(N_1)}_{a_t}$.
    Let $\mathbf{U}_t =$ ApproxBlockSVD $(\widehat{\mathbf{M}}^{(N_1)}_{a_t}, 1/10, \delta')$.        // Lemma 8.5.11
    for $j \in [M]$ do
        Sample $g^{(j)}_t \sim \mathcal{N}(0, \mathrm{Id}_k)$ and define $v^{(j)}_t = \frac{\mathbf{U}_t g^{(j)}_t}{\|\mathbf{U}_t g^{(j)}_t\|_2} \in \mathbb{S}^{d-1}$.
        $a'^{(j)}_t \leftarrow a_t + \eta_t v_j$ for $\eta_t \triangleq \frac{1}{2}k^{-1/4}\cdot\sigma^*_t$.
        Let $\mathcal{F}'^{(j)}_t$ be the univariate mixture of Gaussians which can be sampled from
          by drawing $(x,y) \sim \mathcal{D}$ and computing $y - \langle x, a'^{(j)}_t\rangle$.
        if CompareMinVariances$(\mathcal{F}_t, \mathcal{F}'^{(j)}_t, \overline{\sigma}, \underline{\sigma}, \kappa, 2\kappa, \delta'')$ = True then
            $a_{t+1} \leftarrow a'^{(j)}_t$.
            Break.
```

Lemma 8.5.17 (Running time of FourierMomentDescent)**.** *Let*

$$N_1 = \widetilde{O}(\varepsilon^{-2} p_{\min}^{-2} d k^2 \ln(1/\delta))$$
$$N = p_{\min}^{-4} k \ln(1/\delta) \cdot \mathrm{poly}\left(\sqrt{k}, \ln(1/p_{\min}), \ln(1/\varepsilon)\right)^{O\left(\sqrt{k}\ln(1/p_{\min})\right)}.$$

Then FourierMomentDescent *(Algorithm 31) requires sample complexity $\widetilde{O}(\sqrt{k}e^{\sqrt{k}}(N_1+N))$ and runs in time $\widetilde{O}(\sqrt{k}e^{\sqrt{k}}(dN_1+N))$.*

We defer the proof of Lemma 8.5.17 to Appendix 8.12.6.

We can now complete the proof of Theorem 8.5.1.

Proof of Theorem 8.5.1. By Lemma 8.5.13, FourierMomentDescent outputs a vector $a_T \in \mathbb{R}^d$ for which $\|w_i - a_T\|_2 \leq \varepsilon$ for some $i \in [k]$. The runtime and sample complexity bounds follow from Lemma 8.5.17. □

8.6 Learning All Components Under Zero Noise

In this short section we briefly describe how to use FourierMomentDescent in conjunction with existing techniques for boosting to learn *all* components in a mixture of linear regressions. We remark that the arguments in this section are fairly standard.

We will make use of the following local convergence result of [LL18].

Theorem 8.6.1. *Let $\mathcal{D}$ be a mixture of linear regressions in $\mathbb{R}^d$ with regressors $\{w_j\}$, minimum mixing weight $p_{\min}$, separation Δ, and components whose covariances have eigenvalues all bounded within $[1, \sigma]$. Let $\zeta \triangleq \Delta \cdot \min\left(\frac{1}{2\sigma}, \frac{p_{\min}}{64}\right)$. There is an algorithm* LL-Boost$(\mathcal{D}, v, \varepsilon, \delta)$ *which, given any $\varepsilon > 0$ and $v \in \mathbb{R}^d$ for which there exists $j \in [k]$ with $\|w_j - v\|_2 \leq \zeta/\sigma$, draws $T \cdot M$ samples from $\mathcal{D}$ for*

$$T = O(p_{\min}^{-2} d \ln(\zeta/\varepsilon)) \quad \textit{and} \quad M = \mathrm{poly}(1/\Delta, 1/p_{\min}, \sigma, \log T) \cdot \ln(1/\delta),$$

runs in time $T \cdot M \cdot d$, and outputs $\widetilde{v} \in \mathbb{R}^d$ for which $\|w_j - \widetilde{v}\|_2 \leq \varepsilon$ with probability at least $1 - \delta$.

We give a formal specification of our procedure LearnWithoutNoise for learning all components of a noise-less mixture of linear regressions in Algorithm 32 below. The basic approach is to repeatedly invoke FourierMomentDescent to produce an estimate for one of the regressors of $\mathcal{D}$ to within $O(\Delta p_{\min})$ error, run LL-Boost to refine it to an estimate v with error essentially as small as one would like (because of the exponential convergence rate of LL-Boost), and then filter out all samples (x, y) for which the residual $|y - \langle x, v\rangle|$ is sufficiently small.

```
Algorithm 32: LearnWithoutNoise($\mathcal{D}, \delta, \varepsilon$)
  Input: Sample access to mixture of linear regressions $\mathcal{D}$ with separation $\Delta$ and zero
         noise and regressors $\{w_i\}$, failure probability $\delta$, error $\varepsilon$
  Output: List of vectors $\mathcal{L} \triangleq \{\widetilde{w}_1, ..., \widetilde{w}_k\}$ for which there is a permutation
          $\pi : [k] \to [k]$ for which $\|\widetilde{w}_i - w_{\pi(i)}\|_2 \le \varepsilon$ for all $i \in [k]$, with probability at
          least $1 - \delta$
  $\delta' \leftarrow \delta/2k$.
  $\varepsilon_{\mathrm{FMD}} \leftarrow \Delta p_{\min}/64$.
  $\varepsilon_{\mathrm{boost}} \leftarrow \min\{\varepsilon, \mathrm{poly}(p_{\min}, \Delta, 1/k, 1/d)^{\sqrt{k}\ln(1/p_{\min})}\}$.
  for $i \in [k]$ do
      $w'_i \leftarrow$FourierMomentDescent($\mathcal{D}, \delta', \varepsilon_{\mathrm{FMD}}$).
      $\widetilde{w}_i \leftarrow$LL-Boost($\mathcal{D}, w'_i, \varepsilon_{\mathrm{boost}}, \delta'$).
      Henceforth when sampling from $\mathcal{D}$, ignore all samples $(x, y)$ for which
        $|y - \langle x, \widetilde{w}_i\rangle| \le \varepsilon_{\mathrm{boost}} \cdot \mathrm{poly}(\log d)$.
```

Theorem 8.6.2. *Given $\delta, \varepsilon > 0$ and a mixture of spherical linear regressions $\mathcal{D}$ with separation Δ and zero noise, with probability at least $1 - \delta$,* LearnWithoutNoise*($\mathcal{D}, \delta, \varepsilon$) (Algorithm 32) returns a list of vectors $\mathcal{L} \triangleq \{\widetilde{w}_1, ..., \widetilde{w}_k\}$ for which there is a permutation $\pi : [k] \to [k]$ for which $\|\widetilde{w}_i - w_{\pi(i)}\|_2 \le \varepsilon$ for all $i \in [k]$. Furthermore,* LearnWithoutNoise *requires sample complexity*

$$N = \widetilde{O}\left(d\ln(1/\varepsilon)\ln(1/\delta)p_{\min}^{-4}\Delta^{-2} \cdot \mathrm{poly}\left(k, \ln(1/p_{\min}), \ln(1/\Delta)\right)^{O(\sqrt{k}\ln(1/p_{\min}))}\right)$$

and time complexity $Nd \cdot \mathrm{poly}\log(k, d, 1/\Delta, 1/p_{\min}, 1/\varepsilon)$.

Proof. By Theorem 8.5.1, every w'_i in LearnWithoutNoise is $\frac{\Delta p_{\min}}{64}$-close to a regressor $w_{i'}$ of $\mathcal{D}$, and by Theorem 8.6.1, LL-Boost improves this to a vector $\widetilde{w}_i$ for which $\|\widetilde{w}_i - w_{i'}\|_2 \le$

$\varepsilon_{\mathrm{boost}}$, where

$$\varepsilon_{\mathrm{boost}} \min\{\varepsilon, \mathrm{poly}(p_{\min}, \Delta, 1/k, 1/d)^{\sqrt{k}\ln(1/p_{\min})}\}.$$

As a result, only a $\mathrm{poly}(p_{\min}, \Delta, 1/k, 1/d)^{\sqrt{k}\ln(1/p_{\min})}$ fraction of subsequent samples will be removed, and the resulting error can be absorbed into the sampling error that goes into subsequent calls to L2Estimate and subsequent matrices $\widehat{M}_a^{(N)}$ that we run ApproxBlockSVD on, in the remainder of LearnWithoutNoise. □

8.7 Learning All Components Under Noise

In this section, we describe how to learn all components under the much more challenging setting where there is regression noise. We show that, at the extra cost of running in time exponential in $1/\Delta^2$ in addition to $\sqrt{k}$, there is an algorithm, which we call LearnWithNoise, that can learn mixtures of linear regressions to error ε when $\varsigma = O(\varepsilon)$.

Theorem 8.7.1. *Given $\delta, \varepsilon > 0$ and a mixture of spherical linear regressions $\mathcal{D}$ with regressors $\{w_1, ..., w_k\}$, separation Δ, and noise rate $\varsigma = O(\varepsilon)$, with probability at least $1 - \delta$,* LearnWithNoise *$(\mathcal{D}, \delta, \varepsilon)$ (Algorithm 32) returns a list of vectors $\mathcal{L} \triangleq \{\widetilde{w}_1, ..., \widetilde{w}_k\}$ for which there is a permutation $\pi : [k] \to [k]$ for which $\|\widetilde{w}_i - w_{\pi(i)}\|_2 \le \varepsilon$ for all $i \in [k]$. Furthermore,* LearnWithNoise *requires sample complexity*

$$N = \widetilde{O}\left(d\varepsilon^{-2}\ln(1/\varepsilon)\ln(1/\delta)p_{\min}^{-4}\Delta^{-2} \cdot \mathrm{poly}\left(k, 1/\varepsilon, \ln(1/p_{\min})\right)^{O(\sqrt{k}\ln(1/p_{\min})/\Delta^2)}\right)$$

and time complexity $Nd \cdot \mathrm{poly}\log(k, d, 1/\Delta, 1/p_{\min}, 1/\varepsilon)$.

In Section 8.7.1 we prove the key technical ingredient behind our proof of Theorem 8.7.1, Lemma 8.7.4, which allows us to carefully control the dynamics of Fourier moment descent. In Section 8.7.2 we describe how to get an initialization which satisfies the hypotheses of Lemma 8.7.4. In Section 8.7.3 we give the full specification of LearnWithNoise. In Section 8.7.4 we prove Theorem 8.7.1. Finally, in Section 8.7.5, we briefly describe how to leverage the local convergence result of [KC19] in conjunction with our algorithm to get improved noise tolerance in the setting where the mixing weights are *a priori* known.

8.7.1 Staying on the Same Component

The main result of this section and the primary technical component behind Theorem 8.7.1 is Lemma 8.7.4 below. This is a substantially more refined version of Lemma 8.5.12 in which we control not only the probability we make progress in the t-th step of moment descent, but also the probability that the the component a_{t+1} is closest to is the same as the one a_t is closest to.

We first introduce some preliminary notation and facts that we will use in the proof of Lemma 8.7.4.

For $v \in \mathbb{S}^{d-1}$, define $\mathcal{F}$ and $\mathcal{F}'_v$ respectively to be the distribution of $y - \langle a_t, x\rangle$ and of $y - \langle a_t + \eta v, x\rangle$, where $(x, y) \sim \mathcal{D}$.

Let $\sigma_t^2 \triangleq \varsigma^2 + \min_{i\in[k]}\|w_i - a_t\|_2$. Denote the minimizing index i by i^*.

We record the following application of Lemma 8.5.10 and Fact 1.3.6 which says we have access to $\mathrm{span}(\{w_i - a_t\})$ up to $1/\mathrm{poly}(k)$ additive error.

Lemma 8.7.2. *Let $\delta_{\mathrm{samp}}, \delta' > 0$ and $a_t \in \mathbb{R}^d$. If we draw $N_1 = \widetilde{\Omega}\left(p_{\min}^{-1}\delta_{\mathrm{samp}}^{-1} \cdot d \cdot \ln(k/\delta')\right)$ samples, form $\widehat{\mathbf{M}}_{a_t}^{(N_1)} \in \mathbb{R}^{d\times d}$ as defined in (8.14), and run* ApproxBlockSVD$(\widehat{\mathbf{M}}_{a_t}^{(N_1)}, \delta_{\mathrm{samp}}, \delta')$ *to produce a matrix $\mathbf{U} \in \mathbb{R}^{k\times d}$, then with probability $1-\delta'$ we have that for any $a, b \in \mathbb{S}^{d-1}$ in the row span of $\mathbf{U}$,*

1. *$\langle \mathbf{U}a, \mathbf{U}b\rangle \le \langle a, b\rangle - \delta_{\mathrm{samp}}$*

2. *$1 - \delta_{\mathrm{samp}} \le \|\mathbf{U}(w_i - a_t)\|_2 \le 1 - \delta_{\mathrm{samp}}$.*

Proof. By Lemma 8.5.10 and Fact 1.3.6, with probability $1-\delta'$ we can ensure that

$$\|\mathbf{U}^\top \mathbf{\Lambda} \mathbf{U} - \mathbb{E}_{x,y}[\mathbf{M}_{a_t}^{x,y}]\|_2 \le \delta_{\mathrm{samp}} \cdot p_{\min}/2,$$

where $\mathbf{\Lambda} \in \mathbb{R}^k$ is some diagonal matrix of eigenvectors and $\mathbf{M}_{a_t}^{x,y}$ is defined in (8.13) and satisfies $\mathbb{E}_{x,y}[\mathbf{M}_{a_t}^{x,y}] = \sum_{i=1}^k p_i(w_i - a)(w_i - a)^\top$ by Lemma 8.5.10. Parts 1 and 2 of the lemma then follow by Lemma 1.3.8. □

Lastly, the following elementary fact will be useful:

Fact 8.7.3. *If $x \in \mathbb{R}_{\geq 0}$ satisfies $\frac{1}{2}(x + x^{-1}) \geq 1 + \beta^2$ for some $0 < \beta \leq 1$, then $1 - \beta/2 \leq x \leq 1 + \beta/2$.*

Proof. The solutions to $x + x^{-1} = 2 + 2\beta^2$ are

$$x = 1 + \beta^2 \pm \beta\sqrt{2 + \beta^2}.$$

One can check that $\beta^2 + \beta\sqrt{2 + \beta^2} \geq \beta$ for all $\beta \in \mathbb{R}$, while $-\beta^2 + \beta\sqrt{2 + \beta^2} \geq \beta/2$ for $\beta \in [0, 1]$. □

We are now in a position to state and prove our main result of this section, Lemma 8.7.4. This lemma roughly says that if we sample $M = \exp(\Omega(\sqrt{k}/\Delta^2))$ random steps $v_1, ..., v_M$ at time t of moment descent, then with high probability, if $j^* \in [M]$ is the first index on which CompareMinVariances outputs True, then not only does walking in direction v_{j^*} contract σ_t by a factor of $1 - \Omega(\Delta/\sqrt{k})$ with high probability, but additionally, with at least $1/\mathrm{poly}(k)$ probability, it also keeps us closest to the component we were already closest to, in the following robust sense. Specifically, if we have a $(1 + \Omega(\Delta^2\sqrt{k}))$ gap between $\|w_{i^*} - a_t\|_2$ and all other $\|w_i - a_t\|_2$, then with at least $1/\mathrm{poly}(k)$ probability, after one more iteration of moment descent, the i^*-th component will still be the closest to our new guess $a_{t+1} \in \mathbb{R}^d$, and this gap will persist.

Lemma 8.7.4. *There exist constants a_{LR}, a_{trials}, a_{scale}, constants $\overline{\beta} > \underline{\beta}$, a constant $0 \leq a_{\mathrm{noise}} \leq 1/5$, and a constant $\tau_{\mathrm{gap}} > 0$, such that for all $c < \tau_{\mathrm{gap}}$, the following holds for some $0 < \kappa_1 < \kappa_2 \leq 1$ satisfying $\kappa_2 - \kappa_1 = c\Delta^2 k^{-1/2}$.*

Let $\delta > 0$. Suppose that $\varsigma^2 \leq a_{\mathrm{noise}} \cdot \varepsilon^2$. Suppose that

$$\|w_i - a_t\|_2 \leq a_{\mathrm{scale}} \cdot k^{1/4} \tag{8.20}$$

for all $i \in [k]$. For $M \triangleq e^{a_{\mathrm{trials}}\sqrt{k}/\Delta^2} \ln(3/\delta)$ and $g_1, ..., g_M \sim \mathcal{N}(0, Id_k)$, let $v_j = \frac{g_j \mathbf{U}}{\|g_j \mathbf{U}\|_2} \in \mathbb{S}^{d-1}$ for $j \in [M]$. Let σ^* be a number for which $0.9\sigma_t \leq \sigma^* \leq 1.1\sigma_t$, and let $\eta \triangleq a_{\mathrm{LR}} \cdot \Delta \cdot \sigma_* \cdot k^{-1/4}$.

Then with probability at least $1 - \delta$ over the randomness of $g_1, ..., g_M$ as well as over the behavior of all runs of CompareMinVariances*, the following events hold:*

1. **(Progress detected)** *If* $\min_{i\in[k]}\|w_i - a_t\|_2^2 \geq \varepsilon^2/2$*, then*

$$\textsc{CompareMinVariances}(\mathcal{F}, \mathcal{F}'_v, a_{\text{scale}} \cdot k^{-1/4}, \sigma^*/1.1, \kappa_1, \kappa_2, \delta/3M)$$

   outputs True *for at least one* $j \in [M]$*.*

   Let j^* *be the smallest such* j*, and define*

$$a_{t+1} \triangleq a_t + \eta v_{j^*}.$$

2. **(Make at least some amount of progress)** *If* $\min_{i\in[k]}\|w_i - a_t\|_2^2 \geq \varepsilon^2/2$*, then*

$$\sigma_{t+1}^2 \leq \left(1 - \underline{\beta}\Delta^2/\sqrt{k}\right)\sigma_t^2.$$

3. **(Make at most some amount of progress)** *Regardless of whether* $\min_{i\in[k]}\|w_i - a_t\|_2^2 \leq \varepsilon^2/2$*,*

$$\sigma_{t+1}^2 \geq \left(1 - \overline{\beta}\Delta^2/\sqrt{k}\right)\sigma_t^2.$$

If we assume that for all $i \neq i^*$*,*

$$\|w_i - a_t\|_2 \geq \left(1 + c\Delta^2/\sqrt{k}\right) \cdot \|w_{i^*} - a_t\|_2, \tag{8.21}$$

then crucially, we have that with probability $1/\text{poly}(k)$*, the events above hold and* additionally*:*

4. **(i^* remains closest by same margin)** *If* $\min_{i\in[k]}\|w_i - a_t\|_2^2 \geq \varepsilon^2/2$*, then for all* $i \neq i^*$*,*

$$\|w_i - a_t - \eta v_{j^*}\|_2 \geq \left(1 + c\Delta^2/\sqrt{k}\right) \cdot \|w_{i^*} - a_t - \eta v_{j^*}\|_2.$$

We emphasize that the main content of Lemma 8.7.4 is part 4.

Proof. Henceforth we will say that "CompareMinVariances succeeds and outputs True/False on direction v" to mean that a single run of

$$\textsc{CompareMinVariances}(\mathcal{F}, \mathcal{F}'_v, a_{\mathrm{scale}} \cdot k^{-1/4}, \sigma^*/1.1, \kappa_1, \kappa_2, \delta/3M)$$

is successful (in the language of Corollary 8.5.9, this happens with probability $1 - \delta/3M$) and outputs $\mathsf{True}/\mathsf{False}$.

Recall from (8.16) that we have

$$\|w_i - a_t - \eta v_j\|_2^2 = \|w_i - a_t\|_2^2 + \eta^2 - 2\eta\langle w_i - a_t, v_j\rangle. \tag{8.22}$$

Define $\updelta_i \triangleq w_i - a_t$ and $\widehat{\updelta}_i \triangleq \frac{w_i - a_t}{\|w_i - a_t\|_2}$. For every $i \neq i^*$, define $\updelta_i^\perp \triangleq \widehat{\updelta}_i - \langle \widehat{\updelta}_{i^*}, \widehat{\updelta}_i\rangle \widehat{\updelta}_{i^*}$. Finally, let $\gamma_i^{(j)} = \langle \widehat{\updelta}_i, v_j\rangle$. Where the context is clear, we will omit the superscript (j).

Let $\nu_A, \nu_B, \nu_C > 0$ be absolute constants, and suppose $\nu_A < \nu_B$. For $i \in [k]$ and $j \in [M]$, define the following events:

1. Let $A_j[i]$ be the event that $\gamma_{i^*}^{(j)} \geq \nu_A \Delta k^{-1/4}$.
2. Let $B_j[i]$ be the event that $\gamma_i^{(j)} \leq \nu_B \Delta k^{-1/4}$.
3. Let C_j be the event that $A_j[i^*]$ occurs and also $\langle v_j, \updelta_i^\perp\rangle \leq \nu_C \Delta^2 k^{-1/2} \|\updelta_i^\perp\|_2$ for all $i \neq i^*$.

For $j \in [M]$, also let B_j denote the event that $B_j[i]$ occurs for every $i \in [k]$.

By our assumption on σ_* and the definition of η, we know that $\eta = a'_{\mathrm{LR}} \cdot k^{-1/4} \cdot \|\updelta_i\|_2$, where $a'_{\mathrm{LR}} \in [0.9, 1.1] \cdot a_{\mathrm{LR}}$. It will be useful later in the proof to assume that $\nu_B < a'_{\mathrm{LR}} < 2\nu_A$.

First, we compute the exact distance to v_{i^*} after walking along v_j and, provided the events $B_j[i]$ occur, lower bound the distances to all other components v_i.

Claim 8.7.5. *Let $i \in [k], j \in [M]$, and suppose $B_j[i]$ occurs. Then*

$$\|\updelta_i - \eta v_j\|_2^2 \geq \|\updelta_i\|_2^2 \cdot \left(1 + a'^2_{\mathrm{LR}} k^{-1/2} - 2a'_{\mathrm{LR}} k^{-1/4} \gamma_i^{(j)}\right), \tag{8.23}$$

with equality when $i = i^*$. Furthermore, when $i \neq i^*$ we get from (8.21) that

$$\|\updelta_i - \eta v_j\|_2^2 \geq \|\updelta_{i^*}\|_2^2 \cdot \left(\left(1 + c\Delta^2 k^{-1/2}\right)^2 + a'^2_{\mathrm{LR}} \Delta^2 k^{-1/2} - 2a'_{\mathrm{LR}} \Delta k^{-1/4} \gamma_i^{(j)} - 2ca'_{\mathrm{LR}} \Delta^3 k^{-3/4} \gamma_i^{(j)}\right). \tag{8.24}$$

Proof. We may rewrite (8.22) as

$$\|\delta_i - \eta v_j\|_2^2 = \|\delta_i\|_2^2 + a'^2_{\mathrm{LR}}\Delta^2 k^{-1/2}\|\delta_{i^*}\|_2^2 - 2a'_{\mathrm{LR}}\Delta k^{-1/4}\|\delta_{i^*}\|_2\|\delta_i\|_2 \cdot \gamma_i^{(j)}. \tag{8.25}$$

The right-hand side of (8.25), as a function of $\|\delta_{i^*}\|_2$, is decreasing as long as

$$\|\delta_{i^*}\|_2 \leq a'^{-1}_{\mathrm{LR}}\Delta^{-1}k^{1/4}\|\delta_i\|_2\gamma_i^{(j)}.$$

But this condition holds because event $B_j[i]$ occurs, $\nu_B < a'_{\mathrm{LR}}$, and $\|\delta_{i^*}\|_2 \leq \|\delta_i\|_2$. So (8.23) follows, with equality when $i = i^*$.

When $i \neq i^*$, we additionally know that $\|\delta_i\|_2 \geq \left(1 + c\Delta^2 k^{-1/2}\right) \cdot \|\delta_{i^*}\|_2$. So by the fact that the right-hand side of (8.25) is decreasing as a function of $\|\delta_{i^*}\|_2$ for $\|\delta_{i^*}\|_2 \in (-\infty, \|\delta_i\|_2]$, we get (8.24). □

Using (8.23) of Claim 8.7.5, which is an equality when $i = i^*$, we can upper bound the distance to v_{i^*} after walking along v_j, provided events $A_j[i^*]$ and $B_j[i]$ occur.

Claim 8.7.6. *Let $j \in [M]$, and suppose $A_j[i^*]$ and $B_j[i^*]$ occur. Then there is an absolute constant $\underline{\beta}' > 0$ for which*

$$\|\delta_{i^*} - \eta v_j\|_2^2 \leq \|\delta_{i^*}\|_2^2 \cdot (1 - \underline{\beta}'\Delta^2 k^{-1/2}). \tag{8.26}$$

Proof. By (8.23) which is an equality when $i = i^*$,

$$\|\delta_{i^*} - \eta v_j\|_2^2 \leq \|\delta_{i^*}\|_2^2 \cdot \left(1 - (2a'_{\mathrm{LR}}\nu_A - a'^2_{\mathrm{LR}})\Delta^2 k^{-1/2}\right).$$

The claim follows by taking $\underline{\beta}' \triangleq 2a'_{\mathrm{LR}}\nu_A - a'^2_{\mathrm{LR}}$, which is positive by the assumption that $a'_{\mathrm{LR}} < 2\nu_A$. □

Next, using (8.24) of Claim 8.7.5, we argue that the only way to make progress towards a *different* component $i \neq i^*$ by an amount comparable to that of Claim 8.7.6, is if $A_j[i]$ has occurred. In particular, the following claim is the contrapositive of this.

Claim 8.7.7. *Let $i \neq i^*$ and $j \in [M]$, and suppose $B_j[i]$ occurs and $A_j[i]$* does not occur. *Then*

$$\|\boldsymbol{\delta}_i - \eta v_j\|_2^2 \geq \|\boldsymbol{\delta}_{i^*}\|_2^2 \cdot \left(1 - (\underline{\beta}' - c)\Delta^2 k^{-1/2}\right)$$

Proof. Because $A_j[i]$ does not occur, $\gamma_i^{(j)} < \nu_A k^{-1/4}$. So by (8.24),

$$\begin{aligned}
\|\delta_i - \eta v_j\|_2^2 &\geq \|\delta_{i^*}\|_2^2 \cdot \left(\left(1 + c\Delta^2 k^{-1/2}\right)^2 + a'^2_{\mathrm{LR}}\Delta^2 k^{-1/2} - 2\nu_A a'_{\mathrm{LR}}\Delta^2 k^{-1/2} - 2\nu_A \cdot c \cdot a'_{\mathrm{LR}}\Delta^4 k^{-1}\right) \\
&= \|\delta_{i^*}\|_2^2 \cdot \left(\left(1 + c\Delta^2 k^{-1/2}\right)^2 - \underline{\beta}'^2\Delta^2 k^{-1/2} - 2\nu_A \cdot c \cdot a'_{\mathrm{LR}}\Delta^4 k^{-1}\right) \\
&= \|\delta_{i^*}\|_2^2 \cdot \left(1 - \underline{\beta}'\Delta^2 k^{-1/2} + 2c\Delta^2 k^{-1/2} \cdot \left(1 - \nu_A a'_{\mathrm{LR}}\Delta^2 k^{-1/2}\right) + c^2\Delta^4 k^{-1}\right) \\
&\geq \|\delta_{i^*}\|_2^2 \cdot \left(1 - \underline{\beta}'\Delta^2 k^{-1/2} + c\Delta^2 k^{-1/2}\right),
\end{aligned}$$

where in the last step we used the fact that $1 - \nu_A a'_{\mathrm{LR}}\Delta^2 k^{-1/2} \geq 1/2$ for sufficiently large k. □

Henceforth, let $\kappa_1 = \left(\underline{\beta}' - \frac{3c}{2}\right)\Delta^2 k^{-1/2}$ and $\kappa_2 = \left(\underline{\beta}' - \frac{c}{2}\right)\Delta^2 k^{-1/2}$. In Lemma 8.7.4, we will take $\underline{\beta} \triangleq \underline{\beta}' - \frac{3c}{2}$.

Claims 8.7.6 and 8.7.7 now imply the following about the behavior of CompareMinVariances. The upshot of the following two corollaries is that for any $j \in [M]$, if $B_j[i]$ occurs for every i and CompareMinVariances succeeds and outputs True on direction v_j, the conditional probability of $A_j[i^*]$ happening is at least the conditional probability of $A_j[i]$ happening for any $i \neq i^*$.

Corollary 8.7.8. *Let $j \in [M]$, and suppose $A_j[i^*]$ and $B_j[i^*]$ occur. Then* CompareMinVariances *succeeds and outputs* True *on direction v_j.*

Proof. By adding ς^2 to both sides of (8.26) in Claim 8.7.6, we see that

$$\sigma_{t+1}^2 \leq \|\delta_{i^*} - \eta v_j\|_2^2 + \varsigma^2 \leq \|\delta_{i^*}\|_2^2 \cdot (1 - \underline{\beta}'\Delta^2 k^{-1/2}) + \varsigma^2 \leq \left(1 - (\underline{\beta}' - c/2)\Delta^2 k^{-1/2}\right)\left(\|\delta_{i^*}\|_2^2 + \varsigma^2\right),$$

where in the last step we used the assumptions that $\varsigma^2 \le a_{\text{noise}} \cdot \varepsilon^2$ and $\|w_{i^*} - a_t\|_2^2 \ge \varepsilon^2/2$ for some sufficiently small constant a_{noise}, which we just need to be at most $\frac{c}{4\underline{\beta}'}$ here. □

Corollary 8.7.9. *Let $i \neq i^*$ and $j \in [M]$, and suppose $B_j[i]$ holds and* CompareMinVariances *succeeds and outputs* True *on direction v_j. Then $A_j[i]$ has also occurred.*

Proof. By the contrapositive of Claim 8.7.7, if

$$\|\delta_i - \eta v_j\|_2^2 < \|\delta_{i^*}\|_2^2 \cdot \left(1 - (\underline{\beta}' - c)\Delta^2 k^{-1/2}\right), \tag{8.27}$$

and $B_j[i]$ occurs, then $A_j[i]$ occurs. We would like to show that (8.27) then implies that CompareMinVariances, if it succeeds, outputs True on direction v_j. Adding ς^2 to both sides of this, we conclude that

$$\sigma_{t+1}^2 = \varsigma^2 + \|\delta_i - \eta v_j\|_2^2 < \varsigma^2 + \|\delta_{i^*}\|_2^2 \cdot \left(1 - (\underline{\beta}' - c)\Delta^2 k^{-1/2}\right) \le \left(1 - \left(\underline{\beta}' - \frac{3c}{2}\right)\Delta^2 k^{-1/2}\right),$$

where in the last we used the assumptions that $\varsigma^2 \le a_{\text{noise}} \cdot \varepsilon^2$ and $\|w_{i^*} - a_t\|_2^2 \ge \varepsilon^2/2$ for some sufficiently small constant a_{noise}, which we just need to be at most $\frac{c}{4\underline{\beta}'-c}$ here. □

We also give an upper bound to the amount of progress that any v_j could make in any direction i, provided $B_j[i]$ holds.

Claim 8.7.10. *Let $i \in [k], j \in [M]$, and suppose $B_j[i]$ occurs. Then there is an absolute constant $\overline{\beta} > \underline{\beta}$ for which $\|\delta_i - \eta v_j\|_2^2 \ge \|\delta_{i^*}\|_2^2 \cdot (1 - \overline{\beta}\Delta^2 k^{-1/2})$.*

Proof. By (8.23),

$$\|\delta_i - \eta v_j\|_2^2 \ge \|\delta_i\|_2^2 \cdot \left(1 - (2a'_{\text{LR}}\nu_B - a'^2_{\text{LR}})\Delta^2 k^{-1/2}\right).$$

The claim follows by taking $\overline{\beta} = 2a'_{\text{LR}}\nu_B - a'^2_{\text{LR}}$. Note that we have that $\overline{\beta} > \underline{\beta}' > \underline{\beta}$ because $\nu_A < \nu_B$. □

At this point, we could already use Corollary 8.7.8 and Claim 8.7.10, together with straightforward bounds on the probabilities of the events $A_j[i^*]$ and $B_j[i]$ (see Claims 8.7.12 and 8.7.13 below) to show that parts 1), 2), and 3) of the lemma hold with the claimed probability. Note that the proofs of these steps do not use (8.21), so in particular parts 1), 2), and 3) of the lemma hold with the claimed probability *without assuming* (8.21).

We next lay the foundation for showing part 4) of the lemma holds with at least $1/\mathrm{poly}(k)$ probability, assuming (8.21). Thus far we have not talked about the events C_j. It is at this point that we arrive at the main claim of the proof, namely that if event C_j happens, then the gap of (8.21) between the i^*-th component and all other components persists in the next step.

Claim 8.7.11. *Let $j \in [M]$ and suppose C_j and $B_j[i]$ occur for all $i \in [k]$. Then*

$$\|\delta_i - \eta v_j\|_2 \geq \left(1 + c\Delta^2 k^{-1/2}\right) \cdot \|\delta_{i^*} - \eta v_j\|_2$$

for all $i \neq i^*$.

Proof. Suppose we could show that

$$\gamma_i^{(j)} \leq \gamma_{i^*}^{(j)} \tag{8.28}$$

for all $i \neq i^*$. Then by (8.23), we would conclude that

$$\begin{aligned}\frac{\|\delta_i - \eta v_j\|_2^2}{\|\delta_{i^*} - \eta v_j\|_2^2} &\geq \frac{\|\delta_i\|^2}{\|\delta_{i^*}\|_2^2} \cdot \frac{1 + a_{\mathrm{LR}}'^2 k^{-1/2} - 2a'_{\mathrm{LR}} k^{-1/4}\gamma_i^{(j)}}{1 + a_{\mathrm{LR}}'^2 k^{-1/2} - 2a'_{\mathrm{LR}} k^{-1/4}\gamma_{i^*}^{(j)}} \\ &\geq \left(1 + c\Delta^2 k^{-1/2}\right)^2 \cdot \frac{1 + a_{\mathrm{LR}}'^2 k^{-1/2} - 2a'_{\mathrm{LR}} k^{-1/4}\gamma_i^{(j)}}{1 + a_{\mathrm{LR}}'^2 k^{-1/2} - 2a'_{\mathrm{LR}} k^{-1/4}\gamma_{i^*}^{(j)}} \\ &\geq \left(1 + c\Delta^2 k^{-1/2}\right)^2\end{aligned}$$

as desired.

We now describe the intuition for the remaining argument. (8.28) is not hard to show when the unit vector $\widehat{\delta_i}$ is somewhat far from $\widehat{\delta_{i^*}}$, in which case it is reasonable to imagine a sizable cone of directions around δ_{i^*} such that if v_j lies in that cone, (8.28) holds. On the other hand, suppose $\widehat{\delta_i}$ is close to δ_{i^*}. Then (8.28) can actually be false. But because their non-normalized counterparts δ_i and δ_{i^*} are assumed to be Δ-separated, δ_i and δ_{i^*} must therefore be nearly collinear, in which case there must exist a gap between $\|\delta_i\|_2$ and $\|\delta_{i^*}\|_2$ that's even bigger than the one assumed in (8.21), and furthermore walking in v_j cannot reduce this gap to below that of (8.21) in the next step.

We now proceed with the formal details. First note that

$$\gamma_i^{(j)} = \left\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \right\rangle \cdot \gamma_{i^*}^{(j)} + \left\langle \delta_i^{\perp}, v_j \right\rangle. \tag{8.29}$$

Now if $\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \rangle \leq 0$, then by event C_j, $\gamma_i^{(j)} \leq \langle \delta_i^{\perp}, v_j \rangle \leq \nu_C \Delta^2 k^{-1/2} \cdot \|\delta_i^{\perp}\|_2 < \gamma_{i^*}^{(j)}$ and we'd be done. On the other hand, if $\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \rangle > 0$, then we get that

$$\gamma_i^{(j)} \leq \left\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \right\rangle \cdot \gamma_{i^*}^{(j)} + \nu_C \Delta^2 k^{-1/2} \cdot \|\delta_i^{\perp}\|_2.$$

In this case, to show the desired inequality (8.28), it would suffice to show that

$$\gamma_{i^*}^{(j)} \left(1 - \left\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \right\rangle\right) \geq \nu_C \Delta^2 k^{-1/2} \cdot \|\delta_i^{\perp}\|_2.$$

In particular, because event A_j holds so that $\gamma_{i^*}^{(j)} \geq \nu_A \Delta k^{-1/4}$, we just need to show that

$$\nu_A \left(1 - \left\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \right\rangle\right) \geq \nu_C \Delta k^{-1/4} \cdot \|\delta_i^{\perp}\|_2. \tag{8.30}$$

After squaring both sides of (8.30), making the substitution $\|\delta_i^{\perp}\|_2^2 = 1 - \left\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \right\rangle^2$, and rearranging, (8.30) becomes

$$\nu_A^2 \left(1 - \left\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \right\rangle\right)^2 - \nu_C^2 \Delta^2 k^{-1/2} \cdot \left(1 - \left\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \right\rangle^2\right) \geq 0. \tag{8.31}$$

This is merely a univariate inequality for a quadratic polynomial in $\left\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \right\rangle$. Let $\nu_{CA} \triangleq \nu_C / \nu_A$. One can compute the smaller of the two zeros of the left-hand side of (8.31) and see that the inequality is satisfied provided that $\left\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \right\rangle$ is at most

$$1 - \frac{2\nu_{CA}^2 \Delta^2 k^{-1/2}}{1 + \nu_{CA}^2 \Delta^2 k^{-1/2}} \geq 1 - a_{\mathrm{Del}} \cdot \Delta^2 k^{-1/2}$$

for absolute constant $a_{\mathrm{Del}} \triangleq \frac{2\nu_{CA}^2}{1+\nu_{CA}^2 \Delta^2 k^{-1/2}}$.

It remains to consider the case where $\left\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \right\rangle \geq 1 - a_{\mathrm{Del}} \cdot \Delta^2 k^{-1/2}$. This is where we use the fact that $\|\delta_i - \delta_{i^*}\|_2 = \|w_i - w_{i^*}\|_2 \geq \Delta$ to argue that, even though (8.31) does not hold and we cannot obtain (8.28), $\|\delta_{i^*}\|_2$ is so much smaller than $\|\delta_i\|_2$ that, conditioned on the event $B_j[i]$ for all $j \in [M]$, $\|\delta_i - \eta v_j\|_2$ is far larger than $\|\delta_{i^*} - \eta v_j\|_2$ for any j.

First note that

$$\Delta^2 \leq \|\delta_i - \delta_{i^*}\|_2^2 = \|\delta_i\|_2^2 + \|\delta_{i^*}\|_2^2 - 2\left\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \right\rangle \|\delta_i\|_2 \|\delta_{i^*}\|_2,$$

so

$$\begin{aligned} 1 - a_{\text{Del}}\frac{\Delta^2}{\sqrt{k}} &\leq \left\langle \widehat{\delta}_i, \widehat{\delta}_{i^*} \right\rangle \\ &\leq \frac{1}{2}\left(\frac{\|\delta_i\|_2}{\|\delta_{i^*}\|_2} + \frac{\|\delta_{i^*}\|_2}{\|\delta_i\|_2}\right) - \frac{\Delta^2}{2\|\delta_i\|_2\|\delta_{i^*}\|_2} \\ &\leq \frac{1}{2}\left(\frac{\|\delta_i\|_2}{\|\delta_{i^*}\|_2} + \frac{\|\delta_{i^*}\|_2}{\|\delta_i\|_2}\right) - \frac{\Delta^2}{a_{\text{scale}}\sqrt{k}}, \end{aligned}$$

where the second step follows by the original assumption in (8.20) that $\|\delta_i\|_2, \|\delta_{i^*}\|_2 \leq a_{\text{scale}}\Delta^2/\sqrt{k}$ for some absolute constant $a_{\text{scale}} > 0$. Recalling the relation between a_{Del} and ν_{CA}^2, we conclude that if we pick $\nu_{CA}^2 < 1/a_{\text{scale}}$, then we get that

$$\frac{1}{2}\left(\frac{\|\delta_i\|_2}{\|\delta_{i^*}\|_2} + \frac{\|\delta_{i^*}\|_2}{\|\delta_i\|_2}\right) \geq 1 + \alpha\frac{\Delta^2}{\sqrt{k}}$$

for absolute constant $\alpha \triangleq \frac{1}{a_{\text{scale}}} - a_{\text{Del}} > 0$, from which we conclude, by taking $\beta = \alpha^{1/2}\Delta k^{-1/4}$ in Fact 8.7.3, that

$$\|\delta_i\|_2 \geq \left(1 + \alpha'\Delta k^{-1/4}\right)\|\delta_{i^*}\|_2 \tag{8.32}$$

for $\alpha' = \alpha^{1/2}/2$ which is increasing in α and therefore in $1/a_{\text{scale}}$. But as we showed in Claim 8.7.10,

$$\|\delta_i - \eta v_j\|_2 \geq \left(1 - \overline{\beta}\Delta^2 k^{-1/2}\right) \cdot \|\delta_i\|_2 \tag{8.33}$$

for all $i \in [k], j \in [M]$ if $B_j[i]$ holds. So by taking a_{scale} sufficiently small relative to $\overline{\beta}$, we get from (8.32) and (8.33) that

$$\|\delta_i - \eta v_j\|_2 \geq \left(1 - \overline{\beta}\Delta^2 k^{-1/2}\right) \cdot \left(1 + \alpha'\Delta k^{-1/4}\right)\|\delta_{i^*}\|_2 \gg \left(1 + c\Delta^2 k^{-1/2}\right)\|\delta_{i^*}\|.$$

But because event C_j involves $A_j[i^*]$ happening, we certainly have that $\|\delta_{i^*}\|_2 \le \|\delta_{i^*} - \eta v_j\|_2^2$, so we are done. $\square$

We now proceed to bound the probabilities of the events $A_j[i], B_j[i], C_j$. There are some minor technical complications from the fact that we don't have exact access to $\mathrm{span}(\{w_i - a_t\})$ which we address now.

Define $\alpha^{(i)}_{\mathrm{svd}} \triangleq \frac{\|\mathbf{U}\delta_i\|_2}{\|\delta_i\|_2}$. By the second part of Lemma 8.7.2, $1 - \delta_{\mathrm{samp}} \le \alpha^{(i)}_{\mathrm{svd}} \le 1$. First note that for any $i \in [k]$,

$$\gamma_i^{(j)} = \langle \widehat{\delta}_i, v_j \rangle = \frac{\langle g, \mathbf{U}\delta_i \rangle}{\|g\mathbf{U}\|_2 \cdot \|\delta_i\|_2} = \left\langle \frac{g}{\|g\|_2}, \frac{\mathbf{U}\delta_i}{\|\delta_i\|_2} \right\rangle = \alpha^{(i)}_{\mathrm{svd}} \left\langle \frac{g}{\|g\|_2}, \frac{\mathbf{U}\delta_i}{\|\mathbf{U}\delta_i\|_2} \right\rangle, \tag{8.34}$$

where $g \sim \mathcal{N}(0, \mathrm{Id}_k)$ and the last step follows by the second part of Lemma 8.7.2. The random variable $\left\langle \frac{g}{\|g\|_2}, \frac{\mathbf{U}\delta_i}{\|\mathbf{U}\delta_i\|_2} \right\rangle$ is merely the correlation of a random unit vector with a fixed unit vector; call this random variable X (clearly it does not depend on the fixed vector).

We can now lower bound the probabilities of $A_j[i^*]$ and $B_j[i]$.

Claim 8.7.12. *For any $j \in [M]$, $Pr[A_j[i^*]] \ge e^{-a_{\mathrm{trials}}\sqrt{k}/\Delta^2}$ for some absolute constant $a_{\mathrm{trials}} > 0$.*

Proof. By (8.34), $\Pr[A_j[i^*]] \ge \Pr[X \ge \nu_A \Delta k^{-1/4}]$. By Corollary 1.3.19, $\Pr[X \ge \nu_A k^{-1/4}] \ge e^{-a_{\mathrm{trials}}\sqrt{k}/\Delta^2}$ for some $a_{\mathrm{trials}} > 0$. $\square$

Claim 8.7.13. *For any $i \in [k]$ and $j \in [M]$, $Pr[B_j[i]] \ge 1 - e^{-\overline{a_{\mathrm{trials}}}\sqrt{k}/\Delta^2}$ for some absolute constant $\overline{a_{\mathrm{trials}}} > a_{\mathrm{trials}}$.*

Proof. By (8.34), $\Pr[B_j[i^*]] \ge \Pr[X \le \nu_B \Delta k^{-1/4}]$. If we take $\delta_{\mathrm{samp}} = 1/\mathrm{poly}(k)$ sufficiently small, then because $\nu_B > \nu_A$, we conclude by Corollary 1.3.19 that $\Pr[X \le \nu_B \Delta k^{-1/4}] \ge 1 - e^{-\overline{a_{\mathrm{trials}}}\sqrt{k}/\Delta^2}$ for some $\overline{a_{\mathrm{trials}}} > a_{\mathrm{trials}}$. $\square$

We next lower bound the probability of event C_j relative to that of $A_j[i^*]$. Equivalently, provided $A_j[i^*]$ happens, we lower bound the conditional probability that the gap of (8.21) is preserved. In this proof, we would like to use the fact that $\delta_i^\perp$ is orthogonal to $\widehat{\delta}_{i^*}$ for all $i \ne i^*$ to argue that $\langle g\mathbf{U}, \delta_i^\perp \rangle$ and $\langle g\mathbf{U}, \delta_{i^*} \rangle$ are independent. Again, this is only true if $\mathbf{U}$ is exactly the projector to the span of $\{w_i - a_t\}$, and we need to argue that it suffices to take $\mathbf{U}$ an approximation to that projector.

Claim 8.7.14. *For any $j \in [M]$, $Pr[C_j] \geq \frac{1}{\text{poly}(k)}\, Pr[A_j[i^*]]$.*

Proof. Let $v = g/\|g\|_2 \in \mathbb{R}^k$. Analogous to (8.34), we can define $\beta^{(i)}_{\text{svd}} \triangleq \frac{\|\mathbf{U}\delta^{\perp}_i\|_2}{\|\delta^{\perp}_i\|_2} \in [1-\delta_{\text{samp}}, 1]$ and write

$$\frac{1}{\|\delta^{\perp}_i\|_2}\langle \delta^{\perp}_i, v_j\rangle = \frac{\langle g, \mathbf{U}\delta^{\perp}_i\rangle}{\|g\mathbf{U}\|_2 \cdot \|\delta^{\perp}_i\|_2} = \left\langle \frac{g}{\|g\|_2}, \frac{\mathbf{U}\delta^{\perp}_i}{\|\delta^{\perp}_i\|_2}\right\rangle = \beta^{(i)}_{\text{svd}}\left\langle v, \frac{\mathbf{U}\delta^{\perp}_i}{\|\mathbf{U}\delta^{\perp}_i\|_2}\right\rangle, \tag{8.35}$$

Define $\rho \triangleq \left\langle \mathbf{U}\delta^{\perp}_i, \frac{\mathbf{U}\delta_{i^*}}{\|\mathbf{U}\delta_{i^*}\|_2}\right\rangle$. By the first part of Lemma 8.7.2, if we take $\delta_{\text{samp}} < 1/k^{100}$, then

$$|\langle \mathbf{U}\delta^{\perp}_i, \mathbf{U}\delta_{i^*}\rangle| \leq 1/k^{100}\|\delta_{i^*}\|_2\|\delta^{\perp}_i\|_2,$$

so we conclude that

$$|\rho| \leq \frac{1/k^{100}\|\delta_{i^*}\|_2\|\delta^{\perp}_i\|_2}{\|\mathbf{U}\delta_{i^*}\|_2} \leq 2(1/k^{100})\|\delta^{\perp}_i\|_2. \tag{8.36}$$

So we may write

$$\mathbf{U}\delta^{\perp}_i = \rho \cdot \frac{\mathbf{U}\delta_{i^*}}{\|\mathbf{U}\delta_{i^*}\|_2} + v'$$

for $v' \in \mathbb{R}^k$ lying in the row span of $\mathbf{U}$ and orthogonal to $\mathbf{U}\delta_{i^*}$.

By Corollary 1.3.20, for any absolute constant $a_{\text{perp}} > 0$, with probability at least

$$\frac{1}{\text{poly}(k)}\Pr\left[\langle v, \mathbf{U}\widehat{\delta}_{i^*}\rangle \geq \frac{\nu_A \Delta k^{-1/4}}{(1-\delta_{\text{samp}})}\right] \geq \frac{1}{\text{poly}(k)}\Pr\left[A_j[i^*]\right],$$

we have that

$$\left\langle v, \frac{\mathbf{U}\widehat{\delta}_{i^*}}{\|\mathbf{U}\widehat{\delta}_{i^*}\|_2}\right\rangle \geq \frac{\nu_A \Delta k^{-1/4}}{(1-\delta_{\text{samp}})} \quad \text{and} \quad \langle v, v'\rangle \leq a_{\text{perp}}\Delta^2 k^{-1/2}\|v'\|_2. \tag{8.37}$$

In particular, if this happens, then by (8.34) we get that

$$\langle v, \widehat{\delta}_{i^*}\rangle \geq \alpha^{(i)}_{\text{svd}}\left\langle v, \mathbf{U}\widehat{\delta}_{i^*}\right\rangle \geq \nu_A \Delta k^{-1/4}$$

and by (8.35),

$$\begin{aligned}\frac{1}{\|\delta_i^\perp\|_2}\langle \delta_i^\perp, v_j\rangle &= \beta_{\text{svd}}^{(i)}\left\langle v, \frac{\mathbf{U}\delta_i^\perp}{\|\mathbf{U}\delta_i^\perp\|_2}\right\rangle \\ &\le \frac{\beta_{\text{svd}}^{(i)}}{\|\mathbf{U}\delta_i^\perp\|_2}(\rho + \langle v, v'\rangle) \\ &\le \frac{\beta_{\text{svd}}^{(i)}}{\|\mathbf{U}\delta_i^\perp\|_2}\left(2(1/k^{100})\|\delta_i^\perp\|_2 + a_{\text{perp}}\Delta k^{-1/4}\|\delta_i^\perp\|_2\right) \\ &= 2(1/k^{100}) + a_{\text{perp}}\Delta k^{-1/4} < 2a_{\text{perp}}\Delta k^{-1/4},\end{aligned}$$

where in the third step we used (8.36), the bound on $\langle v, v'\rangle$ in the event that (8.37) holds, and the fact that $\|v'\|_2 \le \|\mathbf{U}\delta_i^\perp\|_2 \le \|\delta_i^\perp\|_2$. So by taking a_{perp} sufficiently small, we conclude that event C_j occurs in the event that (8.37) holds, which is with probability at least $\frac{1}{\text{poly}(k)}\Pr[A_j[i^*]]$. □

Next, we would like to show that for any $j \in [M]$, if we condition on the events $B_j[i]$ holding for all $i \in [k]$, then the conditional probability of $A_j[i]$ is not much more than that of $A_j[i^*]$. Note that by rotational invariance, these conditional probabilities would be identical if $\mathbf{U}$ were exactly the projector to the span of $\{w_i - a_t\}$, and here it is straightforward to see that it suffices to take $\mathbf{U}$ a sufficiently good approximation to that projector.

Claim 8.7.15. *For any $j \in [M]$ and $i \neq i^*$, if $\delta_{\text{samp}} = 1/\text{poly}(k)$ is sufficiently small, $Pr[A_j[i] \wedge B_j] \le k \cdot Pr[A_j[i^*] \wedge B_j]$.*

Proof. By (8.34), we conclude that

$$\frac{\Pr[A_j[i] \wedge B_j]}{\Pr[A_j[i^*] \wedge B_j]} \le \frac{\Pr[\nu_A \Delta k^{-1/4}/\alpha_{\text{svd}}^{(i)} \le X \le \nu_B \Delta k^{-1/4}/\alpha_{\text{svd}}^{(i)}]}{\Pr[\nu_A \Delta k^{-1/4}/\alpha_{\text{svd}}^{(i^*)} \le X \le \nu_B \Delta k^{-1/4}/\alpha_{\text{svd}}^{(i^*)}]}.$$

which can be upper bounded by k by taking a sufficiently small $\delta_{\text{samp}} = 1/\text{poly}(k)$. □

We can now put all of these probability bounds together to show that if CompareMinVariances succeeds and outputs true on some direction v, the conditional probability that the gap of (8.21) has been preserved is at least $1/\text{poly}(k)$.

Claim 8.7.16. *Let B_v be the event that $Pr[\langle\widehat{\delta}_i, v\rangle] \leq \nu_B \Delta k^{-1/4}$ for all $i \in [k]$. Let $\mathsf{detect-progress}_v$ be the event that B_v occurs and additionally that* CompareMinVariances *succeeds and outputs* True *on direction v. Let $\mathsf{gap-preserved}_v$ be the event that B_v occurs and additionally $\|\delta_i - \eta v\|_2 \geq \left(1 + c\Delta^2 k^{-1/2}\right) \cdot \|\delta_{i^*} - \eta v\|_2$ for all $i \neq i^*$. Then*

$$\Pr_v\left[\mathsf{gap-preserved}_v \mid \mathsf{detect-progress}_v\right] \geq \frac{1}{\mathrm{poly}(k)}$$

Proof. For every $i \in [k]$, let S_i denote the set of all v for which B_v occurs, CompareMinVariances succeeds and outputs True on direction v, and $\|\delta_{i'} - \eta v\|_2 \geq \left(1 + c\Delta^2 k^{-1/2}\right) \cdot \|\delta_i - \eta v\|_2$ for all $i' \neq i$.

To show the claim, it suffices to lower bound the quantity

$$\frac{\Pr_v[S_{i^*}]}{\Pr_v\left[\bigcup_{i=1}^k S_i\right]} \geq \frac{\Pr_v[S_{i^*}]}{\sum_{i=1}^k \Pr_v[S_i]},$$

where the probabilities are over v distributed as $\frac{g\mathbf{U}}{\|g\mathbf{U}\|_2}$ for $g \sim \mathcal{N}(0, \mathrm{Id}_k)$, and where the inequality follows by a union bound. Fix any $j \in [M]$. By Corollary 8.7.9, if $v \in S_i$, then it is part of event $(A_j[i] \wedge B_j)$. By Corollary 8.7.8 and Lemma 8.7.11, if v is part of event $(C_j \wedge B_j)$, then $v \in S_{i^*}$. We conclude that

$$\frac{\Pr_v[S_{i^*}]}{\Pr_v\left[\bigcup_{i=1}^k S_i\right]} \geq \frac{\Pr[C_j \wedge B_j]}{\sum_{i=1}^k \Pr[A_j[i] \wedge B_j]} \geq \frac{\frac{1}{2}\Pr[A_j[i^*] \wedge B_j]}{\sum_{i=1}^k \Pr[A_j[i] \wedge B_j]} \geq \frac{1/\mathrm{poly}(k)}{1 + k(k-1)} = \frac{1}{\mathrm{poly}(k)},$$

where the second step follows from Claim 8.7.14 and the third step follows from Claim 8.7.15. □

We are now ready to finish the proof of Lemma 8.7.4. First, condition on the event that all M runs of CompareMinVariances are successful, which happens with probability at least $1-\delta/3$. The probability that $B_j[i]$ holds for all $i \in [k], j \in [M]$ is at least $1 - kMe^{-\overline{a}_{\mathrm{trials}}\sqrt{k}/\Delta^2}$, by Claim 8.7.13. Condition on this happening. The probability that $\mathsf{detect-progress}_{v_j}$ occurs for some $j \in [M]$ is at least the probability that $A_j[i^*]$ occurs for some $j \in [M]$, and this is at least $1 - \left(1 - e^{-a_{\mathrm{trials}}\sqrt{k}/\Delta^2}\right)^M \geq 1 - e^{-Me^{-a_{\mathrm{trials}}\sqrt{k}/\Delta^2}}$ by Claim 8.7.12. So by taking $M = e^{-a_{\mathrm{trials}}\sqrt{k}/\Delta^2}\ln(3/\delta)$, by a union bound we conclude that with probability at least $1-\delta$, every run of CompareMinVariances succeeds, $B_j[i]$ holds for every $i \in [k], j \in [M]$, and furthermore there is some j for which $\mathsf{detect}-\mathsf{progress}_{v_j}$ occurs.

If $\mathsf{detect}-\mathsf{progress}_{v_j}$ occurs for some j, then for that particular j, $\mathsf{gap}-\mathsf{preserved}_{v_j}$ holds with probability at least $1/\mathrm{poly}(k)$. □

8.7.2 Initializing With a Gap

A key assumption in Lemma 8.7.4 is that there is a gap between $\|w_{i^*} - a_t\|_2$ and all other $\|w_i - a_t\|_2$. We next show that this assumption can be made to hold when $t = 0$. The high-level structure of the proof will be very similar to that of Lemma 8.7.11.

Lemma 8.7.17. *There is a constant $\tau'_{\text{gap}} > 0$ such that for all $c' < \tau'_{\text{gap}}$, the following holds for any sufficiently small $\upsilon_* = \mathrm{poly}(\Delta, 1/k)$.*

Fix any $i^* \in [k]$ and suppose that $\|w_{i^*}\|_2 \geq \underline{\sigma}$ for some $\underline{\sigma} > 0$. Let

$$\mathcal{S} \triangleq \{\underline{\sigma} \cdot k^{1/4}, \underline{\sigma} \cdot (1+\upsilon_*) \cdot k^{1/4}, \underline{\sigma} \cdot (1+\upsilon_*)^2 \cdot k^{1/4}, ..., k^{1/4}\}. \tag{8.38}$$

Then any $i \neq i^*$ and $v \in \mathbb{R}^d$, let $\mathcal{E}_v[i]$ denote the event that

$$\|w_i - v\|_2^2 \geq \left(1 + c' \cdot \frac{\Delta^2}{\sqrt{k}}\right) \cdot \|w_{i^*} - v\|_2^2$$

and define $\mathcal{E}_v$ to be the event that $\mathcal{E}_v[i]$ occurs simultaneously for all $i \neq i^*$. There exists $\alpha \in \mathcal{S}$ for which

$$\underset{\|v\|_2 = \alpha}{Pr}[\mathcal{E}_v] \geq \exp(-O(\sqrt{k}/\Delta^2)),$$

where the probability is over v a Haar-random vector in $\mathbb{R}^d$ with norm α.

Proof. By design, there must exist an $\alpha \in \mathcal{S}$ for which $\alpha = (1+\upsilon) \cdot \|w_{i^*}\|_2 \cdot k^{1/4}$ for $\upsilon \in [-\upsilon_*, \upsilon_*]$. Let v be a random vector with norm α.

Define $\widehat{w}_i = w_i/\|w_i\|_2$. For every $i \neq i^*$, define $w_i^\perp \triangleq \widehat{w}_i - \langle \widehat{w}_{i^*}, \widehat{w}_i \rangle \widehat{w}_{i^*}$. Finally, repurposing notation from the proof of Lemma 8.7.4, let $\gamma_i = \langle \widehat{w}_i, v/\|v\|_2 \rangle$. Also, let

$\rho_i \triangleq \|w_i\|_2/\|w_{i^*}\|_2$. Under this notation, we see that

$$\frac{\|w_i - v\|_2^2}{\|w_{i^*} - v\|_2^2} = \frac{\|w_i\|_2^2 + \alpha^2 - 2\alpha\gamma_i\|w_i\|_2}{\|w_{i^*}\|_2^2 + \alpha^2 - 2\alpha\gamma_{i^*}\|w_{i^*}\|_2}$$
$$= \frac{\rho_i^2 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)\rho_i k^{1/4}\gamma_i}{1 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)k^{1/4}\gamma_{i^*}}. \tag{8.39}$$

Using similar terminology as in the proof of Lemma 8.7.4, define the following two types of events over the random vector v:

1. Let $B[i]$ be the event that $\gamma_i \le k^{-1/4} \cdot (1 + c\Delta^2)$ for some absolute constant $c > 0$.

2. Let C be the event that $\gamma_{i^*} \ge k^{-1/4}$ and $\langle v, w_i^\perp \rangle \le \nu_{\mathrm{perp}} k^{-1/2}\|w_i^\perp\|_2$ for all $i \neq i^*$, for some absolute constant ν_{perp}.

The main step will be to show that these events imply $\mathcal{E}_v$.

Claim 8.7.18. *Let $i \neq i^*$. If events $B[i]$ and C occur, then $\mathcal{E}_v[i]$ occurs.*

Proof. Henceforth, condition on $B[i]$, $B[i^*]$, and C occurring.

There are two cases to consider: either $\|w_{i^*}\|_2$ and $\|w_i\|_2$ are quite different, or they are relatively similar.

It turns out that our choice $\alpha = (1+\upsilon) \cdot \|w_{i^*}\|_2 \cdot k^{1/4}$ will allow us to handle the former case quite easily. Indeed, we first show that if $\|w_i\|_2$ is not $(1 \pm O(\Delta))$-close to $\|w_{i^*}\|_2$, then $\mathcal{E}_v[i]$ occurs.

Claim 8.7.19. *Define the interval*

$$\mathcal{I} \triangleq \left[1 - 2c^{1/2}\Delta, 1 + 2c^{1/2}\Delta\right].$$

Then

$$\frac{\|w_i - v\|_2^2}{\|w_{i^*} - v\|_2^2} \ge 1 + \frac{c\Delta^2}{\sqrt{k}}.$$

Proof. From (8.39) and events $B[i]$ and C we get

$$\frac{\|w_i - v\|_2^2}{\|w_{i^*} - v\|_2^2} \geq \frac{\rho_i^2 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)(1+c\Delta^2)\rho_i}{1 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)}. \tag{8.40}$$

If we define $\rho'_i = \rho_i - 1$, we can rewrite (8.40) as

$$\begin{aligned} &1 + \frac{2\rho'_i + \rho'^2_i - 2c\Delta^2(1+\upsilon) - 2(1+\upsilon)(1+c\Delta^2)\rho'_i}{1 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)} \\ &= 1 + \frac{\rho'^2_i - 2c\Delta^2(1+\upsilon) - 2(\upsilon + c\Delta^2 + c\upsilon\Delta^2)\rho'_i}{1 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)}. \end{aligned}$$

For $\upsilon = \mathrm{poly}(\Delta, 1/k)$ sufficiently small, note that if $\rho'^2_i \geq 4c\Delta^2$, then the numerator is at least $c\Delta^2$. And for such an υ,

$$1 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon) \leq \sqrt{k}, \tag{8.41}$$

completing the proof of the claim. □

Next we show that if $\|w_i\|_2$ *is* $(1 \pm O(\Delta))$-close to $\|w_{i^*}\|_2$ and furthermore v is significantly more correlated with $\widehat{w}_{i^*}$ than with any other $\widehat{w}_i$, then $\mathcal{E}_v[i]$ occurs.

Claim 8.7.20. *Let $i \neq i^*$. If $\rho_i \in \mathcal{I}$ and*

$$\gamma_i \leq \gamma_{i^*}(1 - \omega) \tag{8.42}$$

for $\omega \triangleq 2c^2\Delta^4 + c\Delta^2$, then if events $B[i], B[i^*], C$ all occur, then

$$\frac{\|w_i - v\|_2^2}{\|w_{i^*} - v\|_2^2} \geq 1 + \frac{c\Delta^2}{\sqrt{k}}.$$

Proof. From (8.39) and (8.42) we get that

$$\frac{\|w_i - v\|_2^2}{\|w_{i^*} - v\|_2^2} \geq \frac{\rho_i^2 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)\rho_i k^{1/4}\gamma_{i^*}(1-\omega)}{1 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)k^{1/4}\gamma_{i^*}}$$

$$= \frac{\rho_i^2 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)\rho_i k^{1/4}\gamma_{i^*}}{1 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)k^{1/4}\gamma_{i^*}} + \frac{2\omega(1+\upsilon)\rho_i k^{1/4}\gamma_{i^*}}{1 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)k^{1/4}\gamma_{i^*}} \tag{8.43}$$

Next, note that the quantity $\rho_i^2 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)\rho_i k^{1/4}\gamma_{i^*}$, as a function of ρ_i, is minimized by $\rho_i = (1+\upsilon) \cdot k1/4\gamma_{i^*}$, in which case it equals

$$(1+\upsilon)^2\sqrt{k} \cdot (1 - \gamma_{i^*}^2).$$

We conclude that the first of the two terms in (8.43) is at least

$$\frac{(1+\upsilon)^2\sqrt{k} \cdot (1-\gamma_{i^*}^2)}{1 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)k^{1/4}\gamma_{i^*}} = 1 - \frac{(1 - (1+\upsilon)k^{1/4}\gamma_{i^*})^2}{1 + (1+\upsilon)^2\sqrt{k} - 2(1+\upsilon)k^{1/4}\gamma_{i^*}}.$$

Recall that because of event $B[i^*]$, we know $\gamma_{i^*} \le k^{-1/4}(1 + c\Delta^2)$, so

$$(1 - (1+\upsilon)k^{1/4}\gamma_{i^*})^2 \le (|\upsilon| + c\Delta^2 - c|\upsilon|\Delta^2)^2 \le 2c^2\Delta^4 \tag{8.44}$$

for υ sufficiently small. On the other hand, the numerator of the second of the two terms in (8.43) is

$$2\omega(1+\upsilon)\rho_i k^{1/4}\gamma_{i^*} \ge \omega, \tag{8.45}$$

because $\gamma_{i^*} \ge k^{-1/4}$ by event C, and because $(1+\upsilon)\rho_i \ge 1/2$ when υ is sufficiently small and $\rho_i \in \mathcal{I}$. We conclude from (8.41), (8.43), (8.44), and (8.45) that

$$\frac{\|w_i - \upsilon\|_2^2}{\|w_{i^*} - \upsilon\|_2^2} \ge 1 + \frac{\omega - 2c^2\Delta^4}{\sqrt{k}}.$$

In particular, if we took $\omega = 2c^2\Delta^4 + c\Delta^2$, then again we would have $\frac{\|w_i - \upsilon\|_2^2}{\|w_{i^*} - \upsilon\|_2^2} \ge 1 + \frac{c\Delta^2}{\sqrt{k}}$. □

Finally, we show that if $\rho_i \in \mathcal{I}$, then events $B[i]$ and C imply (8.42). We proceed in a manner similar to the proof of (8.28) in Lemma 8.7.11. As with that proof, the intuition is that if the normalized vectors $\widehat{w}_i$ and $\widehat{w}_{i^*}$ are somewhat separated on the unit sphere, then the upper bound on $\langle \upsilon, w_i^{\perp} \rangle$ will ensure the existence of a sizable cone around w_{i^*} for which any υ inside that cone is much closer to w_{i^*} than to w_i. And if instead $\widehat{w}_i$ and $\widehat{w}_{i^*}$ are not separated, the fact that their non-normalized counterparts w_i and w_{i^*} are separated implies that $\widehat{w}_i$ and $\widehat{w}_{i^*}$ are nearly collinear and thus too separated for $\rho_i \in \mathcal{I}$ to hold.

Claim 8.7.21. *Let $i \neq i^*$. If $\rho_i \in \mathcal{I}$ and events $B[i]$ and C occur, then (8.42) must hold.*

Proof. As with (8.29), note that

$$\gamma_i = \langle \widehat{w}_i, \widehat{w}_{i^*} \rangle + \left\langle w_i^\perp, v \right\rangle .$$

Now if $\langle \widehat{w}_i, \widehat{w}_{i^*} \rangle \leq 0$, then by event C, $\gamma_i \leq \langle w_i^\perp, v\rangle \leq \nu_{\text{perp}} k^{-1/2} \|w_i^\perp\|_2 \ll \gamma_{i^*}(1-\omega)$, and we'd be done. On the other hand, if $\langle \widehat{w}_i, \widehat{w}_{i^*} \rangle > 0$, then we get that

$$\gamma_i \leq \langle \widehat{w}_i, \widehat{w}_{i^*} \rangle + \nu_{\text{perp}} k^{-1/2} \|w_i^\perp\|_2.$$

In this case, to show the desired inequality (8.42), it would suffice to show that

$$\gamma_{i^*} \left(1 - \omega - \langle \widehat{w}_i, \widehat{w}_{i^*} \rangle\right) \geq \nu_{\text{perp}} k^{-1/2} \|w_i^\perp\|_2.$$

In particular, because $\gamma_{i^*} \geq k^{-1/4}$, we just need to show that

$$1 - \omega - \langle \widehat{w}_i, \widehat{w}_{i^*} \rangle \geq \nu_{\text{perp}} k^{-1/4} \|w_i^\perp\|_2. \tag{8.46}$$

After squaring both sides of (8.46), making the substitution $\|w_i^\perp\|_2^2 = 1 - \left\langle w_i^\perp, w_{i^*}^\perp \right\rangle^2$, and rearranging, (8.46) becomes

$$\left(1 - \omega - \langle \widehat{w}_i, \widehat{w}_{i^*} \rangle\right)^2 - \nu_{\text{perp}}^2 k^{-1/2} \cdot \left(1 - \left\langle w_i^\perp, w_{i^*}^\perp \right\rangle^2\right) \geq 0 \tag{8.47}$$

This is merely a univariate inequality for a quadratic polynomial in $\langle \widehat{w}_i, \widehat{w}_{i^*} \rangle$. The roots of this polynomial are given by

$$\langle \widehat{w}_i, \widehat{w}_{i^*} \rangle = \frac{1 - \omega \pm \sqrt{\frac{\nu_{\text{perp}}^4}{k} + \frac{2\nu_{\text{perp}}^2 \omega}{\sqrt{k}} - \frac{\nu_{\text{perp}}^2 \omega^2}{\sqrt{k}}}}{1 + \nu_{\text{perp}}^2 / \sqrt{k}}.$$

Observe that

$$\sqrt{\frac{\nu_{\text{perp}}^4}{k} + \frac{2\nu_{\text{perp}}^2\omega}{\sqrt{k}} - \frac{\nu_{\text{perp}}^2\omega^2}{\sqrt{k}}} = \omega \cdot \sqrt{\left(1 + \frac{\nu_{\text{perp}}^2}{\omega\sqrt{k}}\right)^2 - \left(1 + \frac{\nu_{\text{perp}}^2}{\sqrt{k}}\right)}$$
$$\leq \omega \cdot \sqrt{\frac{2\nu_{\text{perp}}^2}{\omega\sqrt{k}} + \frac{\nu_{\text{perp}}^4}{\omega^2 k}} \leq \frac{2\nu_{\text{perp}} \cdot \omega^{1/2}}{k^{1/4}} \leq a_{\text{arb}}\omega,$$

where the last step holds for any absolute constant $a_{\text{arb}} > 0$ for sufficiently large k. We see that the inequality (8.47) is satisfied provided that $\langle \widehat{w}_i, \widehat{w}_{i^*}\rangle$ lies outside the interval

$$\mathcal{J} \triangleq \left[\frac{1-(1+a_{\text{arb}})\omega}{1+\nu_{\text{perp}}^2/\sqrt{k}}, \frac{1-(1-a_{\text{arb}})\omega}{1+\nu_{\text{perp}}^2/\sqrt{k}}\right] \subset \left[1-(1+2a_{\text{arb}})\omega, 1-(1-2a_{\text{arb}})\omega\right].$$

It remains to show that under the hypotheses of the claim, we cannot have $\langle \widehat{w}_i, \widehat{w}_{i^*}\rangle \in \mathcal{J}$. This is where we will crucially use the fact that $\|w_i - w_{i^*}\|_2 \geq \Delta$.

Suppose to the contrary that $\langle \widehat{w}_i, \widehat{w}_{i^*}\rangle \in \mathcal{J}$. In particular, this implies

$$\langle \widehat{w}_i, \widehat{w}_{i^*}\rangle \geq 1 - (1+2a_{\text{arb}})\omega.$$

Now note that

$$\Delta^2 \leq \|w_i - w_{i^*}\|_2^2 = \|w_i\|_2^2 + \|w_{i^*}\|_2^2 - 2\langle \widehat{w}_i, \widehat{w}_{i^*}\rangle \|w_i\|_2 \|w_{i^*}\|_2,$$

so

$$1-(1+2a_{\text{arb}})\omega \leq \langle \widehat{w}_i, \widehat{w}_{i^*}\rangle \leq \frac{1}{2}\left(\rho_i + 1/\rho_i\right) - \frac{\Delta^2}{2\|w_i\|_2\|w_{i^*}\|_2} \leq \frac{1}{2}\left(\rho_i + 1/\rho_i\right) - \Delta^2/2.$$

For c sufficiently small, $\Delta^2/2 - (1+2a_{\text{arb}})\omega \geq \Delta^2/4$, so by taking $\beta = \Delta/2$ in Fact 8.7.3 we conclude that $\rho_i \notin [1-\Delta/4, 1+\Delta/4]$. We get a contradiction upon noting that if $2c^{1/2} < 1/4$, then $\rho_i \notin \mathcal{I}$. □

The proof of Claim 8.7.18 now follows. Take $\tau'_{\text{gap}} = c$ in the statement of Lemma 8.7.17. Then for every $i \neq i^*$, either $\rho_i \in \mathcal{I}$, in which case we are done by Claim 8.7.19. Otherwise, $\rho_i \notin \mathcal{I}$, in which case Claim 8.7.21 implies (8.42) holds, and then we are done by Claim 8.7.21.

□

To complete the proof, we must show that the probability that $B[i]$ and C occur simultaneously for all $i \in [k]$ is at least $\exp(-O(\sqrt{k}/\Delta^2))$. The proofs for these facts are essentially identical to those of Claims 8.7.12, 8.7.13, and 8.7.14 in the proof of Lemma 8.7.4, so we omit them.

Claim 8.7.22. *For any $i \in [k]$, $Pr[B[i]] \geq 1 - e^{-\overline{a}\sqrt{k}/\Delta^2}$ for some absolute constant $\overline{a} > 0$.*

Claim 8.7.23. *For any $i \in [k]$, $Pr[C] \geq \frac{1}{\mathrm{poly}(k)} e^{-\underline{a}\sqrt{k}/\Delta^2}$ for some absolute constant $\underline{a} > 0$ such that $\overline{a} - \underline{a}$ is nonnegative and strictly increasing in Δ.*

Lemma 8.7.17 now follows by a union bound: the probability that all $B[i]$ occur is at least $1 - k \cdot e^{-\overline{a}\sqrt{k}/\Delta^2}$, and the probability that C occurs is $\frac{1}{\mathrm{poly}(k)} e^{-\underline{a}\sqrt{k}/\Delta^2}$, so the probability all of these events occur is at least $\frac{1}{\mathrm{poly}(k)} e^{-\underline{a}\sqrt{k}/\Delta^2} - k \cdot e^{-\overline{a}\sqrt{k}/\Delta^2} = \exp(-O(\sqrt{k^2}/\Delta))$, and then we are done by Claim 8.7.18. □

Lastly, we remark that Lemma 8.7.17 only applies to w_{i^*} for which $\|w_{i^*}\|_2 \geq \underline{\sigma}$. We could for instance take $\underline{\sigma} = \varepsilon/4$ and this would not affect the asymptotics of our runtime. Now for regressors w_i whose norm is less than $\varepsilon/4$, we can simply output an arbitrary vector a of norm $\varepsilon/4$ as an $\varepsilon/2$-close estimate, by triangle inequality. We can also easily check whether there is indeed such a short regressor w_i, e.g. by estimating the minimum variance of the univariate mixture $\mathcal{F}$ given by sampling $(x, y) \sim \mathcal{D}$ and computing $y - \langle a, x \rangle$ (see CheckOutcome).

8.7.3 Algorithm Specification

We are now ready to describe our algorithm LearnWithNoise for learning all components of $\mathcal{D}$. The key subroutines are:

- OptimisticDescent (Algorithm 33): the pseudocode for this is nearly identical to that of FourierMomentDescent, except OptimisticDescent additionally takes as input an initialization, has a different output guarantee, and has slightly different parameters which are tuned to fit the regime of Lemma 8.7.4.

- CheckOutcome (Algorithm 35): CheckOutcome is used to check whether a given estimate is close to any regressor of $\mathcal{D}$. This only needs to be used to check whether there exists a short regressor, as discussed at the end of the previous Section 8.7.2.

8.7.4 Proof of Correctness

We first give a proof of correctness for CheckOutcome.

Lemma 8.7.24. *Let $v \in \mathbb{S}^{d-1}$ and $\mathcal{D}$ be a mixture of linear regressions with noise rate $\varsigma > \Delta p_{\min}^3$, and let $\varepsilon > 2\varsigma$. If there is some component v_{i^*} for which $\|v - v_{i^*}\|_2 \in [-\varepsilon, \varepsilon]$, then* CheckOutcome*$(\mathcal{D}, v, \varepsilon, \delta)$ (Algorithm 37) returns* True *with probability at least $1-\delta$. Otherwise, if $\|v - v_i\|_2 > 2\varepsilon$ for all $i \in [k]$, then* CheckOutcome *returns* False *with probability at least $1-\delta$. Furthermore,* CheckOutcome *has time and sample complexity*

$$\widetilde{O}\left(kp_{\min}^{-4}\ln(1/\delta) \cdot \operatorname{poly}\left(\ln(1/p_{\min}), \ln(1/\Delta)\right)^{\ln(1/p_{\min})}\right).$$

Proof. As usual, $\mathcal{F}$ is a mixture of univariate Gaussians with variances $\{\varsigma^2 + \|w_i - v\|_2^2\}$. By Corollary 8.5.8,

$$\sigma^* \in [0.9^2, 1.1^2] \cdot (\min_{i\in[k]} \|w_i - v\|_2^2 + \varsigma^2).$$

If $\min_{i\in[k]}\|w_i - v\|_2^2 \leq \varepsilon^2$, then we have that

$$(\sigma^*)^2 \leq 1.1^2\left(\varepsilon^2 + \varsigma^2\right) \leq 1.21 \cdot (1 + 1/4)\varepsilon^2 \leq 2\varepsilon^2.$$

If $\min_{i\in[k]}\|w_i - v\|_2^2 \geq 4\varepsilon^2$, then we have that

$$(\sigma^*)^2 \geq 0.9^2\left(4\varepsilon^2 + \varsigma^2\right) \geq 0.9^2 \cdot 4\varepsilon^2 \geq 3\varepsilon^2,$$

which completes the proof. □

We can now prove correctness of LearnWithNoise.

Lemma 8.7.25. *Let $a_{\text{noise}} > 0$ be the constant defined in Lemma 8.7.4, and let $\varepsilon > 0$. Let $\mathcal{D}$ be a mixture of spherical linear regressions with mixing weights $\{p_i\}$, directions $\{w_i\}$,*

Algorithm 33: $\textsc{OptimisticDescent}(\mathcal{D}, a_0, \delta)$

```
Input: Sample access to mixture of linear regressions $\mathcal{D}$ with noise rate
       $\varsigma^2 < a_{\text{noise}} \cdot \varepsilon^2$, initial vector $a_0 \in \mathbb{R}^d$, $\delta > 0$
Output: $a_T \in \mathbb{R}^d$ such that $\|w_{i^*} - a_T\|_2 \le \varepsilon$ with probability at least $1 - \delta$ and,
        with probability at least $\exp(-\widetilde{O}(\sqrt{k}/\Delta^2))$, additionally $\mathcal{E}_{a_T}[i^*]$ holds with
        probability $\exp(-\widetilde{O}(\sqrt{k}/\Delta^2))$ provided $\mathcal{E}_{a_0}[i^*]$ holds for some $i^* \in [k]$
$T \leftarrow \Omega(\sqrt{k} \cdot \ln(1/\varepsilon))$.
$\delta' \leftarrow \frac{\delta}{4T}$.
$M \leftarrow e^{-a_{\text{trials}}\sqrt{k}/\Delta^2} \ln(3/\delta)$.
$\delta'' \leftarrow \frac{\delta}{4MT}$.
$\overline{\sigma} \leftarrow a_{\text{scale}} \cdot k^{-1/4}$.
$\underline{\sigma} \leftarrow \varepsilon/3$.
for $t = 0$ to $T - 1$ do
    Let $\mathcal{F}_t$ be the univariate mixture of Gaussians which can be sampled from by
     drawing $(x, y) \sim \mathcal{D}$ and computing $y - \langle x, a_t \rangle$.
    $p \leftarrow 20 \ln\left(\frac{3}{2p_{\min}}\right) + 1$.
    $\kappa_1 \leftarrow (\underline{\beta}' - 3c/2)\Delta^2 k^{-1/2}$, $\kappa_2 \leftarrow (\underline{\beta}' - c/2)\Delta^2 k^{1/2}$
    $\underline{\sigma}_t^{\text{sharp}} \leftarrow \textsc{EstimateMinVariance}(\mathcal{F}_t, \overline{\sigma}, \underline{\sigma}, p, \delta')$.    // Algorithm 29
    if $\underline{\sigma}_t^{\text{sharp}} < 0.99\varepsilon$ then
        return $a_t$.
    Draw $N_1 \triangleq \widetilde{\Omega}\left(\frac{\overline{\sigma}^2}{(\underline{\sigma}_t^{\text{sharp}})^2} \cdot p_{\min}^{-2} \cdot \text{poly}(k) \cdot d \cdot \ln(k/\delta')\right)$ i.i.d. samples $\{(x_i, y_i)\}_{i \in [N]}$
     from $\mathcal{D}$ and form the matrix $\widehat{\mathbf{M}}_{a_t}^{(N)}$.
    $U_t \leftarrow \textsc{ApproxBlockSVD}\ (\widehat{\mathbf{M}}_{a_t}^{(N)}, 1/\text{poly}(k))$    // Lemma 8.5.11
    for $j \in [M]$ do
        Sample $g_t^{(j)} \sim \mathcal{N}(0, \text{Id}_k)$ and define $v_t^{(j)} = \frac{U_t g_t^{(j)}}{\|U_t g_t^{(j)}\|_2} \in \mathbb{S}^{d-1}$.
        ${a'}_t^{(j)} \leftarrow a_t + \eta_t v_j$ for $\eta_t \triangleq a_{\text{LR}} \cdot \Delta \cdot \underline{\sigma}_t^{\text{sharp}} \cdot k^{-1/4}$
        Let ${\mathcal{F}'}_t^{(j)}$ be the univariate mixture of Gaussians which can be sampled from
         by drawing $(x, y) \sim \mathcal{D}$ and computing $y - \langle x, {a'}_t^{(j)} \rangle$.
        if $\textsc{CompareMinVariances}(\mathcal{F}_t, {\mathcal{F}'}_t^{(j)}, \overline{\sigma}, \underline{\sigma}, \kappa_1, \kappa_2, \delta'') = \text{True}$ then
            $a_{t+1} \leftarrow {a'}_t^{(j)}$
            Break
return $a_T$.
```

```
Algorithm 34: LearnWithNoise($\mathcal{D}, \delta, \varepsilon$)
  Input: Sample access to mixture of linear regressions $\mathcal{D}$ with regressors $\{w_1, ..., w_k\}$
         and separation $\Delta$ and noise rate $\varsigma$, failure probability $\delta$, error $\varepsilon < \Delta/4$
  Output: List of vectors $\mathcal{L} \triangleq \{\widetilde{w}_1, ..., \widetilde{w}_k\}$ for which there is a permutation
          $\pi : [k] \to [k]$ for which $\|w_i - w_{\pi(i)}\|_2 \le \varepsilon$ for all $i \in [k]$, with probability at
          least $1 - \delta$
  $\underline{\sigma} \leftarrow \varepsilon/4$.
  $\mathcal{L} \leftarrow \emptyset$.
  Set $v_*$ to be a sufficiently small $\mathrm{poly}(\Delta, 1/k)$ and define the mesh $\mathcal{S}$ via (8.38)
  Set $v_{\mathrm{tiny}}$ to be a random vector of norm $\varepsilon/4$
  if CheckOutcome($\mathcal{D}, v_{tiny}, \underline{\sigma}, \delta/5$) = True then
      Add $v_{\mathrm{tiny}}$ to $\mathcal{L}$
  $W \leftarrow \exp(O(\sqrt{k}/\Delta^2)) \cdot \ln(2k/\delta)$  $\delta^* \leftarrow \frac{\delta}{2|\mathcal{S}|W}$
  for $\alpha \in \mathcal{S}$ do
      for $0 \le i < W$ do
          Let $v$ be a Haar-random vector in $\mathbb{R}^d$ of norm $\alpha$.
          $\widetilde{v} \leftarrow$ OptimisticDescent($\mathcal{D}, v, \delta^*$)
          if $\|\widetilde{v} - \widetilde{w}\|_2 > 2\varepsilon$ for all $\widetilde{w} \in \mathcal{L}$ then
              Add $\widetilde{v}$ to $\mathcal{L}$
  return $\mathcal{L}$.
```

```
Algorithm 35: CheckOutcome($\mathcal{D}, v, \varepsilon, \delta$)
  Input: Sample access to mixture of linear regressions $\mathcal{D}$ with noise rate $\varsigma$, vector
         $v \in \mathbb{R}^d$, threshold $\varepsilon > 0$, failure probability $\delta$
  Output: True if $\min_{i\in[k]}\|w_i - v\|_2 \le \varepsilon$, False if $\min_{i\in[k]}\|w_i - v\|_2 \ge 2\varepsilon$, with
          probability at least $1 - \delta$
  Let $\mathcal{F}$ be the univariate mixture of Gaussians which can be sampled from by
    drawing $(x, y) \sim \mathcal{D}$ and computing $y - \langle v, x\rangle$.
  $p \leftarrow 20 \ln\left(\frac{3}{2p_{\min}}\right) + 1$.
  $\sigma^* \leftarrow$ EstimateMinVariance($\mathcal{F}, 4, \varsigma, p, \delta$).
  if $(\sigma^*)^2 \le 2\varepsilon^2$ then
      return True.
  return False.
```

and noise rate ς. For any $\varepsilon, \delta > 0$ and $\varsigma^2 \leq \varepsilon^2/10$, with probability at least $1-\delta$, LearnWithNoise*$(\mathcal{D}, \delta, \varepsilon)$ (Algorithm 34) outputs a list of vectors $\mathcal{L} = \{w_1, ..., w_k\}$ such that there exists a permutation $\pi : [k] \to [k]$ for which $\|w_i - w_{\pi(i)}\|_2 \leq \varepsilon$ for all $i \in [k]$.*

Proof. Let $v \in \mathbb{R}^d$, and consider a run of OptimisticDescent$(\mathcal{D}, v, \delta^*)$. Let a_t be the iterate at time t in this run. Let $\rho = 1.1/0.9$.

We first note that if OptimisticDescent breaks out at Line 13, the vector it returns is close to some component of $\mathcal{D}$.

Claim 8.7.26. *If for some $0 \leq t < T$,* OptimisticDescent *(Algorithm 33) breaks out at Line 13 and outputs a_t, then $\min_i \|w_i - a_t\|_2 \leq \varepsilon$.*

Proof. If OptimisticDescent (Algorithm 33) breaks out at Line 13, it is because $\sigma_t^* \leq 0.99\varepsilon$. This implies that $\min_i \|w_i - a_t\|^2 + \varsigma^2 \leq (\sigma_t^*)^2/0.99^2 \leq \varepsilon^2$, so $\min_i \|w_i - a_t\|_2 \leq \varepsilon$ as claimed. □

Next, we show that if a_t is still somewhat far from any component, with high probability over the next iteration either OptimisticDescent will break out at Line 13, or the progress measure will contract.

Claim 8.7.27. *Let i^* be the index minimizing $\|w_i - a_t\|_2$. If $\|w_{i^*} - a_t\|_2^2 \geq \varepsilon^2/2$, then with probability at least $1 - \delta^*/T$ over the next iteration of* OptimisticDescent *(Algorithm 33), either of two things will happen:*

1. $\varepsilon^2/16 \leq \min_{i \in [k]} \|w_i - a_{t+1}\|_2^2 \leq \varepsilon^2/\rho^2 - \varsigma^2$.

2. $\min_{i \in [k]} \|w_i - a_{t+1}\|_2^2 \geq \varepsilon^2/2$ *and*

$$\sigma_{t+1}^2 \leq \left(1 - \underline{\beta} \Delta^2/\sqrt{k}\right) \cdot \sigma_t^2.$$

Condition on either of these outcomes happening. Then additionally, if event $\mathcal{E}_{a_t}[i^*]$ (see Lemma 8.7.17) holds, then with probability at least $1/\mathrm{poly}(k)$, $\mathcal{E}_{a_{t+1}}[i^*]$ holds.

Proof. Condition on outcomes 1), 2), and 3) of Lemma 8.7.4, which all happen with probability at least $1 - \delta^*/T$.

Now if $\min_{i\in[k]}\|w_i - a_{t+1}\|_2^2 \geq \varepsilon^2/2$, then 2) is just a consequence of outcomes 1) and 2) of Lemma 8.7.4.

If $\min_{i\in[k]}\|w_i - a_{t+1}\|_2^2 \leq \varepsilon^2/2$, then by outcome 3) of Lemma 8.7.4, $\sigma_{t+1} \geq (1 - \overline{\beta}\Delta^2/\sqrt{k})\sigma_t \geq 0.99\sigma_t \geq \varepsilon^2/16$.

The last part of the claim is a consequence of outcome 4) of Lemma 8.7.4, which happens in addition to outcomes 1), 2), and 3) with probability $1/\mathrm{poly}(k)$, provided $\mathcal{E}_{a_t}[i^*]$ occurs. □

Next we show that if at some time t there is an $i \in [k]$ for which $\|w_i - a_t\|_2^2 \leq \varepsilon^2/\rho^2 - \varsigma^2$, then OptimisticDescent will break out at Line 13 and correctly output a_t.

Claim 8.7.28. *If for some $0 \leq t < T$ we have*

$$\varepsilon^2/16 \leq \min_{i\in[k]}\|w_i - a_t\|_2^2 \leq \varepsilon^2/\rho^2 - \varsigma^2, \tag{8.48}$$

then OptimisticDescent *(Algorithm 33) breaks out at Line 13 and returns a_t.*

Proof. The lower bound in (8.48) implies that the $\underline{\sigma} = \varepsilon/4$ which is passed to EstimateMinVariance is a valid lower bound for σ_t.

The upper bound in (8.48) implies that

$$(\sigma_t^*)^2 \leq 1.21\sigma_t^2 \leq 1.21 \cdot (\min_{i\in[k]}\|w_i - a_t\|_2^2 + \varsigma^2) \leq 0.99^2\varepsilon^2,$$

so OptimisticDescent breaks out at Line 13 and outputs a_t. The bound on $\|w_i - a_{t+1}\|_2$ immediately follows from (8.48). □

Claims 8.7.26, 8.7.27, and 8.7.28 imply that with high probability the output of OptimisticDescent (Algorithm 33) is close to some component of $\mathcal{D}$.

Claim 8.7.29. *For any $v \in \mathbb{R}^d$,* OptimisticDescent$(\mathcal{D}, v, \delta^*)$ *(Algorithm 33) outputs some vector $\widetilde{v}$ for which $\min_i\|\widetilde{v} - w_i\| \leq \varepsilon$ with probability at least $1 - \delta^*$.*

Proof. By Claims 8.7.26 and 8.7.28, it suffices to consider the case where there does not exist $0 \leq t < T$ for which (8.48) holds. Then $\langle w_i - a_{t}{}_2^2 \geq \varepsilon^2/2$ for every t, so by Claim 8.7.27,

$$\min_{i\in[k]}\|\widetilde{v} - w_i\|_2^2 + \varsigma^2 \leq \left(1 - \underline{\beta}\Delta^2/\sqrt{k}\right)^T \left(\min_{i\in[k]}\|v - w_i\|_2^2 + \varsigma^2\right) \leq \varepsilon^2,$$

where the last inequality follows by taking $T = \frac{2\sqrt{k}}{\underline{\beta}\Delta^2}\ln(1/\varepsilon)$. □

For $v \in \mathbb{R}^d$, denote by Z_v the event in Claim 8.7.29.

We next use Lemma 8.7.17 and part 4) of Claim 8.7.27 to lower bound the probability that v chosen in the inner loop of LearnWithNoise ends up being closest to any given component of $\mathcal{D}$.

Claim 8.7.30. *Take any $i^* \in [k]$ for which $\|w_{i^*}\|_2 \geq \underline{\sigma}$. Then there exists some $\alpha \in \mathcal{S}$ such that if v is a Haar-random vector in $\mathbb{R}^d$ of norm α, then conditioned on $\mathcal{E}_v$, the output $\widetilde{v}$ of* OptimisticDescent*$(\mathcal{D}, v)$ (Algorithm 33) is closest to w_{i^*} with probability at least $\exp(-\widetilde{O}(\sqrt{k}/\Delta^2))$ over the choice of v.*

Proof. Take any $i^* \in [k]$ for which $\|w_{i^*}\|_2 \geq \underline{\sigma}$. Recall the definition of $\mathcal{E}_v[i]$ from Lemma 8.7.17. By Lemma 8.7.17, there is some $\alpha \in \mathcal{S}$ such that if v is a Haar-random vector in $\mathbb{R}^d$ of norm α, then with probability at least $q_1 = \exp(-O(\sqrt{k}/\Delta^2))$ over v, $\mathcal{E}_v[i^*]$ holds. Then by 4) in Lemma 8.7.4, with probability $q_2 = \exp(-O(\sqrt{k}/\Delta^2)) \cdot \mathrm{poly}(k)^{-T}$, $\mathcal{E}_{v^*}[i^*]$ holds for, where $\widetilde{v}$ is the output of OptimisticDescent$(\mathcal{D}, v)$. This completes the proof, as $q_1 \cdot q_2 = \exp(-\widetilde{O}(\sqrt{k}/\Delta^2))$. □

We can now complete the proof of Lemma 8.7.25. Take $\delta^* = \frac{\delta}{2W \cdot |\mathcal{S}|}$ so that Z_v holds for all v sampled in LearnWithNoise with probability at least $1 - \delta/2$, by Claim 8.7.29. In this case, any $\widetilde{v}$ produced in the course of LearnWithNoise must be a ε-close to some component of $\mathcal{D}$.

Then by Claim 8.7.30, for any $i^* \in [k]$ for which $\|w_{i^*}\|_2 \geq \underline{\sigma} = \varepsilon/4$, the probability that some $\widetilde{v}$ produced in the course of LearnWithNoise is ε-close to i^* is at least $1 - (1-q)^W$, where $q \triangleq \exp(-\widetilde{O}(\sqrt{k}/\Delta^2))$. By taking $W = \ln(2k/\delta)/q$, we ensure that this happens with probability at least $\frac{\delta}{2k}$. We conclude by a union bound over $[k]$ that for every $i^* \in [k]$ for which $\|w_{i^*}\|_2 \geq \varepsilon/4$, there is some $\widetilde{v}$ produced in the course of LearnWithNoise which is ε-close to i^*.

Furthermore, by triangle inequality note that we never add vectors $\widetilde{v}$ to $\mathcal{L}$ which are ε-close to a component which is already ε-close to an existing $\widetilde{w} \in \mathcal{L}$.

Lastly, for $i^* \in [k]$ for which $\|w_{i^*}\|_2 \leq \varepsilon/4$, note that any vector v_{tiny} of norm $\varepsilon/4$ is $\varepsilon/2$-close to w_{i^*}. This completes the proof of Lemma 8.7.25. □

Lemma 8.7.31 (Running time of LearnWithNoise). *Let*

$$N_1 = \widetilde{O}(\varepsilon^{-2} p_{\min}^{-2} d k^2 \ln(1/\delta))$$
$$N = p_{\min}^{-4} k \ln(1/\delta) \cdot \mathrm{poly}\left(\overline{\sigma}, \sqrt{k}/\Delta^2, \ln(1/p_{\min}), \ln(1/\underline{\sigma})\right)^{O\left(\sqrt{k}\ln(1/p_{\min})/\Delta^2\right)} .$$

Then LearnWithNoise *(Algorithm 34) requires sample complexity $\widetilde{O}(e^{\widetilde{O}(\sqrt{k}/\Delta^2)}(N_1 + N))$ and runs in time $\widetilde{O}(e^{\widetilde{O}(\sqrt{k}/\Delta^2)}(dN_1 + N))$*

Proof. The complexity of the calls to CheckOutcome is dominated by the calls to OptimisticDescent, whose time and sample complexity are essentially identical to those of FourierMomentDescent called with failure probability parameter $\frac{\delta^*}{2W \cdot |\mathcal{S}|}$, except the complexity of the calls to CompareMinVariances now has exponential dependence on $\widetilde{O}(\sqrt{k}\ln(1/p_{\min})/\Delta^2)$ rather than on $\widetilde{O}(\sqrt{k}\ln(1/p_{\min}))$ because we only make $1 - \sqrt{k}/\Delta^2$ multiplicative progress at each step. Note the complexity of FourierMomentDescent depends only logarithmically on the inverse accuracy, so $\ln\left(\frac{\delta^*}{2W \cdot |\mathcal{S}|}\right)$ is absorbed into the $\widetilde{O}(\cdot)$. We conclude by noting that OptimisticDescent is called $|\mathcal{S}| \cdot W$ times, where $|\mathcal{S}| \leq \mathrm{poly}(k) \cdot \ln(1/\varepsilon)$ and $W = \exp(O(\sqrt{k}/\Delta^2)) \cdot \ln(2k/\delta)$. □

We can now complete the proof of Theorem 8.7.1.

Proof of Theorem 34. By Lemma 8.7.25, LearnWithNoise outputs a list of vectors $\{\widetilde{w}_1, ..., \widetilde{w}_k\}$ for which there exists a permutation $\pi : [k] \to [k]$ such that $\|w_i - \widetilde{w}_i\|_2 \leq \varepsilon$ for all $i \in [k]$. The runtime and sample complexity bounds follow from Lemma 8.7.31. □

8.7.5 Tolerating More Regression Noise

In this subsection we briefly remark that in the case where the mixing weights of $\mathcal{D}$ are known, we can combine Theorem 8.7.1 with the local convergence result of [KC19] to drive error down to ε even in settings where regression noise greatly exceeds ε.

Theorem 8.7.32 ([KC19], Theorem 3.2). *Let $\varepsilon > 0$. If $\mathcal{D}$ is a mixture of linear regressions with known mixing weights, separation Δ, and noise rate $\varsigma \leq O\left(\frac{\Delta}{k^2 \mathrm{poly}\log(k)}\right)$, and $\widetilde{w}_1, ..., \widetilde{w}_k \in \mathbb{R}^d$ is a list of vectors for which there exists a permutation $\pi : [k] \to [k]$ for* *which* $\|\widetilde{w}_i - w_{\pi(i)}\|_2 \le O\left(\frac{\Delta}{k^2}\right)$*, then with high probability finite-sample EM on a batch of* $\widetilde{O}(d) \cdot \mathrm{poly}(k, \ln(1/\varepsilon))$ *samples converges at an exponential rate to* $\widetilde{w}_1^*, ..., \widetilde{w}_k^* \in \mathbb{R}^d$ *for which there exists a permutation* $\pi' : [k] \to [k]$ *such that*

$$\max_{i\in[k]} \|\widetilde{w}_i^* - w_{\pi'(i)}\|_2 \le O(\varepsilon).$$

In particular, this implies the following:

Theorem 8.7.33. *Given* $\delta, \varepsilon > 0$ *and a mixture of spherical linear regressions* $\mathcal{D}$ *with regressors* $\{w_1, ..., w_k\}$*, separation* Δ*, noise rate* $\varsigma \le O\left(\frac{\Delta}{k^2 \mathrm{poly}\log(k)}\right)$*, and* known mixing weights*, there is an algorithm which with high probability outputs a list of vectors* $\mathcal{L} \triangleq \{\widetilde{w}_1, ..., \widetilde{w}_k\}$ *for which there is a permutation* $\pi : [k] \to [k]$ *for which* $\|\widetilde{w}_i - w_{\pi(i)}\|_2 \le \varepsilon$ *for all* $i \in [k]$*. Furthermore,* LearnWithNoise *requires sample complexity*

$$N = \widetilde{O}\left(d \ln(1/\delta) p_{\min}^{-4} \cdot \mathrm{poly}\left(k, 1/\Delta, \ln(1/p_{\min})\right)^{O(\sqrt{k}\ln(1/p_{\min})/\Delta^2)}\right)$$

and time complexity $Nd \cdot \mathrm{poly}\log(k, d, 1/\Delta, 1/p_{\min})$*.*

Proof. Simply run LearnWithNoise to learn to error $\varepsilon' = O(\Delta/k^2)$, which is possible because $\varsigma = O\left(\Delta/(k^2 \mathrm{poly}\log(k))\right) \ll \varepsilon'$. Then run finite-sample EM initialized to the output of LearnWithNoise to learn to error ε. □

8.8 Learning Mixtures of Hyperplanes

In this section we show that our techniques extend to give a sub-exponential time algorithm for learning mixtures of hyperplanes. Formally, we show the following:

Theorem 8.8.1. *Given* $\delta, \varepsilon > 0$ *and a mixture of hyperplanes* $\mathcal{D}$ *with directions* $\{v_1, ..., v_k\}$*, separation* Δ*, with probability at least* $1 - \delta$*,* LearnHyperplanes*(*$\mathcal{D}, \delta, \varepsilon$*) (Algorithm 38) returns a list of unit vectors* $\mathcal{L} \triangleq \{\widetilde{v}_1, ..., \widetilde{v}_k\}$ *for which there is a permutation* $\pi : [k] \to [k]$ *and signs* $\varepsilon_1, ..., \varepsilon_k \in \{\pm 1\}$ *for which* $\|\widetilde{v}_i - \varepsilon_i v_{\pi(i)}\|_2 \le \varepsilon$ *for all* $i \in [k]$*. Furthermore,* LearnHyperplanes *requires sample complexity*

$$N = \widetilde{O}\left(d\ln(1/\varepsilon)\ln(1/\delta)p_{\min}^{-4}\Delta^{-2}\cdot\mathrm{poly}\left(k,\ln(1/p_{\min}),\ln(1/\Delta)\right)^{O(k^{3/5}\ln(1/p_{\min}))}\right)$$

and time complexity $Nd\cdot\mathrm{poly}\log(k,d,1/\Delta,1/p_{\min},1/\varepsilon)$*.*

In Section 8.8.1 we show the key fact that a random step will contract $\min_i\|\Pi_i a_t\|_2$ by a factor of $1-\Theta(k^{-3/5})$ with probability at least $\exp(-k^{3/5})$, provided we use a suitable initialization. In Section 8.8.2 we give the full specification for HyperplaneMomentDescent which can learn a single component in a mixture of hyperplanes. In Section 8.8.3 we prove correctness for HyperplaneMomentDescent. In Section 8.8.4 we show that by properly regarding a mixture of hyperplanes as a mixture of well-conditioned, non-spherical MLRs, we can invoke the boosting result of [LL18] to amplify a warm start obtained by HyperplaneMomentDescent to an estimate with arbitrarily small error. Finally, in Section 8.8.5 we combine all of these primitives to obtain LearnHyperplanes and prove Theorem 8.8.1.

8.8.1 Moment Descent for Hyperplanes

In this section we give the key technical ingredients for showing that a suitable modification of FourierMomentDescent (Algorithm 31) can also be used to learn mixtures of hyperplanes.

Similar to the case of mixtures of linear regressions, here the first step is to estimate the span of the directions $\{v_i\}$. Define the matrix

$$\mathbf{M} \triangleq \mathrm{Id} - \mathbb{E}_{x\sim\mathcal{D}}[xx^\top] = \sum_{i=1}^k p_i v_i v_i^\top$$

and let $\widehat{\mathbf{M}}^{(N)} \triangleq \mathrm{Id} - \frac{1}{N}\sum_{i=1}^N x_i x_i^\top$ for $x_1, ..., x_N$ i.i.d. samples from $\mathcal{D}$. When the context is clear, we will omit the superscript (N).

We will need the following basic concentration inequality, which follows immediately from e.g. Theorem 4.7.1 of [Ver18].

Fact 8.8.2 (Concentration of sample covariance)**.** *For any $\delta_{\mathrm{samp}}, \delta > 0$, we have that*

$$Pr\left[\|\mathbf{M} - \widehat{\mathbf{M}}^{(N)}\|_2 > \Omega(1/p_{\min}) \cdot \left(\sqrt{\frac{d+t}{N}} + \frac{d+t}{N}\right)\right] \le 2e^{-t}.$$

Corollary 8.8.3. *For any $\delta_{\mathrm{samp}}, \delta > 0$, if $N = \widetilde{\Omega}\left(d \cdot \ln(1/\delta) \cdot p_{\min}^{-4} \cdot \delta_{\mathrm{samp}}^{-2}\right)$, then $\|\mathbf{M} - \widehat{\mathbf{M}}^{(N)}\|_2 \le p_{\min} \cdot \delta_{\mathrm{samp}}/2$ with probability at least $1 - \delta$.*

Henceforth, let $\mathbf{W} \in \mathbb{R}^{k\times d}$ be the matrix whose rows are the top k singular vectors of $\mathbf{M}$, let $\mathbf{U} \in \mathbb{R}^{k\times d}$ be the matrix whose rows are the first k singular vectors of $\widehat{\mathbf{M}}^{(N)}$ for N given in Corollary 8.8.3.

We will need the following basic bound which follows straightforwardly from Lemma 1.3.8.

Corollary 8.8.4. *Let $a_t \in \mathbb{S}^{d-1}$ lie in the row span of $\mathbf{U}$, let $v_j = x_j/\|x_j\|_2$, and let Π_j be the projector to the orthogonal complement of v_j. Suppose $\delta_{\mathrm{samp}} < 1/k$ and $\|x_j\| \le 1$. Then $\|\mathbf{U}\Pi_j a_t\|_2 \ge \|\Pi_j a_t\|_2 - \delta_{\mathrm{samp}}^2 - 2\delta_{\mathrm{samp}}$.*

In particular, for any constant $c > 0$, if $\|\Pi_j a_t\|_2 \ge \tau$ for some $\tau = \mathrm{poly}(k)$, then for sufficiently small $\delta_{\mathrm{samp}} = 1/\mathrm{poly}(k)$ we can ensure that $\|\mathbf{U}\Pi_j a_t\|_2 \ge (1 - k^{-c})\|\Pi_j a_t\|$.

Proof. First note that

$$\langle \mathbf{U}a_t, \mathbf{U}v_j\rangle = \langle a_t, v_j\rangle \pm \delta_{\mathrm{samp}} \tag{8.49}$$

by the first part of Lemma 1.3.8 and the fact that $|\langle a_t, v_j\rangle| \le 1$.

We have that

$$\begin{aligned}
\|\mathbf{U}\Pi_j a_t\|_2^2 &= \|\mathbf{U}a_t\|_2^2 + \langle a_t, v_j\rangle^2 \|\mathbf{U}v_j\|_2^2 - 2\langle a_t, v_j\rangle\langle \mathbf{U}a_t, \mathbf{U}v_j\rangle \\
&= 1 + \langle a_t, v_j\rangle^2 \|\mathbf{U}v_j\|_2^2 - 2\langle a_t, v_j\rangle\langle \mathbf{U}a_t, \mathbf{U}v_j\rangle \\
&\ge 1 + \langle a_t, v_j\rangle^2 \cdot (1 - \delta_{\mathrm{samp}}^2) - 2\langle a_t, v_j\rangle \cdot (\langle a_t, v_j\rangle - \delta_{\mathrm{samp}}) \\
&\ge \|\Pi_j a_t\|_2^2 - \delta_{\mathrm{samp}}^2 - 2\delta_{\mathrm{samp}}
\end{aligned}$$

where the second step follows from the fact that $\|\mathbf{U}a_t\|_2^2 = \|a_t\|_2^2 = 1$ as $a_t \in \mathbb{S}^{d-1}$ lies in the row span of $\mathbf{U}$, the third step follows from applying Corollary 8.8.4 and the lower bound of (8.49), and the last step follows from the fact that $\|\Pi_j a_t\|_2^2 = 1 - \langle a_t, v_j\rangle^2$ and $|\langle a_t, v_j\rangle| \le 1$. □

With these preliminary tools in hand, we are ready to prove the main result of this section, the mixture of hyperplanes analogue of Lemma 8.5.12.

Lemma 8.8.5. *Let $0 < c < 1/2$. There are absolute constants $\alpha > 0, \beta > 1, \nu > 0$ such that for any $\delta > \exp(-\nu \cdot k^{1-2c})$, the following holds for k sufficiently large. Let $\sigma_t \triangleq \min_{i\in[k]} \|\Pi_i a_t\|_2$, and suppose $\delta_{\mathrm{samp}} \le \sigma_t^2/9$. Denote the minimizing index i by i^*. For $M \triangleq e^{-\nu \cdot k^{1-2c}} \ln(2/\delta)$ and $g_1, ..., g_M \sim \mathcal{N}(0, Id_k)$, let $z_j = \frac{g_j \mathbf{U}}{g_j \|\mathbf{U}\|_2} \in \mathbb{S}^{d-1}$ for $j \in [M]$. Let σ^* be a number for which $0.9\sigma_t \le \sigma^* \le 1.1\sigma_t$. Let $\eta = k^{-c}\sigma^*$.*

If $|\langle a_t, v_{i^*}\rangle| \ge k^{-c}$, then we have that with probability at least $1-\delta$,

1. *There exists at least one $j \in [M]$ for which $\|\Pi_{i^*}(a_t - \eta \cdot z_j)\|_2^2 \le \left(1 - \frac{\alpha}{k^{3c}}\right)\sigma_t^2$. Denote any one of these indices by j^*.*

2. *For all $j \in [M]$ and $i \in [k]$,*

$$\frac{\|\Pi_i(a_t - \eta z_j)\|_2^2}{\sigma_t^2} \ge \left(\frac{\|\Pi_{i^*}(a_t - \eta z_{j^*})\|_2^2}{\sigma_t^2}\right)^\beta.$$

Proof. Without loss of generality, suppose $\langle a_t, v_{i^*}\rangle \ge 0$. For $j \in [M]$, let $\omega_j \triangleq -\eta^2 + 2\eta\langle z_j, \Pi_i a_t\rangle$ and $a_{t+1}^{(j)} = a_j - \eta \cdot z_j$. We know that

$$\begin{aligned}
\frac{\|\Pi_i a_{t+1}^{(j)}\|_2^2}{\|\Pi_i a_t\|_2^2} &= \frac{\|\Pi_i(a_t - \eta z_j)\|_2^2}{\|a_t - \eta z_j\|_2^2 \cdot \|\Pi_i a_t\|_2^2} \\
&= \frac{\|\Pi_i a_t\|^2 + \eta^2\|\Pi_i z_j\|^2 - 2\eta\langle z_j, \Pi_i a_t\rangle}{(1 + \eta^2 - 2\eta\langle a_t, z_j\rangle) \cdot \|\Pi_i a_t\|_2^2} \\
&= \frac{1 + \|\Pi_i a_t\|_2^{-2} \cdot (\eta^2\|\Pi_i z_j\|^2 - 2\eta\langle z_j, \Pi_i a_t\rangle)}{1 + \eta^2 - 2\eta\langle a_t, z_j\rangle} \\
&= \frac{1 + \|\Pi_i a_t\|_2^{-2} \cdot (\eta^2(1 - \langle v_i, z_j\rangle^2) - 2\eta\langle z_j, \Pi_i a_t\rangle)}{1 + \eta^2 - 2\eta\langle z_j, \Pi_i a_t\rangle - 2\eta\langle a_t, v_i\rangle\langle z_j, v_i\rangle} \\
&= \frac{1 - \|\Pi_i a_t\|_2^{-2}\left(\omega_j + \eta^2\langle v_i, z_j\rangle^2\right)}{1 - \omega_j - 2\eta\langle a_t, v_i\rangle\langle z_j, v_i\rangle}.
\end{aligned}$$

Define the events

$$A_j \triangleq \left\{\langle z_j, \Pi_{i^*} a_t\rangle \ge k^{-c}\|\Pi_{i^*} a_t\|_2 \quad \text{and} \quad \langle z_j, v_{i^*}\rangle \le -k^{-c}\right\}. \tag{8.50}$$

and

$$B_j[i] \triangleq \left\{|\langle z_j, \Pi_i a_t\rangle| \le \xi k^{-c}\|\Pi_i a_t\|_2 \quad \text{and} \quad |\langle z_j, v_i\rangle| \le \xi k^{-c}\right\} \tag{8.51}$$

for some $\xi > 1$ to be specified later. We show in Claim 8.8.6 below that for any given j, there is some absolute constant $\nu > 0$ such that $\Pr[A_j] \ge \exp(-\nu \cdot k^{1-2c})$, and some absolute constant $\xi > 1$ such that $\Pr[B_j[i]] \ge 1 - \exp(-3\nu \cdot k^{1-2c})$.

We will now argue that the event A_j corresponds to making good progress, while the event $B_j[i]$ corresponds to not making too much progress.

Suppose A_j held for some $j = j^*$ and $B_j[i]$ held for all $i \in [k], j \in [M]$. If we took $\eta \triangleq k^{-c}\sigma^*$, we would conclude from the definition of A_{j^*} and the assumption $0.9\sigma_t \le \sigma^* \le 1.1\sigma_t$ that

$$\omega_{j^*} \ge \|\Pi_{i^*} a_t\|_2^2 k^{-2c} \cdot (-0.9^2 + 2 \cdot 0.9) = 0.99\|\Pi_{i^*} a_t\|_2^2 k^{-2c}.$$

Likewise from the definition of $B_j[i]$ we would conclude that

$$\omega_j \le \|\Pi_i a_t\|_2^2 k^{-2c} \cdot (-1.21 + 2.2\xi).$$

for all $j \in [M]$ and $i \in [k]$. So we get that

$$\frac{\|\Pi_i a_{t+1}^{(j)}\|_2^2}{\|\Pi_i a_t\|_2^2} \ge \frac{1 - (2.2\xi - 1.21)k^{-2c} - (1.21\xi^2 k^{-2c}) \cdot (k^{-2c}\xi^2)}{1 - (2.2\xi - 1.21)k^{-2c}\|\Pi_i a_t\|_2^2 + 2 \cdot (0.9\xi k^{-c}) \cdot (k^{-c})\,\|\Pi_i a_t\|_2 \langle a_t, v_i\rangle} \tag{8.52}$$

for every $j \in [M]$ and $i \in [k]$, and for $i = i^*$ and $j = j^*$ we additionally have that

$$\frac{\|\Pi_{i^*} a_{t+1}^{(j^*)}\|_2^2}{\|\Pi_{i^*} a_t\|_2^2} \le \frac{1 - 0.99k^{-2c} - (0.81k^{-2c}) \cdot (k^{-2c})}{1 - 0.99k^{-2c}\|\Pi_{i^*} a_t\|_2^2 + 2 \cdot (1.1k^{-c}) \cdot (k^{-c})\,\|\Pi_{i^*} a_t\|_2 \langle a_t, v_{i^*}\rangle}. \tag{8.53}$$

In particular, we get that

$$\begin{aligned} 1 - \frac{\|\Pi_i a_{t+1}^{(j)}\|_2^2}{\|\Pi_i a_t\|_2^2} &\le \frac{1}{0.99}\left(1.21\xi^2 k^{-4c} + (2.2\xi - 1.21)k^{-2c}\langle a_t, v_i\rangle^2 + 1.8\xi k^{-2c}\|\Pi_i a_t\|_2 \langle a_t, v_i\rangle\right) \\ &\triangleq \overline{g}(\langle a_t, v_i\rangle) \end{aligned}$$

for every $j \in [M]$ and $i \in [k]$, and for $i = i^*, j = j^*$, we additionally have that

$$
\begin{aligned}
1 - \frac{\|\Pi_{i^*} a_{t+1}^{(j^*)}\|_2^2}{\|\Pi_{i^*} a_t\|_2^2} &\geq \frac{1}{1.01}\left(0.81k^{-4c} + 0.99k^{-2c}\langle a_t, v_{i^*}\rangle^2 + 2.2k^{-2c}\|\Pi_{i^*} a_t\|_2 \langle a_t, v_{i^*}\rangle\right) \\
&\triangleq \underline{g}(\langle a_t, v_{i^*}\rangle)
\end{aligned}
$$

where we have used the fact that the denominators of (8.52) and (8.53) are in $[0.99, 1.01]$ for sufficiently large k. Also, we emphasize that these quantities can be expressed as functions $\overline{g}$ and $\underline{g}$ solely in $\langle a_t, v_i\rangle$ because for any $i \in [k]$, $\|\Pi_i a_t\|_2^2 = 1 - \langle v_i, a_t\rangle^2$.

To control these quantities, note that the function $\underline{g}(x)$ is increasing over the interval $[0, \tau^*]$ and decreasing over the interval $[\tau^*, 1]$ for some constant $\tau^* \in [0.91, 0.92]$. When $\langle v_{i^*}, a_t\rangle = 1$, we get that $\underline{g} = \Omega(k^{-2c})$, because the $0.99k^{-2c}\langle a_t, v_{i^*}\rangle^2$ term in the definition of $\underline{g}$ dominates. And when $\langle v_{i^*}, a_t\rangle = k^{-c}$, we get that $\underline{g} = \Omega(k^{-3c})$, because the $2.2k^{-2c}\|\Pi_{i^*} a_t\|_2 \langle a_t, v_{i^*}\rangle$ term in the definition of $\underline{g}$ dominates. So the first part of the lemma follows.

On the other hand, there is some absolute constant $\beta' > 1$ such that $\underline{g}(x) \leq \overline{g}(x) \leq \beta' \underline{g}(x)$ for all $x \in [0, 1]$. There is some constant $\beta'' > 1$ for which $\underline{g}(x)/\underline{g}(y) < \beta''$ for all $0 \leq x \leq y \leq 1$. The reason is that $\underline{g}(x)$ is increasing over the interval $[0, \tau^*]$ and decreasing over the interval $[\tau^*, 1]$, and $\underline{g}(x) = 1 - \Omega(k^{-2c})$ for $x \in [\tau^*, 1]$.

It follows that for any $j \in [M], i \in [k]$

$$
1 - \frac{\|\Pi_i a_{t+1}^{(j)}\|_2^2}{\|\Pi_i a_t\|_2^2} \leq \overline{g}(\langle a_t, v_i\rangle) \leq \beta' \underline{g}(\langle a_t, v_i\rangle) \leq \beta'\beta'' \cdot \underline{g}(\langle a_t, v_{i^*}\rangle) \leq \beta'\beta'' \cdot \left(1 - \frac{\|\Pi_{i^*} a_{t+1}^{(j^*)}\|_2^2}{\|\Pi_{i^*} a_t\|_2^2}\right),
$$

so by taking β in the statement of the lemma to be $\beta' \cdot \beta''$ and invoking the elementary inequality $1 - a \cdot x \leq (1 - x)^a$ for $a > 1$, we get the second part of the lemma.

We conclude that if event (8.50) held for some $j \in [M]$ and event (8.51) held for all $j \in [M]$ and $i \in [k]$, then parts 1 and 2 of Lemma 8.8.5 would hold.

Furthermore,

$$
\Pr\left[\bigwedge_{j \in [M]} \overline{A}_j\right] \leq \left(1 - \exp(-\nu \cdot k^{1-2c})\right)^M
$$

and

$$\Pr\left[\bigvee_{i\in[k],j\in[M]} \overline{B}_j[i]\right] \le kM\cdot\exp(-3\nu\cdot k^{1-2c}),$$

so by taking $M = \exp(\nu\cdot k^{1-2c})\ln(2/\delta)$, we get that with probability at least $1-\delta/2$, the event A_j occurs for some $j\in[M]$. And with probability at least $1-k\exp(-2\nu\cdot k^{1-2c})\ln(2/\delta)\ge 1-\delta/2$, the event $B_j[i]$ occurs for all $i\in[k]$ and $j\in[M]$, where we have used the fact that $\delta > \exp(-\nu\cdot k^{1-2c})$. □

It remains to lower bound the probabilities of events A_j and $B_j[i]$.

Claim 8.8.6. *There are absolute constants $\nu > 0, \xi > 1$ such that for any $j\in[M], i\in[k]$,*

$$Pr[A_j] \ge \exp(-\nu\cdot k^{1-2c}) \quad \textit{and} \quad Pr[B_j[i]] \ge 1-\exp(-3\nu\cdot k^{1-2c}).$$

Proof. While (8.50) is defined with respect to v_{i^*}, the argument below holds for general i. Henceforth, fix an arbitrary $i\in[k]$ and $j\in[M]$. The key fact we will use is that the two quantities $\langle z,\Pi_i a_t\rangle$ and $\langle z, v_i\rangle$ are approximately independent (if U consisted of the k top singular vectors of $\mathbf{M}$ itself, these random variables would be exactly independent).

Note that

$$\begin{aligned}\rho \triangleq \left\langle \mathbf{U}\Pi_i a_t, \frac{\mathbf{U}v_j}{\|\mathbf{U}v_j\|_2}\right\rangle &= \frac{1}{\|\mathbf{U}v_j\|_2}\cdot\left[\langle \mathbf{U}a_t, \mathbf{U}v_j\rangle - \langle a_t, v_j\rangle\|\mathbf{U}v_j\|_2^2\right]\\ &\le \frac{1}{(1-\delta_{\text{samp}}^2)^{1/2}}\cdot\left[(\langle a_t, v_j\rangle + \delta_{\text{samp}}) - \langle a_t, v_j\rangle\cdot(1-\delta_{\text{samp}}^2)\right]\\ &\le \frac{\delta_{\text{samp}}+\delta_{\text{samp}}^2}{(1-\delta_{\text{samp}}^2)^{1/2}} \le 2\delta_{\text{samp}}.\end{aligned}$$

where the second step follows from the second part of Lemma 1.3.8 and (8.49). Likewise we have that

$$\rho \ge (\langle a_t, v_j\rangle - \delta_{\text{samp}}) - \langle a_t, v_j\rangle \ge -\delta_{\text{samp}}.$$

So we may write

$$\mathbf{U}\Pi_i a_t = \rho\cdot\frac{\mathbf{U}v_j}{\|\mathbf{U}v_j\|_2} + v^{\perp} \tag{8.54}$$

for $v^\perp$ lying in the row span of $\mathbf{U}$ and orthogonal to $\mathbf{U}v_j$, and satisfying

$$\|v^\perp\|^2 \geq \|\mathbf{U}\Pi_i a_t\|^2 - \rho^2 \geq \left(\|\Pi_i a_t\|_2^2 - \delta_{\text{samp}}^2 - 2\delta_{\text{samp}}\right) - 4\delta_{\text{samp}}^2 \geq \frac{1}{2}\|\Pi_i a_t\|_2^2$$

by Lemma 8.8.5 and the assumption that $\|\Pi_i a_t\| \geq 3\delta_{\text{samp}}^{1/2}$. As $\langle g, \mathbf{U}v_j\rangle$ and $\langle g, v^\perp\rangle$ are *independent* Gaussians with variances at least $1 - \delta_{\text{samp}}^2 \geq 1/2$ and $\frac{1}{2}\|\Pi_i a_t\|_2^2$ respectively, by the same argument as in Corollary 1.3.20 we can show that $\left\langle \frac{g}{\|g\|_2}, \mathbf{U}v_j \right\rangle \leq -2 \cdot k^{-c}$ and $\left\langle \frac{g}{\|g\|_2}, v^\perp \right\rangle \geq 2 \cdot k^{-c} \cdot \|\Pi_i a_t\|_2$ with probability at least $\exp(-\nu \cdot k^{1-2c})$ for some absolute constant $\nu > 0$. By another application of Corollary 1.3.19, there is some absolute constant $\xi' > 0$ for which $\left|\left\langle \frac{g}{\|g\|_2}, \mathbf{U}v_j \right\rangle\right| \leq \xi' \cdot k^{-c}$ and $\left|\left\langle \frac{g}{\|g\|_2}, v^\perp \right\rangle\right| \leq \xi \cdot k^{-c} \cdot \|\Pi_i a_t\|_2$ with probability at least $1 - \exp(-3\nu \cdot k^{1-2c})$.

If this is the case, then by (8.54),

$$\left\langle \frac{g}{\|g\|_2}, \mathbf{U}\Pi_i a_t \right\rangle \geq \frac{-4\delta_{\text{samp}} \cdot k^{-c}}{(1 - \delta_{\text{samp}}^2)^{1/2}} + 2 \cdot k^{-c} \cdot \|\Pi_i a_t\|_2 \geq \frac{3}{2} k^{-c} \cdot \|\Pi_i a_t\|_2 \geq k^{-c} \cdot \|\Pi_i a_t\|_2,$$

where we have used that $\|\Pi_i a_t\| \geq 3\delta_{\text{samp}}^{1/2} \geq 4 \cdot \frac{4\delta_{\text{samp}}}{(1-\delta_{\text{samp}}^2)^{1/2}}$ for any δ_{samp} smaller than some absolute constant. We also get that

$$\left|\left\langle \frac{g}{\|g\|_2}, \mathbf{U}\Pi_i a_t \right\rangle\right| \leq 2\delta_{\text{samp}} \cdot \xi' \cdot k^{-c} + \xi' \cdot k^{-c} \cdot \|\Pi_i a_t\|_2 \leq 19\xi' \cdot k^{-c} \cdot \|\Pi_i a_t\|_2,$$

where we have used that $\|\Pi_i a_t\| \geq \|\Pi_i a_t\|^2 \geq 9\delta_{\text{samp}}$.

Noting that $\|g\mathbf{U}\|_2 = \|g\|$ by orthonormality of the columns of $\mathbf{U}$ so that

$$\left\langle \frac{g}{\|g\|_2}, \mathbf{U}v_j \right\rangle = \langle z, v_j \rangle \quad \text{and} \quad \left\langle \frac{g}{\|g\|_2}, \mathbf{U}\Pi_i a_t \right\rangle = \langle z, \Pi_i a_t \rangle,$$

we conclude that with probability at least $\exp(-\nu \cdot k^{1-2c})$, both events in (8.50) hold, and likewise with probability at least $1 - \exp(-3\nu \cdot k^{1-2c})$, both events in (8.51) hold for $\xi = \xi'/19$. □

8.8.2 Algorithm Specification– Single Component

We are now ready to describe our algorithm HyperplaneMomentDescent for learning a single components of $\mathcal{D}$. The key subroutines are:

- HyperplaneMomentDescent (Algorithm 33): the pseudocode for this is very similar to that of FourierMomentDescent, the key differences being 1) the matrix on which we run ApproxBlockSVD, 2) the definition of $\mathcal{F}_t$, 3) the fact that we maintain that a_t are unit vectors, 4) the parameters which are tuned towards detecting $1 - \Omega(k^{-3/5})$ multiplicative progress instead of $1 - \Omega(k^{-1/2})$, and most importantly, 5) the outer loop over $i \in [S]$ which tries many random initializations, runs a full T rounds of moment descent on each of them, and checks whether the final estimate in any of these runs is close to a component of $\mathcal{D}$.

- CheckOutcomeHyperplanes (Algorithm 35): CheckOutcome is used to check whether a given estimate is close to any component of $\mathcal{D}$.

8.8.3 Proof of Correctness

We first give a proof of correctness for CheckOutcomeHyperplanes.

Lemma 8.8.7. *Let $v \in \mathbb{S}^{d-1}$ and $\mathcal{D}$ be a mixture of hyperplanes, and let $\varepsilon > 0$. If there is some component v_{i^*} for which $\|v - v_{i^*}\|_2 \in [-\varepsilon, \varepsilon]$, then* CheckOutcomeHyperplanes*$(\mathcal{D}, v, \varepsilon, \delta)$ (Algorithm 37) returns* True *with probability at least $1 - \delta$. Otherwise, if $\|v - v_i\|_2 > 2\varepsilon/p_{\min}$ for all $i \in [k]$, then* CheckOutcomeHyperplanes *returns* False *with probability at least $1 - \delta$.*

Proof. First suppose there is some $i^* \in [k]$ for which $\|v - v_{i^*}\|_2 \leq \varepsilon$. Then $\mathcal{F}$ is a mixture of Gaussians with one of its components having variance at most ε^2. So for $x \sim \mathcal{F}$, we get that

$$\Pr[|x| \leq \varepsilon/2] \geq p_{\min} \cdot \int_{-\varepsilon/2}^{\varepsilon/2} e^{-\frac{x^2}{2\varepsilon^2}} \, \mathrm{d}x \geq p_{\min}/3.$$

On the other hand if we had that $\|v - v_i\|_2 > 2\varepsilon/p_{\min}$ for all $i \in [k]$, then $\mathcal{F}$ is a mixture of Gaussians whose components have variances exceeding $\frac{4\varepsilon^2}{p_{\min}^2}$. So for $x \sim \mathcal{F}$, we would get

```
Algorithm 36: HyperplaneMomentDescent($\mathcal{D}, \delta, \varepsilon$)
Input: Sample access to mixture of hyperplanes $\mathcal{D}$, failure probability $\delta$, error $\varepsilon$
Output: $v^* \in \mathbb{S}^{d-1}$ satisfying $\min_{i\in[k]} \|\Pi_i v^*\|_2 \le \varepsilon$, with probability at least $1-\delta$.
$\varepsilon' \leftarrow \varepsilon \cdot p_{\min}/2$.
$S \leftarrow \exp(-\Omega(k^{1-2c})) \cdot \ln(2/\delta)$.
$T \leftarrow \Omega(k^{3/5} \cdot \ln(\mu/\varepsilon'))$.
$\delta_{\mathrm{samp}} \leftarrow 1/\mathrm{poly}(k)$ sufficiently small.
$\delta' \leftarrow \frac{\delta}{50T}$.
$M \leftarrow e^{-\nu \cdot k^{1-2c}} \ln(2/\delta')$.
$\delta'' \leftarrow \frac{\delta}{50MT}$.
$\delta^* \leftarrow \frac{\delta}{2S}$.
$\underline{\sigma} \leftarrow 2 \cdot (\varepsilon'/2)^{\beta}$, where $\beta > 1$ is the constant from Lemma 8.8.5.
$\overline{\sigma} \leftarrow 4$.
$N_1 \leftarrow \Omega\left(d \cdot \ln(1/\delta') \cdot p_{\min}^{-2} \cdot \delta_{\mathrm{samp}}^{-2}\right)$
Draw $N_1$ i.i.d. samples $\{x_i\}_{i\in[N_1]}$ from $\mathcal{D}$ and form the matrix
  $\widehat{\mathbf{M}}^{(N_1)} \triangleq \mathrm{Id} - \frac{1}{N_1}\sum_{i=1}^{N_1} x_i x_i^\top$.
$\mathbf{U} \leftarrow$ ApproxBlockSVD($\widehat{\mathbf{M}}^{(N_1)}, \delta_{\mathrm{samp}}/2, 1/50$)
for $0 \le i < S$ do
    Sample $g \sim \mathbb{N}(0, \mathrm{Id}_k)$ and let $a_0 = \frac{g\mathbf{U}}{\|g\mathbf{U}\|_2}$.
    for $0 \le t < T$ do
        Let $\mathcal{F}_t$ be the univariate mixture of Gaussians which can be sampled from by
          drawing $x \sim \mathcal{D}$ and computing $\langle a_t, x\rangle$.
        $p \leftarrow 20 \ln\left(\frac{3}{2p_{\min}}\right)$.
        $\kappa \leftarrow \Theta(k^{-3/5})$ as in the proof of Lemma 8.8.8.
        $\underline{\sigma}_t^{\mathrm{sharp}} \triangleq$ EstimateMinVariance($\mathcal{F}_t, \overline{\sigma}, \underline{\sigma}, p, \delta'$).        // Algorithm 29
        for $j \in [M]$ do
            Sample $g_t^{(j)} \sim \mathbb{N}(0, \mathrm{Id}_k)$ and define $v_t^{(j)} = \frac{g_t^{(j)}\mathbf{U}}{\|g_t^{(j)}\mathbf{U}\|_2} \in \mathbb{S}^{d-1}$.
            ${a'}_t^{(j)} \leftarrow \frac{a_t - \eta_t v_t^{(j)}}{\|a_t - \eta_t v_t^{(j)}\|_2}$ for $\eta_t \triangleq k^{-1/5} \cdot \underline{\sigma}_t^{\mathrm{sharp}}$.
            Let ${\mathcal{F}'}_t^{(j)}$ be the univariate mixture of Gaussians which can be sampled
              from by drawing $x \sim \mathcal{D}$ and computing $\langle a_t, x\rangle$.
            if CompareMinVariances($\mathcal{F}_t, {\mathcal{F}'}_t^{(j)}, \overline{\sigma}, \underline{\sigma}, \kappa, \delta''$) = True then
                Set $a_{t+1} = {a'}_t^{(j)}$
                Break
    if CheckOutcomeHyperplanes($\mathcal{D}, a_t, \varepsilon', \delta^*$) = True then
        $v^* \leftarrow a_t$.
        return $v^*$.
```

Algorithm 37: CheckOutcomeHyperplanes($\mathcal{D}, v, \varepsilon, \delta$)

```
Input: Sample access to mixture of hyperplanes $\mathcal{D}$, direction $v \in \mathbb{S}^{d-1}$, threshold
       $\varepsilon > 0$, failure probability $\delta$
Output: True if $\min_{i\in[k]}\|\Pi_i v\|_2 \le \varepsilon$, False if $\|\Pi_i v\|_2 \ge 2\varepsilon/p_{\min}$, with probability at
        least $1-\delta$
Let $\mathcal{F}$ be the univariate mixture of Gaussians which can be sampled from by
  drawing $x \sim \mathcal{D}$ and computing $\langle v, x\rangle$.
Draw $N_2 \triangleq O\left(\ln(1/\delta)p_{\min}^{-2}\right)$ samples from $\mathcal{F}$.
if $\ge \frac{4p_{\min}}{15}\cdot N_2$ samples lie in $[-\varepsilon,\varepsilon]$ then
    return True
return False
```

that

$$\Pr[|x| \le \varepsilon/2] \le \sum_{i=1}^k p_i \cdot \frac{0.8\cdot\varepsilon/2}{2\varepsilon/p_{\min}} = p_{\min}/5,$$

where in the first step we have used the fact that $\int_{-\tau}^{\tau} e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x \le \sqrt{2/\pi}\cdot(\tau/\sigma) \le 0.8\tau/\sigma$ for any $\tau, \sigma > 0$.

We need to take enough samples for our empirical estimate of $\Pr[|x| \le \varepsilon]$ to be $p_{\min}/15$-additively close to the true value with probability at least $1-\delta$, for which it suffices to take $O(\ln(1/\delta)p_{\min}^{-2})$ samples. □

We can now prove correctness of HyperplaneMomentDescent.

Lemma 8.8.8. *Let $\mathcal{D}$ be a mixture of hyperplanes with mixing weights $\{p_i\}$ and directions $\{v_i\}$. With probability at least $1-\delta$,* HyperplaneMomentDescent*($\mathcal{D}, \delta, \varepsilon$) (Algorithm 36) outputs direction $a_T \in \mathbb{S}^{d-1}$ for which $(\varepsilon/2)^C \le \min_{i\in[k]}\|w_i - a_T\|_2 \le \varepsilon$ for some absolute constant $C > 0$.*

Proof. Henceforth, take c in Lemma 8.8.5 to be $c = 1/5$. Let $\sigma_t \triangleq \min_{i\in[k]}\|w_i - a_t\|_2$. Naively we have that $\sigma_t \le 2$.

By a simple union bound, we first upper bound the probability that the steps of moment descent in the i-th iteration of the outer loop all succeed.

Claim 8.8.9. *Let $i \in [S]$. With probability at least $9/10$, the randomized components of the inner loop (over t) of the i-th iteration of the outer loop of* HyperplaneMomentDescent *all succeed.*

Proof. Each t-th iteration of the second loop in HYPERPLANEMOMENTDESCENT has the following randomized components: 1) empirically estimating $\mathbf{M}$, 2) running APPROXBLOCKSVD on this empirical estimate, 3) running ESTIMATEMINVARIANCE, 4) trying the Gaussian vectors g in the innermost loop over $j \in [M]$, and 5) running COMPAREMINVARIANCES in this innermost loop.

Because the failure probability δ' for 1), 3), 4) were chosen to be $\frac{1}{50T}$, the failure probability δ'' for 5) was chosen to be $\frac{1}{50MT}$, and the failure probability for 2) was chosen to be $1/50$, we can bound by $1/10$ the overall failure probability of these tasks in a single i-th iteration of the outer loop of HYPERPLANEMOMENTDESCENT. □

Call the event in Claim 8.8.9 $\mathcal{E}_i$. Next, we show that provided $\mathcal{E}_i$ occurs and the initial point a_0 for the i-th iteration of the outer loop is close to some v_j, then we can bound the extent to which every step of the subsequent inner loop (over t) contracts σ_t^2.

Claim 8.8.10. *Let $i \in [S]$ and condition on $\mathcal{E}_i$. If in the i-th iteration of* HYPERPLANEMOMENTDESCENT, *$|\langle a_0, v_j\rangle| \geq k^{-c}$, then for each $0 \leq t < T$:*

1. *(Completeness) There exists some $j \in [M]$ for which*

$$\textsc{CompareMinVariances}(\mathcal{F}_t, \mathcal{F}_t'^{(j)}, \overline{\sigma}, \underline{\sigma}, \kappa, 2\kappa, \delta'')$$

outputs True *for some $\kappa = \Theta(k^{-3c})$.*

2. *(Soundness) For any such $j \in [M]$ for which* COMPAREMINVARIANCES *outputs* True*,*

$$\left(1 - \frac{\beta}{k^{3c}}\right)\sigma_t^2 \leq \sigma_{t+1}^2 \leq \left(1 - \frac{\underline{\alpha}}{k^{3c}}\right)\sigma_t^2. \tag{8.55}$$

for some $\underline{\alpha} < \alpha$, where α, β are the constants in Lemma 8.8.5.

Proof. Suppose inductively that $|\langle a_t, v_j\rangle| \geq k^{-c}$ for some $j \in [k]$. By the first part of Lemma 8.8.5, there exists some $j \in [M]$ for which $\sigma_t^{(j)} \triangleq \min_{i\in[k]}\|w_i - a_t'^{(j)}\|_2$ satisfies $(\sigma_t^{(j)})^2 \leq \left(1 - \frac{\alpha}{k^{3c}}\right)(\sigma_t)^2$, and because $1 - \frac{\alpha}{k^{3c}} \leq \left(\frac{1}{1+2\kappa}\right)^2$ for some $\kappa = \Theta(k^{-3c})$,

$$\textsc{CompareMinVariances}(\mathcal{F}_t, \mathcal{F}_t'^{(j)}, \overline{\sigma}, \underline{\sigma}, \kappa, 2\kappa, \delta'')$$

would return True, completing the proof of completeness.

For soundness, note that for any such j, by Corollary 8.5.9 we know that

$$(\sigma_t^{(j)})^2 \le (1+\kappa)^{-2} \cdot \sigma_t^2 \le (1-\kappa/2) \cdot \sigma_t^2 \le \left(1 - \frac{\alpha}{k^{3c}}\right) \cdot \sigma_t^2,$$

which gives the upper bound in (8.55). The lower bound follows from the second part of Lemma 8.8.5. This completes the proof of soundness as well as the inductive step, as the upper bound of (8.55) implies that $\max_{j\in[k]} |\langle a_{t+1}, v_j\rangle| \ge \max_{j\in[k]} |\langle a_t, v_j\rangle| \ge k^{-c}$. □

Lastly, we lower bound the probability that in the i-th iteration of the outer loop, the randomly chosen initial point a_0 is sufficiently close to some v_j.

Claim 8.8.11. *Let $i \in [S]$. With probability at least $\exp(-\Omega(k^{1-2c}))$, the following holds. In the i-th iteration of the outer loop of* HyperplaneMomentDescent*, $|\langle a_0, v_j\rangle| \ge k^{-c}$ for some $j \in [k]$, where a_0 is the initial iterate in the inner loop over t.*

Proof. We know by Corollary 1.3.19 that for $g \sim \mathcal{N}(0, \mathrm{Id}_k)$ and $a_0 \triangleq \frac{g\mathbf{U}}{\|g\mathbf{U}\|_2}$, for any $j \in [k]$ we have that $\Pr_g[|\langle a_0, v_j\rangle| \ge k^{-c}] \ge \exp(-\Omega(k^{1-2c}))$. □

We are ready to complete the proof of Lemma 8.8.8. If we take $T = k^{3c}\ln(8\varepsilon^{-2}p_{\min}^{-2})/\underline{\alpha}$, then in an iteration $i \in [S]$ for which $\mathcal{E}_i$ holds and $|\langle a_0, v\rangle| \ge k^{-c}$, by Claim 8.8.10 we are guaranteed that

$$2/(\varepsilon^2 p_{\min}^2/8)^C \le \sigma_T^2 \le \varepsilon^2 \cdot p_{\min}^2/4 \tag{8.56}$$

for some absolute constant $C > 0$. We remark that the lower bound on σ_T^2 ensures that throughout the course of HyperplaneMomentDescent, the parameter $\underline{\sigma}$ passed to CompareMinVariances is a valid lower bound on σ_t for all $0 \le t \le T$.

Now for $i \in [S]$, let A_i be the event that the inner loop breaks out with a direction a for which $\|\Pi_j a\|_2 \le \varepsilon \cdot p_{\min}/2$. Also, let B_i be the event that CheckOutcomeHyperplanes runs successfully. Note that A_i and B_i are independent. By Claim 8.8.9, Claim 8.8.11, and (8.56), we know $\Pr[A_i] \ge \frac{9}{10}\exp(-\Omega(k^{1-2c})) \triangleq q$. By Lemma 8.8.7, we know $\Pr[B_i] \ge 1-\delta^*$.

The probability that A_i occurs for at least one $i \in [S]$ is at least $1-(1-q)^S \ge 1-e^{-qS}$, while the probability that B_i holds for all i is at least $1-S\delta^*$. By taking $S = q^{-1}\ln(2/\delta) =$

$\exp(-\Omega(k^{1-2c})) \cdot \ln(2/\delta)$ and $\delta^* = \frac{\delta}{2S}$, we conclude that the output of HyperplaneMomentDescent is some v^* for which $\min_{i\in[k]} \|\Pi_i v^*\|_2 \leq \varepsilon$. □

The analysis for the runtime and sample complexity of HyperplaneMomentDescent is essentially the same as that of FourierMomentDescent:

Lemma 8.8.12 (Running time of HyperplaneMomentDescent)**.** *Let*

$$\begin{aligned}
N_1 &= \frac{dk^2 \ln(1/\delta)}{\varepsilon^2 p_{\min}^2} \\
N &= p_{\min}^{-4} k \ln(1/\delta) \cdot \mathrm{poly}\left(k^{3/5}, \ln(1/p_{\min}), \ln(1/\varepsilon)\right)^{O\left(k^{3/5} \ln(1/p_{\min})\right)} \\
N_2 &= O\left(p_{\min}^{-2} k^{3/5} \ln(1/\delta)\right) \\
S &= \exp(\Omega(k^{3/5}) \ln(1/\delta)\ .
\end{aligned}$$

Then HyperplaneMomentDescent *(Algorithm 31) requires sample complexity*

$$\widetilde{O}\left(N_1 + S \cdot (k^{3/5} e^{k^{3/5}} N + N_2)\right)$$

and runs in time

$$\widetilde{O}\left(dN_1 + S \cdot (k^{3/5} e^{k^{3/5}} N + N_2)\right).$$

8.8.4 Boosting for Mixtures of Hyperplanes

As with FourierMomentDescent, HyperplaneMomentDescent cannot be used on its own to obtain an arbitrarily good estimate for a component of the mixture, as the runtime and sample complexity of the primitives used for estimating minimum variance increase rapidly as the minimum variance of the univariate projections decreases. So at some point we need to switch over to a boosting algorithm.

In this section, we describe how to regard mixtures of hyperplanes as mixtures of non-spherical but fairly well-conditioned linear regressions. With this in place, we can then run either the boosting algorithm of [LL18] or the one introduced in our work in this chapter (see Section 8.9), all of which can tolerate the condition numbers of such mixtures.

Let $w \in \mathbb{S}^{d-1}$ be some direction and let Π_w denote the projection to the orthogonal complement of w. Given $x \sim \mathcal{D}$, we may regard this as a sample from a mixture of linear regressions as follows. Consider the tuple $(\Pi_w x, \langle x, w\rangle)$. By identifying Π_w with $\mathbb{R}^{d-1}$, we may regard $\Pi_w x$ as a vector in $\mathbb{R}^{d-1}$ and $\langle x, w\rangle$ as the response.

Concretely, up to a change of basis we can assume without loss of generality that $w = (0, ..., 0, 1)$, in which case $\Pi_w x$ is simply identified with the first $d-1$ coordinates of x, and the response $\langle x, w\rangle$ is simply the last coordinate of x. Then the covariance matrix of the hyperplane orthogonal to v_j is merely the upper $(d-1) \times (d-1)$ submatrix of Π_j, and because any x sampled from that hyperplane satisfies $\langle v_j, x\rangle = 0$, we have that

$$x_d = \left\langle -\frac{(v_j)_{1:d-1}}{(v_j)_d}, x_{1:d-1} \right\rangle,$$

where we use the notation of Section 8.2.2. For simplicity, denote $(v_j)_{1:d-1}$ by v'_j, $(v_j)_d$ by a_j. We may further assume without loss of generality that $a_j = \langle v_j, w\rangle$ is nonnegative for every j, as the directions $\{v_j\}$ for a mixture of hyperplanes are only specified up to sign.

Altogether, this yields the following basic claim.

Lemma 8.8.13. *Given a mixture $\mathcal{D}$ of hyperplanes with mixing weights $\{p_j\}$ and directions $\{v_j\}$, let $\mathcal{D}'$ be the mixture of linear regressions with mixing weights $\{p_j\}$, components $\{\mathcal{N}(0, Id - v'_j v'^{\top}_j)\}$, and regressors $\{-v'_j/a_j\}$. Then $\mathcal{D}$ and $\mathcal{D}'$ are identical as distributions over $\mathbb{R}^d$.*

We will choose w randomly by sampling $g \sim \mathcal{N}(0, \mathrm{Id}_k)$ and defining $w = \frac{g\mathbf{U}}{\|g\mathbf{U}\|_2}$. We need a basic estimate on the condition number of the covariances $\mathrm{Id} - v'_j v'^{\top}_j$ for a typical such w, keeping in mind that v'_j is defined with respect to an orthonormal basis under which v'_j is the d-th standard basis vector.

Lemma 8.8.14. *For $g \sim \mathcal{N}(0, Id_k)$ and $w = \frac{g\mathbf{U}}{\|g\mathbf{U}\|_2}$, the eigenvalues of $Id - v'_j v'^{\top}_j$ lie in $[\Omega(1/k^3), 1]$ for all $j \in [k]$ with probability at least $4/5$.*

Proof. Let $\widetilde{v}_j \triangleq \frac{v'_j}{\|v'_j\|_2} \in \mathbb{S}^{d-2}$. Note that

$$\mathrm{Id} - v'_j v'^{\top}_j = \mathrm{Id} - \widetilde{v}_j \widetilde{v}^{\top}_j \cdot \|v'_j\|_2^2 = \mathrm{Id} - \widetilde{v}_j \widetilde{v}^{\top}_j \cdot (1 - \langle w, v_j\rangle^2),$$

so the eigenvalues of $\mathrm{Id} - v'_j v'^{\top}_j$ are 1 with multiplicity $d-2$ and $\langle w, v_j \rangle^2$ with multiplicity 1. Fact 8.8.15 below allows us to conclude that with probability at least 4/5, $\langle w, v_j \rangle^2 \geq \Omega(1/k^3)$ for all $j \in [k]$. □

Fact 8.8.15. *There is some constant $a_{\mathrm{anti}} > 0$ such that for the random vector w defined in Lemma 8.8.14,*

$$Pr\left[\langle w, v_j \rangle^2 \geq \frac{a_{\mathrm{anti}}}{k^3} \ \forall\ j \in [k]\right] \geq 4/5.$$

Proof. For any $j \in [k]$, we have that

$$\langle w, v_j \rangle = \frac{1}{\|g\mathbf{U}\|_2} \langle g, \mathbf{U} v_j \rangle = \frac{1}{\|g\|_2} \langle g, \mathbf{U} v_j \rangle .$$

By the second part of Lemma 1.3.8, $\|\mathbf{U} v_j\|_2 \geq (1 - \delta^2_{\mathrm{samp}})^{1/2} \geq 1/2$, and by Fact 1.3.18, $\|g\mathbf{U}\|_2 \leq 1.1\sqrt{k}$ with probability at least $1 - e^{-c_{\mathrm{shell}} d/100}$. $\langle g, \mathbf{U} v_j \rangle$ is distributed as a zero-mean Gaussian with variance at least $1 - \delta^2_{\mathrm{samp}}$, so for any $\tau > 0$, with probability at least $1 - \frac{\tau}{(1-\delta^2_{\mathrm{samp}})^{1/2}} \geq 1 - 2\tau$ we have that $\langle g, \mathbf{U} v_j \rangle^2 \geq \tau^2$. The proof is completed by taking $\tau = \frac{1}{10k}$ and $a_{\mathrm{anti}} = 1/121$. □

We can now invoke the boosting result of [LL18] stated in Theorem 8.6.1.

Corollary 8.8.16. *Let $\mathcal{D}$ be a mixture of hyperplanes in $\mathbb{R}^d$ with directions $\{v_j\}$, minimum mixing weight $p_{\min}$, and separation Δ. There exist constants $a_{\mathrm{sep}}, a_{\mathrm{eig}} > 0$ for which the following holds.*

Let $\zeta \triangleq a_{\mathrm{sep}} \Delta \cdot k \cdot \min\left(\frac{a_{\mathrm{eig}}}{k^3}, \frac{p_{\min}}{64}\right)$. There is an algorithm $(\mathcal{D}, v, \varepsilon, \delta)$ which, given any $\varepsilon > 0$, $\delta > 0$, and $v \in \mathbb{R}^d$ for which there exists $j \in [k]$ with $\|w_j - v\|_2 \leq \frac{\zeta}{a_{\mathrm{eig}} k^3}$, draws $T \cdot M$ samples from $\mathcal{D}$ for

$$T = O\left(p^{-2}_{\min} d \ln(\zeta/\varepsilon)\right) \quad \textit{and} \quad M = \mathrm{poly}\left(\Delta^{-1}, p^{-1}_{\min}, k, \log T\right) \cdot \ln(1/\delta),$$

runs in time $T \cdot M \cdot d$, and outputs $\widetilde{v} \in \mathbb{R}^d$ for which either $\|v_j - \widetilde{v}\|_2 \leq \varepsilon$ or $\|v_j + \widetilde{v}\|_2 \leq \varepsilon$ with probability at least $1 - \delta$.

Proof. By the same argument as in Fact 8.8.15, we know there exists some $a'_{\mathrm{anti}} > 0$ for which $\Pr\left[|\langle v, w\rangle| \geq a'_{\mathrm{anti}}/k^2\right] \geq 1/40$. For every $i \neq j$, we know there exists some $a_{\mathrm{conc}} > 0$ for which

$$\Pr\left[|\langle v_i - v_j, w\rangle| \leq \frac{a_{\mathrm{conc}}}{\sqrt{k}}\|v_i - v_j\|_2\right] \geq 1 - \frac{1}{40k^2}.$$

By a union bound over the former event, the latter event for every $i \neq j$, and the event in Fact 8.8.15, the probability all of these events happen is at least $3/4$. Condition on these events.

Given $v \in \mathbb{S}^{d-1}$ satisfying $\|v - v_j\|_2 \leq \delta$, and $w = \frac{g\mathbf{U}}{\|g\mathbf{U}\|_2}$ for $g \sim \mathcal{N}(0, \mathrm{Id}_k)$, note that $u \triangleq -v'/\langle v, w\rangle \in \mathbb{R}^{d-1}$ satisfies

$$\begin{aligned}
\|u - v'_j\|_2 &\leq \|-\frac{v'}{\langle v, w\rangle} + \frac{v'_j}{\langle v, w\rangle}\|_2 + \|-\frac{v'_j}{\langle v, w\rangle} + \frac{v'_j}{\langle v_j, w\rangle}\| \\
&= \frac{\delta}{|\langle v, w\rangle|} + \|v'_j\|_2 \cdot \left|\frac{1}{\langle v, w\rangle} - \frac{1}{\langle v_j, w\rangle}\right| \\
&\leq \frac{\delta}{|\langle v, w\rangle|} + \frac{|\langle v - v_j, w\rangle|}{|\langle v, w\rangle| \cdot |\langle v_j, w\rangle|} \\
&\leq \frac{\delta}{|\langle v, w\rangle|} + \frac{\|v - v_j\|_2 \cdot \|w\|_2}{|\langle v, w\rangle| \cdot |\langle v_j, w\rangle|} \\
&\leq O(\delta \cdot k^2),
\end{aligned}$$

where in the first step we use the triangle inequality, in the fourth step we use Cauchy-Schwarz, and in the fifth step we use the events we conditioned on. In other words, when $\mathcal{D}$ is regarded as a mixture of linear regressions $\mathcal{D}'$ under the direction w, v' is a warm start close to v'_j.

Next, we check that this mixture of linear regressions $\mathcal{D}'$ is well-separated. For any $i \neq j$, let $v''_i, v''_j \in \mathbb{R}^d$ be the vectors $(-v'_i, \langle v_i, w\rangle)$ and $(-v'_j, \langle v_j, w\rangle)$ respectively. Then

$$\begin{aligned}
\|\frac{v'_i}{\langle v_i, w\rangle} - \frac{v'_j}{\langle v_j, w\rangle}\|_2^2 &= \|\frac{v''_i}{\langle v_i, w\rangle} - \frac{v''_j}{\langle v_j, w\rangle}\|_2^2 \\
&= \frac{\|v''_i\|_2^2}{\langle v_i, w\rangle^2} + \frac{\|v''_j\|_2^2}{\langle v_j, w\rangle^2} - \frac{2\langle v''_i, v''_j\rangle}{\langle v_i, w\rangle\langle v_j, w\rangle}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\|v_i\|_2^2}{\langle v_i, w\rangle^2} + \frac{\|v_j\|_2^2}{\langle v_j, w\rangle^2} - \frac{2\langle v_i, v_j\rangle}{\langle v_i, w\rangle\langle v_j, w\rangle} \\
&= \frac{1}{\langle v_i, w\rangle^2} + \frac{1}{\langle v_j, w\rangle^2} - \frac{2 - \|v_i - v_j\|_2^2}{\langle v_i, w\rangle\langle v_j, w\rangle} \\
&\geq \left(\frac{1}{\langle v_i, w\rangle} - \frac{1}{\langle v_j, w\rangle}\right)^2 + \frac{\|v_i - v_j\|_2^2}{\langle v_i, w\rangle\langle v_j, w\rangle}, \qquad (8.57)
\end{aligned}
$$

where in the third step we used the fact that v_i and v_j are the same as v_i'' and v_j'' up to a change of basis and a change of sign of the entry corresponding to the w direction. Recall that we are assuming without loss of generality that $\langle v_i, w\rangle \geq 0$ for all $i \in [k]$, so (8.57) is at least $\frac{\|v_i - v_j\|_2^2}{\langle v_i, w\rangle\langle v_j, w\rangle} \geq \Omega(\Delta^2 \cdot k^2)$.

Lastly, by Lemma 8.8.14, we have that the covariances of the components of this mixture of linear regressions $\mathcal{D}'$ have eigenvalues all lying in $[\Omega(1/k^3), 1]$. So that the scaling is consistent with Theorem 8.6.1, consider the mixture of linear regressions $\widetilde{\mathcal{D}}$ from which one can sample by drawing (x, y) from $\mathcal{D}'$ and taking $(x \cdot \Theta(k^3), y \cdot \Theta(k^3))$. $\mathcal{D}'$ has the same regressors as $\mathcal{D}'$ and thus the same separation $\Omega(\Delta \cdot k)$, but its components' covariances have eigenvalues all lying in $[1, \Theta(k^3)]$. By Theorem 8.6.1, if we take $\zeta = \Omega(\Delta\cdot k)\cdot\min\left(\frac{1}{\Theta(k^3)}, \frac{p_{\min}}{64}\right)$, then the algorithm of [LL18] converges to an ε-close estimate for v_j' provided $\|v' - v_j'\|_2 \leq \zeta/\Theta(k^3)$. □

8.8.5 Learning All Hyperplanes

With HyperplaneMomentDescent and HyperplaneBoost in hand, it is now straightforward to obtain an algorithm that learns all components of a mixture of hyperplanes, see Algorithm 38.

We can complete the proof of Theorem 8.8.1.

Proof of Theorem 8.8.1. By Lemma 8.8.8, every v_i' in LearnHyperplanes is $\frac{\zeta}{a_{\mathrm{eig}}k^3}$-close (up to signs) to a direction $v_{i'}$ of $\mathcal{D}$, and by Corollary 8.8.16, HyperplaneBoost improves this to a vector $\widetilde{v}_i$ for which $\|\widetilde{v}_i - v_{i'}\|_2 \leq \varepsilon_{\mathrm{boost}}$, where

$$\varepsilon_{\mathrm{boost}} = \min\{\varepsilon, \mathrm{poly}(p_{\min}, \Delta, 1/k, 1/d)^{k^{3/5}\ln(1/p_{\min})}\}.$$

```
Algorithm 38: LearnHyperplanes($\mathcal{D}, \delta, \varepsilon$)
  Input: Sample access to mixture of hyperplanes $\mathcal{D}$ with separation $\Delta$ and
         directions $\{v_i\}$, failure probability $\delta$, error $\varepsilon$
  Output: List of vectors $\mathcal{L} \triangleq \{\widetilde{v}_1, ..., \widetilde{v}_k\}$ for which there is a permutation
          $\pi : [k] \to [k]$ and signs $\varepsilon_1, ..., \varepsilon_k \in \{\pm 1\}$ for which $\|\widetilde{v}_i - \varepsilon_i v_{\pi(i)}\|_2 \le \varepsilon$ for all
          $i \in [k]$, with probability at least $1 - \delta$.
  $\delta' \leftarrow \delta/2k$
  $\zeta \leftarrow a_{\mathrm{sep}}\Delta \cdot k \cdot \min\left(\frac{a_{\mathrm{eig}}}{k^3}, \frac{p_{\min}}{64}\right)$
  $\varepsilon_{\mathrm{HMD}} \leftarrow \frac{\zeta}{a_{\mathrm{eig}}k^3}$.
  $\varepsilon_{\mathrm{boost}} \leftarrow \min\{\varepsilon, \mathrm{poly}(p_{\min}, \Delta, 1/k, 1/d)^{k^{3/5}\ln(1/p_{\min})}\}$.
  for $i \in [k]$ do
      $v'_i \leftarrow$ HyperplaneMomentDescent($\mathcal{D}, \delta', \varepsilon_{\mathrm{FMD}}$)
      $\widetilde{v}_i \leftarrow$ HyperplaneBoost($\mathcal{D}, v'_i, \varepsilon_{\mathrm{boost}}, \delta'$)
      Henceforth when sampling from $\mathcal{D}$, ignore all samples $x \in \mathbb{R}^d$ for which
        $|\langle \widetilde{v}_i, x\rangle| \le \varepsilon_{\mathrm{boost}} \cdot \mathrm{poly}(\log d)$.
```

As a result, only a $\mathrm{poly}(p_{\min}, \Delta, 1/k, 1/d)^{k^{3/5}\ln(1/p_{\min})}$ fraction of subsequent samples will be removed, and the resulting error can be absorbed into the sampling error that goes into subsequent calls to L2Estimate and subsequent matrices $\widehat{\mathbf{M}}_a^{(N)} \in \mathbb{R}^{d\times d}$ that we run ApproxBlockSVD on, in the remainder of LearnWithoutNoise. □

8.9 Boosting Down the Cosine Integral

The main result that we show in this section is the following local convergence guarantee for Boost.

Theorem 8.9.1. *There are absolute constants $C, C' > 0$ such that the following holds. Let $\varepsilon > 0$, and let $\mathcal{D}$ be any mixture of spherical linear regressions with separation Δ, noise rate $\varsigma \le C' \cdot (p_{\min} \cdot \varepsilon \cdot \Delta^4)^{1/5}$. Suppose $\|v - w_{i^*}\|_2 \le \Delta/\gamma$ for $\gamma = C \cdot p_{\min}^{1/4}$. Then* Boost($\mathcal{D}, v, \varepsilon, \delta$) *(Algorithm 39) returns v^* satisfying $\|v^* - w_{i^*}\|_2 \le \varepsilon$. Additionally, it has sample complexity*

$$\widetilde{O}\left(d \cdot \mathrm{poly}(1/\varepsilon, 1/\Delta) \cdot (\ln(1/\varepsilon) \cdot \ln(1/p_{\min}))^{O(\ln(1/p_{\min}))}\right)$$

and runtime

$$\widetilde{O}\left(d^2 \cdot \mathrm{poly}(1/\varepsilon, 1/\Delta) \cdot (\ln(1/\varepsilon) \cdot \ln(1/p_{\min}))^{O(\ln(1/p_{\min}))}\right).$$

Remark 8.9.2. *Note that our boosting algorithm can tolerate a warm start at distance $O(\Delta p_{\min}^{-1/4})$, whereas that of [LL18] can only tolerate one at distance $O(\Delta p_{\min})$ (see Theorem 8.6.1). Our algorithm can also tolerate regression noise as large as $O(p_{\min}^{1/5}\varepsilon^{1/5}\Delta^{4/5})$. In particular, if $\varepsilon = o(p_{\min}^{1/20}\Delta^{1/5})$, then our algorithm* Boost *can tolerate noise rate $\varsigma = \omega(\varepsilon)$.*

In Section 8.9.1 we recall the boosting algorithm of [LL18] to motivate the high-level blueprint for our argument. In Section 8.9.2 we give the full specification of our boosting algorithm and a proof of Theorem 8.9.1.

8.9.1 Background: Gravitational Allocation

In [LL18], Li and Liang boost a warm start to a fine estimate for one of the w_i's by performing stochastic gradient descent on the (regularized) gravitational potential objective

$$h(v) = \mathbb{E}_{x,y}[\ln(|\langle x, v\rangle - y| + \xi)]$$

for some $\xi > 0$ which is introduced to ensure smoothness even when $v = w_i$ for some $i \in [k]$. We emphasize that this objective is *concave*. For any $i^* \in [k]$, the inner product between the expected gradient step and $w_{i^*} - v^{(t)}$, where $v^{(t)}$ is the current iterate, is given by

$$\begin{aligned}\langle -\Delta h(v^{(t)}), w_{i^*} - v^{(t)}\rangle &= -\mathbb{E}_{x,y}\left[\frac{\operatorname{sgn}(\langle x, v\rangle - y)\cdot\langle x, w_{i^*} - v^{(t)}\rangle}{|\langle x, v\rangle - y| + \xi}\right] \\ &= \frac{1}{k}\sum_{i=1}^{k} p_i \cdot \mathbb{E}_{x\sim\mathcal{N}(0,\mathrm{Id})}\left[\frac{\operatorname{sgn}(\langle x, w_i - v^{(t)}\rangle)\cdot\langle x, w_{i^*} - v^{(t)}\rangle}{|\langle x, w_i - v^{(t)}\rangle| + \xi}\right].\end{aligned}$$

They argue that provided $\|v^{(0)} - w_{i^*}\|_2 \le O(\Delta/k)$, the contribution of the i^*-th summand dominates that of all other summands, so the correlation of the gradient step with $w_{i^*} - v^{(t)}$ is sufficiently large that each step contracts the distance to w_{i^*} appreciably.

8.9.2 Boosting via the Cosine Integral

Here we argue that a warm start of $\|v^{(0)} - w_{i^*}\|_2 \le O(\Delta \cdot p_{\min}^{1/4})$ is sufficient if we run gradient descent not on the gravitational potential objective, but on the *cosine integral objective.*

Concretely, we propose Algorithm 39 below for boosting.

```
Algorithm 39: BOOST($\mathcal{D}, v, \varepsilon, \delta$)
Input: Mixture of linear regressions $\mathcal{D}$ with separation $\Delta$ and noise rate
       $\varsigma \leq O((p_{\min} \cdot \varepsilon \cdot \Delta^4)^{1/5})$, warm start $v$, accuracy $\varepsilon$, failure probability $\delta$
$\gamma \leftarrow C \cdot p_{\min}^{1/4}$.
$v^{(0)} \leftarrow v$, $T \leftarrow O(d \cdot \Delta^8/\varepsilon^8 \cdot \ln(\Delta/\gamma\varepsilon)$.
$\delta' \leftarrow \frac{\delta}{2T}$.
$\overline{\sigma} \leftarrow 4$.
for $t = 0, ..., T-1$ do
    $\xi_t \leftarrow$ ESTIMATEMINVARIANCE$(\mathcal{F}_t, \overline{\sigma}, \varepsilon/10, \Omega(\ln(1/p_{\min})), \delta')/1.1$
    if $\xi_t \cdot (1.1/0.9) \leq \varepsilon$ then
        break
    Draw $N = \mathrm{poly}(1/\xi_t, 1/\Delta, \ln(\delta'))$ fresh samples from $\mathcal{D}$, call them
      $(x_1, y_1), ..., (x_N, y_N)$.
    Form the empirical gradient
        $$\delta_t = -\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[|\langle x_i, v^{(t)}\rangle - y_i| \geq \xi_t] \cdot \frac{\cos(\xi_t^{-1}\pi|\langle x_i, v^{(t)}\rangle - y_i|)}{\langle x_i, v^{(t)}\rangle - y_i} \cdot x_i.$$
    Set learning rate $\eta_t \leftarrow \frac{\xi_t^5}{2d\Delta^4} \cdot \|w_{i^*} - v^{(t)}\|_2$ and define
        $$v^{(t+1)} \leftarrow v^{(t)} - \eta_t \delta_t. \qquad (8.58)$$
return $v^{(T)}$.
```

Remark 8.9.3. *An obvious caveat for our result is the exponential dependence on* $\ln(1/p_{\min})$, *which comes from the need to compute the regularization parameter* ξ_t *at each step. Similar to the* ξ *in the gravitational potential objective of [LL18], the* ξ_t *in* BOOST *is to ensure smoothness. In our case, we need* ξ_t *to be a lower bound for* $\|w_{i^*} - v^{(t)}\|_2$*, and the rate of contraction decreases as* ξ_t *decreases (see Lemma 8.9.4 below).*

To show Theorem 8.9.1, we first show that if ξ_t is chosen to be sufficiently small at each step, $\|w_{i^*} - v^{(t)}\|_2$ is guaranteed to contract.

Lemma 8.9.4. *Let* $C, C' > 0$ *be the constants in Theorem 8.9.1.*

For any t *and* $\delta > 0$, *if* $\varepsilon/10 \leq \|w_{i^*} - v^{(t)}\|_2 \leq \Delta/\gamma$ *for* $\gamma = C \cdot p_{\min}^{1/4}$ *and* ξ_t *satisfies* $\xi_t \leq \|w_{i^*} - v^{(t)}\|_2$*, then for* $N = \mathrm{poly}(1/\xi_t, 1/\Delta, \ln(1/\delta))$*, we have with probability at least*

$1-\delta$ over the N samples used to form the empirical gradient that

$$\left(1-\frac{\xi_t^4}{\sqrt{d}\Delta^4}\right)\|w_{i^*}-v^{(t)}\|_2^2 \le \|w_{i^*}-v^{(t+1)}\|_2^2 \le \left(1-\frac{\xi_t^8}{4d\Delta^8}\right)\cdot\|w_{i^*}-v^{(t)}\|_2^2.$$

Proof. The key step is to lower bound the correlation between the negative gradient $-\mathbb{E}[\delta_t]$ and the direction $w_{i^*}-v^{(t)}$ in which we would like to move. We have that

$$\begin{aligned}
&\langle -\mathbb{E}[\delta_t], w_{i^*}-v^{(t)}\rangle \\
&= \mathbb{E}_{x,y}\left[\mathbb{1}[|\langle x, v^{(t)}\rangle - y| \ge \xi_t]\cdot\frac{\cos\left(\xi_t^{-1}\pi|\langle x, v\rangle - y|\right)\cdot\langle x, w_{i^*}-v^{(t)}\rangle}{\langle x, v^{(t)}\rangle - y_i}\right] \\
&= \sum_{i=1}^k p_i\mathbb{E}_{\substack{x\sim\mathcal{N}(0,\mathrm{Id})\\ g\sim\mathcal{N}(0,\varsigma^2)}}\left[\mathbb{1}[|\langle x, w_i-v^{(t)}\rangle - g| \ge \xi_t]\cdot\frac{-\cos\left(\xi_t^{-1}\pi|\langle x, w_i-v^{(t)}\rangle - g|\right)\cdot\langle x, w_{i^*}-v^{(t)}\rangle}{\langle x, w_i-v^{(t)}\rangle - g}\right]
\end{aligned}\tag{8.59}$$

where the last step follows from the fact that (x,y) comes from component i with probability p_i, in which case $\langle x, v^{(t)}\rangle - y_i = -\langle x, w_i - v^{(t)}\rangle$.

For every $i\in[k]$, define

$$\beta_i \triangleq \left(\|w_i-v^{(t)}\|_2^2+\varsigma^2\right)^{1/2} \quad \text{and} \quad \nu_i \triangleq \frac{\|w_i-v^{(t)}\|_2^2}{\|w_i-v^{(t)}\|_2^2+\varsigma^2}.$$

We have the naive bounds

$$\|w_i-v^{(t)}\|_2 \le \beta_i \le \max\left\{\|w_i-v^{(t)}\|_2, \varsigma\right\}\cdot\sqrt{2} \tag{8.60}$$

and

$$\min\left\{\frac{1}{2}, \frac{\|w_i-v^{(t)}\|_2^2}{2\varsigma^2}\right\} \le \nu_i \le 1 \tag{8.61}$$

for all i. Then by Lemma 8.9.5 below, we can bound the $i\neq i^*$ and $i=i^*$ terms of (8.59) to get

$$\langle -\mathbb{E}[\delta_t], w_{i^*}-v^{(t)}\rangle \ge p_{i^*}\frac{0.22\xi_t^3\nu_{i^*}}{\beta_{i^*}^3} - \sum_{i\neq i^*}p_i\frac{\|w_{i^*}-v^{(t)}\|_2}{\|w_i-v^{(t)}\|_2}\cdot\frac{0.26\xi_t^3\nu_i}{\beta_i^3}$$

$$
\begin{aligned}
&\geq \xi_t^3 \left[p_{i^*} \cdot \frac{0.22\nu_{i^*}}{\beta_{i^*}^3} - \sum_{i \neq i^*} p_i \frac{0.26\|w_{i^*} - v^{(t)}\|_2}{\|w_i - v^{(t)}\|_2^4} \right] \\
&\geq \xi_t^3 \left[p_{i^*} \cdot \frac{0.22\nu_{i^*}}{\beta_{i^*}^3} - \frac{0.26\|w_{i^*} - v^{(t)}\|_2}{(\Delta/2)^4} \right], \qquad (8.62)
\end{aligned}
$$

where in the second step we invoked the lower and upper bounds of (8.60) and (8.61) respectively, and in the third step we used the fact that for every $i \neq i^*$,

$$
\|w_i - v^{(t)}\|_2 \geq \|w_i - w_{i^*}\|_2 - \|w_{i^*} - v^{(t)}\|_2 \geq \Delta/2.
$$

We proceed by casework based on the relation between $\|w_{i^*} - v^{(t)}\|_2$ and ς.

Case 1. $\|w_{i^*} - v^{(t)}\|_2 \geq \varsigma$.

In this case, we know that $\beta_{i^*} \leq \sqrt{2}\|w_{i^*} - v^{(t)}\|_2 \leq \sqrt{2}\Delta/\gamma$ and $\nu_{i^*} \geq 1/2$ by (8.60) and (8.61). From (8.62) we get that

$$
\langle -\mathbb{E}[\delta_t], w_{i^*} - v^{(t)} \rangle \geq \xi_t^3 \|w_{i^*} - v^{(t)}\|_2 \cdot \left[p_{\min} \cdot \frac{0.11 \cdot (2^{-3/2})}{(\Delta/\gamma)^4} - \frac{0.26}{(\Delta/2)^4} \right]
$$

So there is an absolute constant $C > 0$ such that for $\gamma = C \cdot p_{\min}^{1/4}$, we have that

$$
\langle -\mathbb{E}[\delta_t], w_{i^*} - v^{(t)} \rangle \geq \xi_t^3 \|w_{i^*} - v^{(t)}\|_2 \cdot \Delta^{-4} \qquad (8.63)
$$

Case 2. $\|w_{i^*} - v^{(t)}\|_2 \leq \varsigma$.

In this case, we know that $\beta_{i^*} \leq \sqrt{2}\varsigma$ and $\nu_{i^*} \geq \frac{\|w_{i^*} - v^{(t)}\|_2^2}{2\varsigma^2}$ by (8.60) and (8.61). From (8.62) we get that

$$
\begin{aligned}
\langle -\mathbb{E}[\delta_t], w_{i^*} - v^{(t)} \rangle &\geq \xi_t^3 \cdot \left[p_{\min} \cdot \frac{0.22}{\varsigma^3 \cdot 2^{3/2}} \cdot \frac{\|w_{i^*} - v^{(t)}\|_2^2}{2\varsigma^2} - \frac{0.26\|w_{i^*} - v^{(t)}\|_2}{(\Delta/2)^4} \right] \\
&\geq \xi_t^3 \|w_{i^*} - v^{(t)}\|_2 \cdot \left[p_{\min} \cdot \frac{0.22 \cdot (2^{-5/2})}{\varsigma^5} \cdot \|w_{i^*} - v^{(t)}\|_2 - \frac{0.26}{(\Delta/2)^4} \right] \\
&\geq \xi_t^3 \|w_{i^*} - v^{(t)}\|_2 \cdot \left[p_{\min} \cdot \frac{0.22 \cdot (2^{-5/2}) \cdot (\varepsilon/3)}{\varsigma^5} - \frac{0.26}{(\Delta/2)^4} \right].
\end{aligned}
$$

By taking $\gamma = C \cdot p_{\min}^{1/4}$ as in the previous case, we see that there exists some absolute constant $C' > 0$ such that for $\varsigma \leq C' \cdot (p_{\min} \cdot \varepsilon \cdot \Delta^4)^{1/5}$, (8.63) still holds.

To show moving in the direction opposite the *empirical* gradient suffices, we need concentration. First note that for every sample (x, y),

$$\|\mathbb{1}[|\langle x, v^{(t)}\rangle - y| \geq \xi_t] \cdot \frac{\cos(\xi_t^{-1}\pi|\langle x, v^{(t)}\rangle - y|)}{\langle x, v^{(t)}\rangle - y} \cdot x\|_2 \leq \frac{\|x\|}{\xi_t},$$

and likewise

$$\left|\left\langle \mathbb{1}[|\langle x, v^{(t)}\rangle - y| \geq \xi_t] \cdot \frac{\cos(\xi_t^{-1}\pi|\langle x, v^{(t)}\rangle - y|)}{\langle x, v^{(t)}\rangle - y} \cdot x, \frac{w_{i^*} - v^{(t)}}{\|w_{i^*} - v^{(t)}\|_2}\right\rangle\right| \leq \frac{\left|\left\langle x, \frac{w_{i^*} - v^{(t)}}{\|w_{i^*} - v^{(t)}\|_2}\right\rangle\right|}{\xi_t}.$$

Furthermore, by (8.63), the expected gradient satisfies

$$\left\langle -\mathbb{E}[\delta_t], \frac{w_{i^*} - v^{(t)}}{\|w_{i^*} - v^{(t)}\|_2}\right\rangle \geq \xi_t^3 \Delta^{-4}.$$

By standard Gaussian concentration, for some $N \geq \mathrm{poly}(\xi_t^{-1}, \Delta^{-1}, \ln(1/\delta))$, we get that with probability at least $1 - \delta/3$,

$$\|\delta_t\| \leq \frac{2\sqrt{d}}{\xi_t} \quad \text{and} \quad \left\langle -\delta_t, \frac{w_{i^*} - v^{(t)}}{\|w_{i^*} - v^{(t)}\|_2}\right\rangle \geq \frac{1}{2}\xi_t^3 \Delta^{-4}.$$

By (8.58),

$$\|w_{i^*} - v^{(t+1)}\|_2^2 = \|w_{i^*} - v^{(t)}\|_2^2 + \eta_t^2 \|\delta_t\|_2^2 - 2\eta_t \langle -\delta_t, w_{i^*} - v^{(t)}\rangle,$$

so by taking learning rate $\eta_t \triangleq \frac{\xi_t^5}{2d\Delta^4} \cdot \|w_{i^*} - v^{(t)}\|_2$, we ensure that

$$\|w_{i^*} - v^{(t+1)}\|_2^2 \leq \left(1 - \frac{\xi_t^8}{4d\Delta^8}\right) \|w_{i^*} - v^{(t)}\|_2^2.$$

At the same time, from the naive bounds $\|\delta_t\|_2^2 \geq 0$ and $2\eta_t \langle -\delta_t, w_{i^*} - v^{(t)}\rangle \leq \frac{\xi_t^4}{\sqrt{d}\Delta^4} \|w_{i^*} - v^{(t)}\|_2^2$ which follows by Cauchy-Schwarz, we also have

$$\|w_{i^*} - v^{(t+1)}\|_2^2 \geq \left(1 - \frac{\xi_t^4}{\sqrt{d}\Delta^4}\right) \|w_{i^*} - v^{(t)}\|_2^2.$$

□

To complete the proof of Lemma 8.9.4, it remains to prove the following lemma which was crucial to establishing (8.62).

Lemma 8.9.5. *For any vectors $a, b \in \mathbb{R}^d$ and $\xi \le \|b\|_2$, we have that*

$$\left|\mathbb{E}_{\substack{x\sim\mathcal{N}(0,Id)\\ g\sim\mathcal{N}(0,\varsigma^2)}}\left[\mathbb{1}[|\langle b,x\rangle + g| \ge \xi]\cdot\frac{-\cos\left(\xi^{-1}\pi|\langle b,x\rangle+g|\right)}{\langle b,x\rangle+g}\cdot\langle a,x\rangle\right]\right| \le \frac{\|a\|_2}{\|b\|_2}\cdot\frac{\|b\|_2^2}{\varsigma^2+\|b\|_2^2}\cdot\frac{0.26\xi^3}{(\varsigma^2+\|b\|_2^2)^{3/2}}. \tag{8.64}$$

Furthermore, we have that for $a = b$,

$$\mathbb{E}_{\substack{x\sim\mathcal{N}(0,Id)\\ g\sim\mathcal{N}(0,\varsigma^2)}}\left[\mathbb{1}[|\langle b,x\rangle + g| \ge \xi]\cdot\frac{-\cos\left(\xi^{-1}\pi|\langle b,x\rangle+g|\right)}{\langle b,x\rangle+g}\cdot\langle b,x\rangle\right] = \frac{\|b\|_2^2}{\varsigma^2+\|b\|_2^2}\cdot\frac{[0.22,0.26]\cdot\xi^3}{(\varsigma^2+\|b\|_2^2)^{3/2}}.$$

Proof. For notational convenience, given $x \sim \mathcal{N}(0,\mathrm{Id})$, let $\mathcal{E}_\xi$ denote the event that $|\langle b,x\rangle + g| \ge \xi$. We may write

$$a = \frac{\rho\|a\|_2}{\|b\|_2}\cdot b + \sqrt{1-\rho^2}\cdot b^\perp$$

for $\rho = \frac{\langle a,b\rangle}{\|a\|_2\|b\|_2}$ and $b^\perp \in \mathbb{S}^{d-1}$ orthogonal to b. Then the left-hand side of (8.64) can be written as

$$\begin{aligned}
&\mathbb{E}_{\substack{x\sim\mathcal{N}(0,\mathrm{Id})\\ g\sim\mathcal{N}(0,\varsigma^2)}}\left[\mathbb{1}[\mathcal{E}_\xi]\cdot\frac{-\cos\left(\xi^{-1}\pi|\langle b,x\rangle+g|\right)}{\langle b,x\rangle+g}\cdot\langle a,x\rangle\right]\\
&= \mathbb{E}_{\substack{x\sim\mathcal{N}(0,\mathrm{Id})\\ g\sim\mathcal{N}(0,\varsigma^2)}}\left[\mathbb{1}[\mathcal{E}_\xi]\cdot\frac{-\cos\left(\xi^{-1}\pi|\langle b,x\rangle+g|\right)}{\langle b,x\rangle+g}\cdot\frac{\rho\|a\|_2}{\|b\|_2}\cdot\langle b,x\rangle\right]\\
&= -\frac{\rho\|a\|_2}{\|b\|_2}\mathbb{E}_{\substack{x\sim\mathcal{N}(0,\mathrm{Id})\\ g\sim\mathcal{N}(0,\varsigma^2)}}\left[\mathbb{1}[\mathcal{E}_\xi]\frac{\cos\left(\xi^{-1}\pi|\langle b,x\rangle+g|\right)}{\langle b,x\rangle+g}\cdot\langle b,x\rangle\right]\\
&= -\frac{\rho\|a\|_2}{\|b\|_2}\mathbb{E}_{\substack{g\sim\mathcal{N}(0,\varsigma^2)\\ g'\sim\mathcal{N}(0,\|b\|_2^2)}}\left[\mathbb{1}[\mathcal{E}_\xi]\frac{\cos\left(\xi^{-1}\pi(g+g')\right)}{g+g'}\cdot g'\right]\\
&= -\frac{\rho\|a\|_2}{\|b\|_2}\cdot\frac{\|b\|_2^2}{\varsigma^2+\|b\|_2^2}\cdot\mathbb{E}_{\substack{g\sim\mathcal{N}(0,\varsigma^2)\\ g'\sim\mathcal{N}(0,\|b\|_2^2)}}\left[\mathbb{1}[\mathcal{E}_\xi]\cos\left(\xi^{-1}\pi(g+g')\right)\right]\\
&= -\frac{\rho\|a\|_2}{\|b\|_2}\cdot\frac{\|b\|_2^2}{\varsigma^2+\|b\|_2^2}\cdot\mathbb{E}_{g\sim\mathcal{N}(0,\varsigma^2+\|b\|_2^2)}\left[\mathbb{1}[|g|\ge\xi]\cos\left(\xi^{-1}\pi g\right)\right]
\end{aligned}$$

where the first step follows from the fact that $\langle b,x\rangle$ and $\langle b^\perp,x\rangle$ are independent mean-zero random variables, the third step follows by the fact that $\cos(\cdot)$ is even, and the penultimate step follows from the fact that we may decompose g' in terms of $g+g'$ as

$$g' = \frac{\|b\|_2^2}{\varsigma^2 + \|b\|_2^2}(g+g') + h$$

for Gaussian h independent of $g+g'$.

We conclude the proof of the first half of the lemma by noting that $|\rho| \le 1$ and appealing to the upper bound in Lemma 8.9.6, where we take $\beta = (\varsigma^2 + \|b\|_2^2)^{1/2}$.

Next, the upper bound in the second half of the lemma follows immediately from Lemma 8.9.6. Finally, for the lower bound in the second half of the lemma, the lower bound in Lemma 8.9.6 gives

$$\mathbb{E}_{\substack{x\sim\mathcal{N}(0,\mathrm{Id})\\ g\sim\mathcal{N}(0,\varsigma^2)}}\left[\mathbb{1}[|\langle b,x\rangle + g| \ge \xi]\cdot\frac{-\cos\left(\xi^{-1}\pi|\langle b,x\rangle+g|\right)}{\langle b,x\rangle+g}\cdot\langle b,x\rangle\right] \ge \frac{0.23\xi^3}{(\varsigma^2+\|b\|_2^2)^{3/2}} - \exp\left(-\frac{\pi^2(\varsigma^2+\|b\|_2^2)}{2\xi^2}\right),$$

and we conclude by noting that for $\xi \le \beta$, $\exp(-\frac{\beta^2}{2\xi^2}) \le \exp(-\pi^2/2)/\beta^3 \le 0.01/\beta^3$. □

Lemma 8.9.6. *For any $\beta,\xi > 0$ for which $\xi \le \beta$, we have that*

$$\exp(-\frac{\pi^2\beta^2}{2\xi^2}) - \mathbb{E}_{g\sim\mathcal{N}(0,\beta^2)}\left[\mathbb{1}[\mathcal{E}_\xi]\cdot\cos(\xi^{-1}\pi|g|)\right] \in \left[\frac{0.23\xi^3}{\beta^3}, \frac{0.26\xi^3}{\beta^3}\right]. \tag{8.65}$$

Proof. We can rewrite the LHS in (8.65) as follows:

$$\mathrm{LHS} = \underbrace{\exp(-\frac{\pi^2\beta^2}{2\xi^2}) - \frac{1}{\beta\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{x^2}{2\beta^2}\right)\cos(\pi x/\xi)\mathrm{d}x}_{\textcircled{I}} + \underbrace{\frac{2}{\beta\sqrt{2\pi}}\int_0^{\xi}\exp\left(-\frac{x^2}{2\beta^2}\right)\cos(\pi x/\xi)\mathrm{d}x}_{\textcircled{II}}$$

Using Claim 8.9.7, we can show $\textcircled{I} = 0$. Using Claim 8.9.8, we can upper and lower bound $\textcircled{II}$. □

Claim 8.9.7. *We have*

$$\frac{1}{\beta\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{x^2}{2\beta^2}}\cos(\pi x/\xi)\mathrm{d}x = \exp\left(-\frac{\pi^2\beta^2}{2\xi^2}\right).$$

Proof. Let $\upsilon = \frac{\xi^2}{2\pi^2\beta^2}$. Noting that $\cos(x) = \mathrm{Re}(e^{-\mathbf{i}x})$, one can compute LHS by a standard contour integral.

$$\begin{aligned}
\mathrm{LHS} = & \frac{1}{\beta\sqrt{2\pi}} \cdot \frac{\xi}{\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{\xi^2 x^2}{2\pi^2\beta^2}\right) \cdot \cos(x) \mathrm{d}x \\
= & \frac{1}{\beta\sqrt{2\pi}} \cdot \frac{\xi}{\pi} \int_{-\infty}^{\infty} \exp(-\upsilon x^2) \cdot \cos(x) \mathrm{d}x \\
= & \frac{1}{\beta\sqrt{2\pi}} \cdot \frac{\xi}{\pi} \cdot \mathrm{Re} \int_{-\infty}^{\infty} \exp\left(-\upsilon x^2 - \mathbf{i}x\right) \mathrm{d}x \\
= & \frac{1}{\beta\sqrt{2\pi}} \cdot \frac{\xi}{\pi} \cdot \mathrm{Re} \int_{-\infty}^{\infty} \exp(-\upsilon(x + \mathbf{i}/(2\upsilon))^2 - 1/(4\upsilon)) \mathrm{d}x \\
= & \frac{\exp(-1/(4\upsilon))}{\beta\sqrt{2\pi}} \cdot \frac{\xi}{\pi} \cdot \mathrm{Re} \int_{-\infty}^{\infty} \exp(-\upsilon(x + \mathbf{i}/(2\upsilon))^2) \mathrm{d}x \\
= & \frac{\exp(-1/(4\upsilon))}{\beta\sqrt{2\pi}} \cdot \frac{\xi}{\pi} \cdot \mathrm{Re} \int_{-\infty+\mathbf{i}/(2\upsilon)}^{\infty+\mathbf{i}/(2\upsilon)} \exp(-\upsilon x^2) \mathrm{d}x \\
= & \frac{\exp(-1/(4\upsilon))}{\beta\sqrt{2\pi}} \cdot \frac{\xi}{\pi} \cdot \frac{\sqrt{\pi}}{\sqrt{\upsilon}} \\
= & \exp\left(-\frac{\pi^2\beta^2}{2\xi^2}\right),
\end{aligned}$$

where the second step follows from definition of υ, the third step follows from $\cos(x) = \mathrm{Re}(\exp(-\mathbf{i}x))$, the fourth step follows from $-\upsilon x^2 - \mathbf{i}x = -\upsilon(x^2 + \mathbf{i}x/\upsilon - 1/(4\upsilon^2)) - 1/4\upsilon = -\upsilon(x + \mathbf{i}/(2\upsilon))^2 - 1/(4\upsilon)$, the fifth step follows from pulling the term $\exp(-1/(4\upsilon))$ out of integral, the sixth step follows from shifting the integral range, the seventh step follows from Cauchy's theorem[2], and the last step follows from definition of υ.

Thus, we complete the proof. □

Claim 8.9.8. *Let $\xi \leq \beta$. We have*

$$\frac{2}{\beta\sqrt{2\pi}} \int_{0}^{\xi} \exp\left(-\frac{x^2}{2\beta^2}\right) \cos(\pi x/\xi) \mathrm{d}x \ \in \ [0.23\xi^3/\beta^3, 0.26\xi^3/\beta^3].$$

[2] By Cauchy's theorem, the integral around the box in the complex plane with vertices $-R$, R, $-R+\mathbf{i}/(2\upsilon)$, and $R+\mathbf{i}/(2\upsilon)$ is zero. The sum of the contributions of the edges between $-R$ and $-R+\mathbf{i}/(2\upsilon)$ and between R and $R+\mathbf{i}/(2\upsilon)$is imaginary and thus contributes 0 to the real part. If we take $R \to \mathbf{i}nfty$, we see that the integral we want to compute is the same as the one where you ignore the $\mathbf{i}/(2\upsilon)$ terms, which is a standard Gaussian integral.

Proof. We will use the bound

$$1 - \frac{x^2}{2\beta^2} \le \exp\left(-\frac{x^2}{2\beta^2}\right) \le 1 - \frac{x^2}{2\beta^2} + \frac{x^4}{8\beta^4}$$

to obtain upper and lower bounds.

Noting that $\cos(\pi x/\xi) \ge 0$ for $x \in [0, \xi/2]$ and $\cos(\pi x/\xi) \le 0$ for $x \in [\xi/2, \xi]$, we get that

$$\begin{aligned}\text{LHS} &\ge \frac{2}{\beta\sqrt{2\pi}} \int_0^{\xi/2} \cos(\pi x/\xi) \cdot \left(1 - \frac{x^2}{2\beta^2}\right) \mathrm{d}x + \frac{2}{\beta\sqrt{2\pi}} \int_{\xi/2}^{\xi} \cos(\pi x/\xi) \cdot \left(1 - \frac{x^2}{2\beta^2} + \frac{x^4}{8\beta^4}\right) \mathrm{d}x \\ &\ge \frac{2}{\beta\sqrt{2\pi}} \left(\frac{\xi^3}{\pi\beta^2} - 0.021 \cdot \frac{\xi^5}{\beta^4}\right) \ge \frac{0.23\xi^3}{\beta^3}\end{aligned}$$

and

$$\begin{aligned}\text{LHS} &\le \frac{2}{\beta\sqrt{2\pi}} \int_0^{\xi/2} \cos(\pi x/\xi) \cdot \left(1 - \frac{x^2}{2\beta^2} + \frac{x^4}{8\beta^4}\right) \mathrm{d}x + \frac{2}{\beta\sqrt{2\pi}} \int_{\xi/2}^{\xi} \cos(\pi x/\xi) \cdot \left(1 - \frac{x^2}{2\beta^2}\right) \mathrm{d}x \\ &\le \frac{2}{\beta\sqrt{2\pi}} \left(\frac{\xi^3}{\pi\beta^2} + 0.0002 \cdot \frac{\xi^5}{\beta^4}\right) \le \frac{0.26\xi^3}{\beta^3},\end{aligned}$$

where in the last steps we used the fact that $\xi \le \beta$. □

We can now complete the proof of Theorem 8.9.1.

Proof of Theorem 8.9.1. The only probabilistic components of BOOST is the invocation of ESTIMATEMINVARIANCE and the event of Lemma 8.9.4 holding at each step. For a given t, with probability $1 - 2\delta' = 1 - \delta/T$ these two events both hold, so by a union bound over all T iterations, the failure probability of BOOST is at most δ as desired.

We now proceed to show correctness of BOOST. Conditioned on making progress in every step of BOOST, note that $\max_i \|w_i - v^{(t)}\|_2 \le \Delta/\gamma + \|w_i - w_{i^*}\| \le \Delta/\gamma + 2 \le 4$, so $\overline{\sigma} = 4$ is always a valid upper bound for the maximum variance of any component of a univariate mixture of Gaussians $\mathcal{F}_t$ encountered over the course of BOOST. So we conclude by Lemma 8.5.7 that $\xi_t \le \|w_{i^*} - v^{(t)}\|_2$. Then because of the lower bound of Lemma 8.9.4, the inequality $\xi_t \cdot (1.1/0.9) \ge \|w_i - v^{(t)}\|_2$, and the fact that BOOST breaks out of its main loop if $\xi_t \cdot (1.1/0.9)$, we know that at all times in main loop of BOOST, $\|w_{i^*} - v^{(t)}\|_2 \ge \varepsilon/10$.

So by the upper bound of Lemma 8.9.4 and the fact that $\xi_t = \Omega(\varepsilon)$, we conclude that after $T \triangleq O\left(d \cdot \Delta^8/\varepsilon^8 \cdot \ln(\Delta/\gamma\varepsilon)\right)$ iterations, $\|w_{i^*} - v^{(T)}\|_2 \leq \varepsilon$.

For the time and sample complexity, at every time step t we must draw

$$N = \mathrm{poly}(1/\xi_t, 1/\Delta, \ln(1/\delta')) \leq \mathrm{poly}(1/\varepsilon, 1/\Delta, \ln(T), \ln(1/\delta))$$

samples to form the empirical gradient in time $d \cdot N$. We also know that each invocation of EstimateMinVariance, by Lemma 8.5.7, requires time and sample complexity

$$N' \triangleq \widetilde{O}\left((\mu_0 \cdot \ln(1/\varepsilon)\ln(1/p_{\min}))^{O(\ln(1/p_{\min}))} \cdot \ln(2T/\delta)\right).$$

So Boost requires

$$T \cdot (N + N') = \widetilde{O}\left(d \cdot \mathrm{poly}(1/\varepsilon, 1/\Delta) \cdot (\ln(1/\varepsilon) \cdot \mu_0 \cdot \ln(1/p_{\min}))^{O(\ln(1/p_{\min}))}\right)$$

samples and

$$T(d \cdot N + N') = \widetilde{O}\left(d^2 \cdot \mathrm{poly}(1/\varepsilon, 1/\Delta) \cdot (\ln(1/\varepsilon) \cdot \mu_0 \cdot \ln(1/p_{\min}))^{O(\ln(1/p_{\min}))}\right)$$

time. □

8.10 Appendix: Failure of Low-Degree Identifiability

In this section, we exhibit a pair of mixtures of spherical linear regressions which are far in parameter distance but which agree on all degree-$\Omega(k)$ moments. This demonstrates that any method which hopes to achieve sample complexity which is subexponential in k cannot rely solely on low order moments of the MLR.

First, we exhibit a pair of non-identical univariate mixtures of zero-mean Gaussians whose moments match up to degree $2k - 1$ and whose variances and mixing weights satisfy reasonable bounds. We remark that the proof, in particular the application of Borsak-Ulam, is largely inspired by that of Lemma 2.9 in [HP15].

Lemma 8.10.1. *There exist $\sigma_1, ..., \sigma_k, \sigma'_1, ..., \sigma'_k \geq 0$ such that the following holds. Let D_1 (resp. D_2) be the uniform mixture of univariate Gaussians with components $\mathcal{N}(0, \sigma_1^2), ..., \mathcal{N}(0, \sigma_k^2)$ (resp. $\mathcal{N}(0, \sigma_1'^2), ..., \mathcal{N}(0, \sigma_k'^2)$). Then*
1) there is some $i \in [k]$ for which $|\sigma_i - \sigma'_j| > \Omega(1/\sqrt{k})$ for all $j \in [k]$,
2) $|\sigma_i - \sigma_j|, |\sigma'_i - \sigma'_j| > 1/2$ for all $i \neq j$,
3) $\sigma_i, \sigma'_i \in [1/2, k+1]$ for all $i \in [k]$, and
4) D_1 and D_2 match on all moments of degree at most $2k-1$.

Proof. For each $i \in [k]$, define $\sigma_i(z) = i + \alpha z$ for $\alpha = 1/4$ and consider the map $M : \mathbb{S}^{k-1} \to \mathbb{R}^{k-1}$ given by

$$M(z)_\ell = \sum_{i=1}^{k} \sigma_i(z)^{2\ell} \quad \ell = 1, ..., k-1.$$

M is clearly continuous, so by Borsak-Ulam, there exists $z \in \mathbb{S}^{k-1}$ for which $M(z) = M(-z)$. For each $i \in [k]$, define $\sigma_i \triangleq \sigma_i(z)$ and $\sigma'_i \triangleq \sigma_i(-z)$. Then because $\alpha = 1/4$ and $\|z\|_\infty \leq 1$, $\sigma_i, \sigma'_i \in [i - 1/4, i + 1/4]$, 2) and 3) are immediately satisfied. Furthermore, this implies that for any $i \in [k]$, $|\sigma_i - \sigma'_j| > 1/2$ for all $j \neq i$. For $j = i$, $|\sigma_i - \sigma'_i| = 2|z_i|$, and because $\|z\|_2 = 1$, there must exist some i for which $|z_i| \geq \frac{1}{\sqrt{k}}$, from which 1) follows.

To see that 4) is satisfied, first note that D_1 and D_2 are mixtures of zero-mean Gaussians and thus both have odd-degree moments equal to zero. Then for any $1 \leq \ell \leq k-1$, note that the the 2ℓ-th moment of D_1 is

$$\frac{1}{k} \sum_{i=1}^{k} \sigma_i^{2\ell} \cdot (2\ell - 1)!! = \frac{1}{k}(2\ell - 1)!! \cdot M(z)_\ell.$$

Likewise, the 2ℓ-th moment of D_2 is

$$\frac{1}{k}(2\ell - 1)!! \cdot M(-z)_\ell.$$

So because $M(z)_\ell = M(-z)_\ell$ for all $\ell = 1, ..., k-1$, we conclude that 4) is satisfied. □

We can now exhibit a moment-matching example for mixtures of linear regressions. Let the parameters $\sigma_1, ..., \sigma_k, \sigma'_1, ..., \sigma'_k$ be as in Lemma 8.10.1.

Lemma 8.10.2. *Take any mixture of spherical linear regressions $\mathcal{D}$ in $\mathbb{R}^d$ with mixing weights $p_1, ..., p_k$ and $\Omega(1)$-separated regressors $v_1, ..., v_k \in \mathbb{R}^d$ satisfying $\|v_i\|_2 \le \mathrm{poly}(k)$ for all $i \in [k]$. Take any additional direction $v \in \mathbb{S}^{d-1}$, and any $\lambda \ge 0$. Let $\mathcal{D}_1$ (resp. $\mathcal{D}_2$) be the mixture of $3k$ linear regressions with regressors $v_1, ..., v_k, \pm\sigma_1 v, ..., \pm\sigma_k v$ (resp. regressors $v_1, ..., v_k, \pm\sigma'_1 v, ..., \pm\sigma'_k v$) and mixing weights $\frac{p_1}{Z}, ..., \frac{p_k}{Z}, \frac{\lambda/2k}{Z}, ..., \frac{\lambda/2k}{Z}$, where $Z = \lambda + 1$.*

Then $\mathcal{D}_1, \mathcal{D}_2$ satisfy the following:

1. *Both are mixtures of $\Omega(1)$-separated linear regressions whose regressors are $\mathrm{poly}(k)$-bounded in L_2 norm*

2. *They match on all moments of degree at most $2k - 1$*

3. *$d_{TV}(\mathcal{D}_1, \mathcal{D}_2) = \frac{\lambda}{\lambda+1}$*

4. *There exists a regressor w of $\mathcal{D}_1$ such that for any regressor w' of $\mathcal{D}_2$, $\|w - w'\|_2 = \Omega(\sqrt{k})$.*

Proof. 1) follows by 2) and 3) from Lemma 8.10.1. 4) follows by 1) from Lemma 8.10.1. For 3), note that the components of $\mathcal{D}_1, \mathcal{D}_2$ in direction v all have disjoint support, so $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) = \frac{\lambda}{\lambda+1}$.

It remains to check that $\mathcal{D}_1, \mathcal{D}_2$ match on moments of degree at most $2k-1$. As $\mathcal{D}_1, \mathcal{D}_2$ are identical on the components they share with $\mathcal{D}$, it suffices to show this for the mixtures $\mathcal{D}'_1, \mathcal{D}'_2$ obtained by conditioning out the components appearing in $\mathcal{D}$, that is, the two mixtures of $2k$ spherical linear regressions with uniform mixing weights and directions $\pm\sigma_1 v, ..., \pm\sigma_k v$ and directions $\pm\sigma'_1 v, ..., \pm\sigma'_k v$ respectively.

Equivalently, we must show that for any direction $(\mathbf{x}, y) \in \mathbb{R}^{d+1}$, where $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$, the univariate Gaussian mixtures D_1, D_2 obtained from projecting $\mathcal{D}'_1, \mathcal{D}'_2$ in the direction $(\mathbf{x}, y)$ have identical degree-s moment for any $s \le 2k - 1$. These moments will be zero for odd s. For $s = 2\ell$, noting that for any $\sigma \ge 0$,

$$(\mathbf{x}, y)^\top \Sigma(\sigma v)(\mathbf{x}, y) = \|x\|_2^2 + \sigma^2 y^2 + 2\sigma y \langle v, x\rangle.$$

Without loss of generality, assume $\|x\|_2 = 1$, and let $\gamma \triangleq \langle v, x\rangle$. We see that the projection

D_1 has 2ℓ-th moment

$$\frac{1}{k}(2\ell-1)!!\cdot\left[\sum_{i=1}^{k}(\sigma_i^2y^2+1+2\sigma_i\gamma y)^{\ell}+(\sigma_i^2y^2+1-2\sigma_i\gamma y)^{\ell}\right]. \tag{8.66}$$

Note that there is some degree-ℓ polynomial p for which the i-th summand in (8.66) is $p(\sigma_i^2)$. In the same way, we can see that the projection D_2 has 2ℓ-th moment $\frac{1}{k}(2\ell-1)!!\cdot\sum_{i=1}^{k}p(\sigma_i'^2)$. As the univariate mixtures in Lemma 8.10.1 match on all 2ℓ-th moments for $\ell \le k-1$, we know that $\sum_{i=1}^{k}p(\sigma_i^2)=\sum_{i=1}^{k}p(\sigma_i'^2)$ for all polynomials p of degree at most $k-1$, so the projections D_1, D_2 indeed match on all moments up to degree $2k-1$. □

8.11 Appendix: Integrating Against Fourier Transforms of Piecewise Polynomials

In this section we prove Lemma 8.5.6. Note that there are indeed explicit expressions for the Fourier moments of piecewise polynomials in terms of hypergeometric functions, but we avoid explicitly describing these for simplicity.

We will show Lemma 8.5.6 in a couple of steps. First, we show:

Lemma 8.11.1. *Let r be a nonnegative integer. Then, we have that*

$$\int_0^1 x^r\cos(ax)\mathrm{d}x=\frac{A_r(a,\sin(a),\cos(a))}{a^r}\,,\qquad \int_0^1 x^r\sin(ax)\mathrm{d}x=\frac{B_r(a,\sin(a),\cos(a))}{a^r}\,,$$

where A_r, B_r are degree-r polynomials over $\mathbb{R}^3$ whose coefficients can be computed in time $O(r^2)$.

Proof. We proceed by induction on r. The base case is trivial: if $r=0$, then

$$\int_0^1 \cos(ax)\mathrm{d}x=\frac{\sin(a)}{a}\,,$$

and

$$\int_0^1 \cos(ax)\mathrm{d}x=\frac{1-\cos(a)}{a}\,,$$

Now assume $r > 0$, and that the claim holds for $r-1$. Then by integration by parts,

$$\begin{aligned}\int_0^1 x^r \cos(ax)\mathrm{d}x &= \frac{\sin(a)}{a} - \frac{r}{a}\int \alpha_r(x)\sin(ax)\mathrm{d}x \\ &= \frac{1}{a^r}\left(a^{r-1}\sin(a) - rB_{r-1}(a, \sin(a), \cos(a))\right) \ ,\end{aligned}$$

and similarly

$$\begin{aligned}\int_0^1 x^r \sin(ax)\mathrm{d}x &= \frac{1-\cos(a)}{a} + \frac{r}{a}\int \alpha_r(x)\cos(ax)\mathrm{d}x \\ &= \frac{1}{a^r}\left(a^{r-1}(1-\cos(a)) - rA_{r-1}(a, \sin(a), \cos(a))\right) \ .\end{aligned}$$

This establishes that these are of the desired form. Moreover, this recurrence demonstrates that given the coefficients to A_{r-1}, B_{r-1}, one can obviously compute the coefficients to A_r, B_r using at most $O(r)$ additional time. This completes the proof. □

Note that we must have $\frac{A_r(a,\sin(a),\cos(a))}{a^r}$ and $\frac{B_r(\sin(a),\cos(a))}{a^r}$ converge to a finite value as $a \to 0$, as they must both converge to $\int_0^1 x^r \mathrm{d}x = r-1$. In particular, they are both analytic functions over the entire real line, if we take the convention that these functions evaluate to $r-1$ at 0, which we will. We now show:

Lemma 8.11.2. *Let $\tau > 0$, and let r, ℓ be non-negative integers. Let $\alpha_r(x) = x^r(x) \cdot \mathbf{1}_{[0,1]}(x)$. There is an algorithm that runs in time $O(r^2)$ and outputs*

$$\int_\tau^\tau \widehat{\alpha_r}[\omega] \cdot \omega^\ell \mathrm{d}\omega \ .$$

Proof. By Lemma 8.11.1, we know that there exist A_{r-1}, B_{r-1} which are degree r polynomials whose coefficients we can compute in $O(r^2)$ time so that

$$\widehat{\alpha_r}[w] = \frac{A_r(2\pi\omega, \sin(2\pi\omega), \cos(2\pi\omega))}{(2\pi\omega)^r} + \mathbf{i}\frac{B_r(2\pi\omega, \sin(2\pi\omega), \cos(2\pi\omega))}{(2\pi\omega)^r} \ .$$

Therefore we may evaluate the integral

$$\int_\tau^\tau \frac{A_r(2\pi\omega, \sin(2\pi\omega), \cos(2\pi\omega))}{(2\pi\omega)^r} \cdot \omega^\ell \mathrm{d}\omega$$

in additional $O(\ell r^2)$ time by first applying integration by parts $r - \ell$ times to remove the denominator, and then solving the trigonometric integral. and similarly we can evaluate

$$\int_{\tau}^{\tau} \frac{B_r(2\pi\omega, \sin(2\pi\omega), \cos(2\pi\omega))}{(2\pi\omega)^r} \cdot \omega^{\ell} \mathrm{d}\omega \, .$$

in time $O(r^2)$. This completes the proof. □

We now have all the tools necessary to prove Lemma 8.5.6.

Proof of Lemma 8.5.6. Note that a piecewise polynomial can be written as $\sum_{i=1}^{s} p_i(x)\mathbf{1}_{I_i}(x)$, where p_i is a degree d polynomial and $\mathbf{1}_{I_i}$ are indicator variables for intervals. By linearity of the Fourier transform, it suffices to compute the Fourier moment of $\alpha_i(x) = p_i(x)\mathbf{1}_{I_i}(x)$ for each $i = 1, \ldots, s$, and to do so, it suffices to compute $x^j\mathbf{1}_{I_j}(x)$ for every monomial $j = 0, \ldots, d$. By a change of variables, Lemma 8.11.2 gives an algorithm that runs in time $O(d^2)$ to compute the ℓ-th Fourier moment of $x^j\mathbf{1}_{I_j}(x)$. Thus we can compute the ℓ-th Fourier moment of $\alpha_i(x)$ in time $O(d^3)$, and hence of the entire piecewise polynomial in time $O(sd^3)$, as claimed. □

8.12 Appendix: Deferred Proofs

8.12.1 Proof of Lemma 8.3.1

We first require the following inequality.

Fact 8.12.1 (Rosenthal Bound, see e.g. Theorems 6.1 and 6.2 of [Pin94]). *Let $X_1, ..., X_n$ be independent random variables for which $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[|X_i|^t] < \infty$ for some $t \geq 2$. If we define $X = \frac{1}{n}\sum_{i=1}^{n} X_i$, then*

$$\mathbb{E}[|X|^t] \leq \frac{1}{n^t} \cdot \left[C_1(t) \cdot \left(\sum_{i=1}^{n} \mathbb{E}[|X_i|^t] \right) + C_2(t) \cdot \left(\sum_{i=1}^{n} \mathbb{E}[X_i^2] \right)^{t/2} \right] ,$$

where $C_1(t) = (c\gamma)^t$ and $C_2(t) = (c\sqrt{\gamma}e^{t/\gamma})^t$ for any $\gamma \in [1, t]$ and universal constant $c > 0$. In particular, we can take γ for which $\gamma \ln \gamma = 2t$ to get $C_1(t) = C_2(t) = (c' t/\ln(t))^t$ for *some other universal constant $c' > 0$.*

We can apply this to get a moment bound on the deviation of the empirical p-th moment from the true p-th moment. Define the random variable $X = \frac{1}{N}\sum_{i=1}^{N} Z_i^p - \mathbb{E}_{Z\sim\mathcal{F}}[Z^p]$, where recall that $\mathcal{F}$ is a mixture of k univariate Gaussians, and $Z_1, ..., Z_N$ are N draws from $\mathcal{F}$.

Lemma 8.12.2 (Moment bound for empirical deviation of p-th moment)**.** *There is an absolute constant $c' > 0$ for which the following holds for any p. For all $t \in \mathbb{N}$ we have that*

$$\mathbb{E}[X^t] \leq \left(\frac{c'}{\sqrt{N}} \cdot \sigma_{\max}(\mathcal{F})^p \cdot p^{p/2} \cdot t^{p/2+1}\right)^t.$$

Then for any $r, \gamma > 0$, we have that for $N = \left(\frac{\alpha}{\gamma r}\right)^2$,

$$\underset{Z_1,\cdots,Z_N}{Pr}\left[\left|\frac{1}{N}\sum_{i=1}^{N} Z_i^p - \mathbb{E}_{Z\sim\mathcal{F}}[Z^p]\right| > r\right] \leq \gamma^t.$$

Proof. For simplicity, we define

$$\sigma_{\max} = \sigma_{\max}(\mathcal{F}).$$

For every $i \in [N]$, define the random variable $X_i \triangleq Z_i^p - \mathbb{E}[Z_i^p]$. To apply Fact 8.12.1, we must compute moments of X_i. First note that for any even $d \in \mathbb{N}$ and $Z \sim \mathcal{F}$,

$$\mathbb{E}[Z^d] = \sum_{j=1}^{k} p_j \cdot \mathcal{M}_d(\mathcal{N}(0, \sigma_j^2)) \leq \sum_{j=1}^{k} p_j \cdot \sigma_j^d \cdot d^{d/2}.$$

In particular, we have that $\mathbb{E}[Z^d] \leq \sigma_{\max}^d \cdot d^{d/2}$. So for all $i \in [N]$ and even $t \in \mathbb{N}$, we get that

$$\begin{aligned}\mathbb{E}[X_i^t] &= \mathbb{E}[(Z_i^p - \mathbb{E}[Z_i^p])^t] \\ &= \sum_{\ell=0}^{t} (-1)^\ell \binom{t}{\ell} \cdot \mathbb{E}[Z_i^{p\ell}] \cdot \mathbb{E}[Z_i^p]^{t-\ell} \\ &\leq \sum_{\ell\in[t] \text{ even}} \binom{t}{\ell} \cdot (p\ell)^{p\ell/2} \sigma_{\max}^{p\ell} \cdot p^{(p/2)(t-\ell)} \sigma_{\max}^{p(t-\ell)}\end{aligned}$$

$$
\begin{aligned}
&= p^{pt/2}\sigma_{\max}^{pt} \sum_{\ell\in[t] \text{ even}} \binom{t}{\ell} \ell^{p\ell/2} \\
&\leq (2t^{p/2} \cdot p^{p/2} \cdot \sigma_{\max}^{p})^t.
\end{aligned}
$$

where the third step follows from the fact that for any degree d, $\mathbb{E}[Z_i^d] = 0$ if d is odd, and $\mathbb{E}[Z_i^d] \leq \mathbb{E}_{g\sim\mathcal{N}(0,\sigma_{\max})}[g^d] \leq d^{d/2} \cdot \sigma_{\max}^d$ if d is even; and the last step follows by naively upper bounding the terms $\ell^{p\ell/2}$ by $t^{pt/2}$.

In particular,

$$\frac{1}{N^t}\sum_{i=1}^N \mathbb{E}[X_i^t] \leq \frac{1}{N^{t-1}} \cdot \left(2t^{p/2} \cdot p^{p/2} \cdot \sigma_{\max}^p\right)^t.$$

For $\mathbb{E}[X_i^2]$, note that

$$\mathbb{E}[X_i^2] = \mathbb{E}[Z_i^{2p}] - \mathbb{E}[Z_i^p]^2 \leq \mathbb{E}[Z_i^{2p}] \leq \sigma_{\max}^{2p} \cdot (2p)^p,$$

so

$$\frac{1}{N^t}\left(\sum_{i=1}^N \mathbb{E}[X_i^2]\right)^{t/2} = \frac{1}{N^{t/2}}(\sigma_{\max}^p \cdot (2p)^{p/2})^t.$$

We conclude by Fact 8.12.1 that the random variable X satisfies the following moment bound:

$$
\begin{aligned}
\mathbb{E}[X^t] &\leq \frac{(c't/\ln(t))^t}{N^{t-1}} \cdot \left(2t^{p/2} \cdot p^{p/2} \cdot \sigma_{\max}^p\right)^t + \frac{(c't/\ln(t))^t}{N^{t/2}} \cdot \left(\sigma_{\max}^p \cdot (2p)^{p/2}\right)^t \\
&= \sigma_{\max}^{pt} \cdot p^{pt/2} \cdot \left(\frac{(c't/\ln(t))^t}{N^{t-1}} \cdot 2^t t^{pt/2} + \frac{(c't/\ln(t))^t}{N^{t/2}} \cdot 2^{pt/2}\right) \\
&\leq \sigma_{\max}^{pt} \cdot p^{pt/2} \cdot \left(\frac{(t/\ln t)^t}{N^{t/2}} \cdot t^{pt/2}\right) \cdot (c')^t \\
&\leq \left(c' \cdot \sigma_{\max}^p \cdot p^{p/2} \cdot t^{p/2+1}/\sqrt{N}\right)^t
\end{aligned}
$$

as claimed, where the third step follows from choosing $c' > 1$ to be sufficiently large constant.

□

Finally, we need some standard facts about Orlicz norms.

Definition 8.12.3 (Orlicz norms). *Let $\Psi : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ be a convex, increasing function* *satisfying* $\Psi(0) = 0$ *and* $\Psi(x) \to \infty$. *We call such a function* Ψ *a* Young function. *Let* X *be a random variable over* $\mathbb{R}_{>0}$. *The* Orlicz norm of X with respect to Ψ *is defined by*

$$\|X\|_\Psi \triangleq \inf\{c > 0 : \mathbb{E}[\Psi(X/c)] \leq 1\}.$$

Fact 8.12.4 (Sufficient condition for bound)**.** *If* $\alpha, \beta > 0$ *satisfies* $\mathbb{E}[\Psi(X/\alpha)] \leq \beta$, *then* $\|X\|_\Psi \leq \alpha \cdot \beta$.

Fact 8.12.5 (Tail bound given Orlicz norm bound)**.** *Let* X *be a random variable over* $\mathbb{R}_{>0}$. *If* $\sigma = \|X\|_\Psi < \infty$, *then*

$$Pr[X \geq \beta\|X\|_\Psi] \leq 1/\Psi(\beta)$$

Fact 8.12.6 (Approximation of e^{x^α} by Young function)**.** *For any* $0 < \alpha < 1$, *the function* Ψ_α *given by*

$$\Psi_\alpha(x) \triangleq \begin{cases} (\alpha e)^{1/\alpha} \cdot x & x < (1/\alpha)^{1/\alpha} \\ e^{x^\alpha} & x \geq (1/\alpha)^{1/\alpha} \end{cases}$$

is a Young function satisfying

$$\Psi_\alpha(x) \leq e^{x^\alpha}.$$

We can now complete the proof of Lemma 8.3.1.

Proof of Lemma 8.3.1. Take $\alpha = \frac{1}{p+2}$. Note that

$$\begin{aligned} \mathbb{E}[\Psi(X/c)] &\leq \mathbb{E}[e^{(X/c)^\alpha}] \\ &= \sum_{t=0}^{\infty} \frac{1}{t!}\mathbb{E}[(X/c)^{\alpha t}] \\ &\leq \sum_{t=0}^{\infty} \frac{c^{-\alpha t}}{t!}\mathbb{E}[X^t]^\alpha \\ &\leq \sum_{t=0}^{\infty} \frac{1}{t!} \cdot \left(c^{-1} \cdot c' \cdot \sigma_{\max}(\mathcal{F})^p \cdot p^{p/2} \cdot t^{p/2+1}/\sqrt{N}\right)^{\alpha t} \\ &= \sum_{t=0}^{\infty} \frac{1}{t!} \cdot \left(c^{-1} \cdot c' \cdot \sigma_{\max}^p \cdot p^{p/2}/\sqrt{N}\right)^{\alpha t} t^{t/2}, \end{aligned}$$

where the third step follows by Jensen's and concavity of $x \mapsto x^\alpha$ when $0 < \alpha < 1$, the fourth step follows by Lemma 8.12.2, and the last step follows by the fact that $\alpha(p/2+1) = 1/2$.

So if we take $c = c' \cdot \sigma_{\max}^p \cdot p^{p/2}/\sqrt{N}$, then $\mathbb{E}[\Psi(X/c)] = O(1)$, so we conclude by Fact 8.12.4 that

$$\|X\|_\Psi = c'' \cdot \sigma_{\max}^p \cdot p^{p/2}/\sqrt{N}$$

for some absolute constant $c'' > 0$. We can now apply Fact 8.12.5 to get that

$$\Pr\left[X \geq \beta \cdot c'' \cdot \sigma_{\max}^p \cdot p^{p/2}/\sqrt{N}\right] \leq 1/\Psi(\beta).$$

If we take $\beta = \gamma\sqrt{N}/(c'' \cdot \sigma_{\max}^p \cdot p^{p/2})$, then provided

$$N \geq (c'')^2 \cdot \gamma^{-2} \cdot (\max\{p+2, \ln(1/\delta)\})^{2p+4} \cdot p^p \cdot \sigma_{\max}^{2p},$$

we have that $\beta \geq (p+2)^{p+2}$ so that $1/\Psi(\beta) \leq \exp(\beta^{1/(p+2)})$, and $\beta \geq (\ln(1/\delta))^{p+2}$ so that $\exp(-\beta^{1/(p+2)}) \leq \delta$, so we get that

$$\Pr_{Z_1,\ldots,Z_N}\left[\left|\frac{1}{N}\sum_{i=1}^N Z_i^p - \mathbb{E}_{Z\sim\mathcal{F}}[Z^p]\right| \leq \gamma \cdot \sigma_{\max}^p \cdot p^{p/2}\right] \leq \delta.$$

Finally, we would like to relate the deviation term $\gamma \cdot \sigma_{\max}^p \cdot p^{p/2}$ to $\beta \cdot \mathbb{E}_{Z\sim\mathcal{F}}[Z^p]$. By (8.1), if we take $\gamma = \beta \cdot p_{\min}$, the lemma follows. □

8.12.2 Proof of Fact 8.2.3

Proof. Define the function $F(\sigma) \triangleq \int_{[-\tau,\tau]^c} \mathcal{N}(0,\sigma^2;x) \cdot x^p \, \mathrm{d}x$. It suffices to show that $F'(\sigma) > 0$ for all $\sigma \in (0, \sigma^*]$. We have that

$$\begin{aligned}\frac{\partial}{\partial\sigma}F(\sigma) &= \int_{[-\tau,\tau]^c} \frac{e^{-x^2/(2\sigma^2)}(x^2-\sigma^2)}{\sqrt{2\pi}\sigma^4} \cdot x^p \, \mathrm{d}x \\ &= \frac{1}{\sigma^3} \cdot \left(\int_{[-\tau,\tau]^c} \mathcal{N}(0,\sigma^2;x) \cdot (x^{p+2} - \sigma^2 x^p) \, \mathrm{d}x\right) \\ &= ((p+1)!! - (p-1)!!)\,\sigma^p - \left(\int_{[-\tau,\tau]} \mathcal{N}(0,\sigma^2;x) \cdot (x^{p+2} - \sigma^2 x^p) \, \mathrm{d}x\right).\end{aligned}$$

As the above expression tends to zero as $\tau \to \infty$, and because $\mathcal{N}(0, \sigma^2; x) \cdot (x^{p+2} - \sigma^2 x^p)$ is even, it suffices to show that the function $G(t) \triangleq \int_0^\tau \mathcal{N}(0, \sigma^2; x) \cdot (x^{p+2} - \sigma^2 x^p)\, \mathrm{d}x$ is increasing in τ. By the fundamental theorem of calculus,

$$G'(\tau) = \mathcal{N}(0, \sigma^2, \tau) \cdot (\tau^{p+2} - \sigma^2 \tau^p) > 0,$$

by the assumption that $\sigma < \sigma^* < \tau$. □

8.12.3 Proof of Corollary 1.3.19

Proof. Let $c = 1/2 - \gamma$. Note that $\langle g, w\rangle \sim \mathcal{N}(0, 1)$, so by Fact 1.3.13 we have that for sufficiently large d,

$$\Pr[\langle g, w\rangle \geq 1.1\underline{\alpha} d^c] \geq \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{2.2\underline{\alpha} d^c} \cdot e^{-1.21\underline{\alpha}^2 d^{2c}/2} \geq 2e^{-\underline{\beta} \cdot d^{2c}}$$

for $\underline{\beta} = 1.21\underline{\alpha}^2$, and

$$\Pr[\langle g, w\rangle \leq 0.9\overline{\alpha} d^c] \geq 1 - \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{0.9\overline{\alpha} d^c} \cdot e^{-0.81\overline{\alpha}^2 d^{2c}/2} \geq 1 - \frac{1}{2} \cdot e^{-\overline{\beta} \cdot d^{2c}}$$

for $\overline{\beta} = 0.81\overline{\alpha}^2/2$. On the other hand, by Fact 1.3.18, $\Pr[\|g\|_2 \in [0.9, 1.1] \cdot \sqrt{d}] \geq 1 - 2e^{-c_{\mathrm{shell}} d/100}$. By a union bound, we conclude that

$$\Pr\left[\langle v, w\rangle \geq d^{c-1/2}\right] \geq 2e^{-\underline{\beta} \cdot d^{2c}} - 2e^{-c_{\mathrm{shell}} d/100} \geq e^{-\underline{\beta} \cdot d^{2c}},$$

and similarly

$$\Pr\left[\langle v, w\rangle \leq \overline{\beta} \cdot d^{c-1/2}\right] \geq 1 - 2e^{-c_{\mathrm{shell}} d/100} - \frac{1}{2} e^{-\overline{\beta} \cdot d^{2c}} \geq 1 - e^{-\overline{\beta} \cdot d^{2c}}.$$

□

8.12.4 Proof of Corollary 1.3.20

Proof. Note that $\langle g, w_1\rangle$ and $\langle g, w_2\rangle$ are independent and distributed as $\mathcal{N}(0, 1)$.

Decompose $g \in \mathbb{R}^d$ as

$$g = \langle g, w_1\rangle w_1 + \langle g, w_2\rangle w_2 + g^{\perp},$$

where $g^{\perp} \in \mathbb{R}^d$ is a standard Gaussian vector in the subspace orthogonal to $w_1, w_2 \in \mathbb{R}^d$.

We will first lower bound the probability of the event on the left-hand side of (1.14).

By Fact 1.3.18, we have that for some absolute constant $t > 0$,

$$\Pr\left[\|g^{\perp}\|_2^2 = d \pm t\right] \geq 1/2.$$

Call this event $\mathcal{E}$. Let $\mathcal{E}'$ be the event that $\langle g, w_2\rangle \leq \sqrt{d+1} \cdot \alpha_2 d^{-1/2} = O(1)$. We know that $\Pr[\mathcal{E} \wedge \mathcal{E}'] \geq \Omega(1)$.

Conditioning on $\mathcal{E}$ and $\mathcal{E}'$, first note that $\langle v, w_2\rangle \leq \frac{1}{\sqrt{d+1}} \leq \alpha_2 \cdot d^{-1/4}$. We also have that

$$\langle v, w_1\rangle \geq \frac{\langle g, w_1\rangle}{\sqrt{\langle g, w_1\rangle^2 + 1 + d + t}},$$

so if we take $\alpha' > \alpha_1$ to be the solution to

$$\frac{\alpha' d^{1/4}}{\sqrt{\alpha'^2 \sqrt{d} + 1 + d + t}} = \alpha_1 \cdot d^{-1/4}, \tag{8.67}$$

we conclude that

$$\Pr\left[\left(\langle v, w_1\rangle \geq \alpha_1 \cdot d^{-1/4}\right) \wedge \left(\langle v, w_2\rangle \leq \alpha_2 \cdot d^{-1/4}\right)\right] \geq \Omega(1) \cdot \Pr_{h \sim \mathcal{N}(0,1)}[h \geq \alpha' d^{1/4}].$$

Furthermore, squaring both sides of (8.67) and rearranging, we see that

$$\alpha_1'^2 - \alpha_1^2 = \frac{\alpha_1^2 \alpha'^2}{\sqrt{d}} + \frac{\alpha_1^2}{d(1+t)} = O(1/\sqrt{d}),$$

so in particular $\Pr[h \geq \alpha' d^{1/4}] \geq \frac{1}{\mathrm{poly}(d)} \cdot \Pr[h \geq \alpha_1 d^{1/4}]$.

We next upper bound the probability of the event on the right-hand side of (1.14).

Write g as $g = \langle g, w_1\rangle w_1 + g'^{\perp}$ for $g'^{\perp}$ a standard Gaussian vector orthogonal to w_1. Then the event on the right-hand side of (1.14) is the event that $\langle g, w_1\rangle \geq \alpha_1 d^{-1/4}\|g\|_2$, or equivalently, that

$$\langle g, w_1\rangle \geq \frac{\alpha_1 d^{-1/4}}{\sqrt{1-\alpha_1^2 d^{-1/2}}}\|g'^{\perp}\|_2.$$

Let $\alpha'' > \alpha_1$ be the solution to

$$\frac{\alpha_1 d^{-1/4}}{\sqrt{1-\alpha_1^2 d^{-1/2}}} = \alpha'' \cdot d^{-1/4}. \tag{8.68}$$

Then the above event has probability given by the integral

$$\int_0^\infty \Pr_{h\sim\mathcal{N}(0,1)}[h \geq \alpha'' d^{-1/4}\cdot \beta^{1/2}]\cdot \mu(\beta)\,\mathrm{d}\beta, \tag{8.69}$$

where $\mu(\beta)$ is the density of the random variable $\|g'^{\perp}\|_2^2$. By Fact 1.3.13,

$$\Pr_h[h \geq \alpha'' d^{-1/4}\cdot \beta^{1/2}] \leq \frac{d^{1/4}}{\alpha''\beta^{1/2}}\cdot e^{-\frac{1}{2}\alpha''^2\beta/\sqrt{d}}.$$

For $\beta \in [0.9d, 1.1d]$, this quantity is at most $1/\mathrm{poly}(d)\cdot e^{-\frac{1}{2}\alpha''^2\beta/\sqrt{d}}$. So we may write (8.69) as

$$\begin{aligned}
&\int_{[0.9d,1.1d]} \Pr_{h\sim\mathcal{N}(0,1)}[h \geq \alpha'' d^{-1/4}\cdot \beta^{1/2}]\cdot \mu(\beta)\,\mathrm{d}\beta \\
&\quad + \int_{[0.9d,1.1d]^c} \Pr_{h\sim\mathcal{N}(0,1)}[h \geq \alpha'' d^{-1/4}\cdot \beta^{1/2}]\cdot \mu(\beta)\,\mathrm{d}\beta \\
&\leq \frac{1}{\mathrm{poly}(d)}\int_0^\infty e^{-\frac{1}{2}\alpha''^2\beta/\sqrt{d}}\cdot \mu(\beta)\,\mathrm{d}\beta \\
&\quad + \exp(-\Omega(d)) \\
&= \frac{1}{\mathrm{poly}(d)}\cdot\frac{1}{(2\pi)^{(d-1)/2}}\int e^{-\frac{g_1^2+\cdots+g_{d-1}^2}{2}\cdot\left(1+\alpha''^2/\sqrt{d}\right)}\mathrm{d}g_1\cdots\mathrm{d}g_{d-1} \\
&= \frac{1}{\mathrm{poly}(d)}\cdot(1+\alpha''^2/\sqrt{d})^{-(d-1)/2}.
\end{aligned}$$

Finally, we observe that

$$\begin{aligned}
(1+\alpha''^2/\sqrt{d})^{-(d-1)/2} &= \left((1+\alpha''^2/\sqrt{d})^{\sqrt{d}/\alpha''^2}\right)^{-\frac{d-1}{\sqrt{d}}\alpha''^2/2} \\
&\leq \left(e - O(1/\sqrt{d})\right)^{-\frac{d-1}{\sqrt{d}}\alpha''^2/2}
\end{aligned}$$

$$\leq O(e^{-\sqrt{d}\alpha''^2/2}).$$

We are done by (1.3.13) if we can show that $\Pr[h \geq \alpha'' d^{1/4}] \leq \mathrm{poly}(d)\Pr[h \geq \alpha d^{1/4}]$. But by squaring both sides of (8.68) and rearranging, we see that

$$\alpha''^2 - \alpha_1^2 = \frac{\alpha''^2\alpha_1^2}{\sqrt{d}} = O(1/\sqrt{d}),$$

so in particular $\Pr[h \geq \alpha'' d^{1/4}] \leq \mathrm{poly}(d) \cdot \Pr[h \geq \alpha d^{1/4}]$ as desired. □

8.12.5 Proof of Lemma 8.5.10

Proof. Suppose $\mathcal{D}$ has parameters $(\{p_i\}), \{w_i\})$. For the first part of the lemma, first assume that the noise rate $\varsigma = 0$ so that with probability p_i, $y = \langle w_i, x\rangle$. For entry $(j,j') \in [d]^2$, we may write

$$\begin{aligned}
2\mathbb{E}[\mathbf{M}_a^{x,y}]_{j,j'} &= \sum_{i=1}^k p_i \mathbb{E}_{x\sim\mathcal{N}(0,\mathrm{Id})}\left[\langle w_i - a, x\rangle^2 \cdot x_j x_{j'} - \mathbb{1}[j=j'] \cdot \langle w_i - a, x\rangle^2\right] \\
&= \sum_{i=1}^k p_i \mathbb{E}_{x\sim\mathcal{N}(0,\mathrm{Id})}\left[\langle w_i - a, x\rangle^2 \cdot x_j x_{j'}\right] - \mathbb{1}[j=j'] \cdot \sum_{i=1}^k p_i \|w_i - a\|_2^2 \\
&= \begin{cases} \sum_{i=1}^k p_i \left((w_i - a)_j^2 \mathbb{E}[x_j^4] + \sum_{\ell \neq j} (w_i - a)_{\ell'}^2 \mathbb{E}[x_j^2 x_\ell^2] \right) - \sum_{i=1}^k p_i \|w_i - a\|_2^2 & \text{if } j = j' \\ \sum_{i=1}^k p_i \left(2(w_i - a)_j (w_i - a)_{j'} \mathbb{E}[x_j^2 x_{j'}^2] \right) & \text{if } j \neq j' \end{cases} \\
&= \begin{cases} \sum_{i=1}^k p_i \left(3(w_i - a)_j^2 + \sum_{\ell \neq j} (w_i - a)_{\ell'}^2 \right) - \sum_{i=1}^k p_i \|w_i - a\|_2^2 & \text{if } j = j' \\ \sum_{i=1}^k p_i \left(2(w_i - a)_j (w_i - a)_{j'} \right) & \text{if } j \neq j' \end{cases} \\
&= 2\sum_{i=1}^k p_i (w_i - a)_j (w_i - a)_{j'},
\end{aligned}$$

as claimed. Now if the noise rate ς is nonzero so that with probability p_i, let y' be the random variable which equals $\langle w_i, x\rangle$ with probability p_i, so that $y = y' + g$ for $g \sim \mathcal{N}(0, \varsigma^2)$, then

$$\begin{aligned}
2\mathbb{E}[\mathbf{M}_a^{x,y}] &= \mathbb{E}\left[(y' - \langle a, x\rangle + g)^2 xx^\top - (y'+g)^2 \cdot \mathrm{Id}\right] \\
&= 2\sum_{i=1}^{k} p_i (w_i - a)(w_i - a)^\top + \mathbb{E}[g^2 \cdot xx^\top] - \mathbb{E}[g^2]\cdot \mathrm{Id} \\
&= 2\sum_{i=1}^{k} p_i (w_i - a)(w_i - a)^\top,
\end{aligned}$$

where the second step follow by the fact that g is independent of the random variables x, y'.

The second part of the lemma follows from the following fact, which quantifies the extent to which the matrix $\mathbf{M}_a^{x,y}$ concentrates in spectral norm. This is already proven in the noiseless case, see e.g. Eq. (34) in [YCS16], and the noisy version follows from a straightforward modification of that proof using Theorem 4.7.1 in [Ver18].

Fact 8.12.7 (Concentration of Empirical Moments)**.**

$$Pr\left[\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_a^{x_i,y_i} - \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{M}_a^{x,y}\right]\|_2 \geq \Omega\left(\max_{i\in[k]}\|w_i - a\|_2^2 \cdot \frac{\ln\left(p_{\min}N\right)}{\sqrt{p_{\min}N}} \cdot \sqrt{d\ln(k/\delta)}\right)\right] \leq \delta.$$

□

8.12.6 Proof of Lemma 8.5.17

Proof. We bound the sample complexity and runtime of each of the $O(MT)$ iterations. In each iteration, we first sample N_1 points, and perform an approximate k-SVD on a $N_1 \times d$ matrix, where N_1 is defined as in Line 14. By Corollary 8.5.8, $\underline{\sigma}_t^{\mathrm{sharp}}$ is at most a constant factor smaller than $\underline{\sigma} = \Omega(\varepsilon)$. Therefore the sample complexity of this step is at most

$$N_1 = \widetilde{O}(\varepsilon^{-2} p_{\min}^{-2} dk^2 \ln(1/\delta))\ .$$

and the runtime is at most $\widetilde{O}(N_1 kd)$ by Lemma 8.5.11. The other contribution to the sample complexity and runtime of each iteration (at least in most regimes) is from CompareMinVariances. By our choice of parameters and Corollary 8.5.9, the sample complexity of

CompareMinVariances is

$$N = p_{\min}^{-4} k \ln(1/\delta) \cdot \mathrm{poly}\left(\sqrt{k}, \ln(1/p_{\min}), \ln(1/\varepsilon)\right)^{O\left(\sqrt{k}\ln(1/p_{\min})\right)} .$$

and the runtime is bounded by $\widetilde{O}(N)$. Since we run for $MT = \widetilde{O}\left(\sqrt{k}e^{\sqrt{k}}\ln(1/\varepsilon)\right)$ iterations, this completes the proof.

□

8.12.7 Proof of Lemma 8.8.12

Proof. Prior to the outer loop, we first sample N_1 points, and perform an approximate k-SVD on a $N_1 \times d$ matrix, where N_1 is defined as in Line 14.

Therefore the sample complexity of this step is at most

$$N_1 = \widetilde{O}\left(d \cdot \mathrm{poly}(k)/p_{\min}^2\right) .$$

and the runtime is at most $\widetilde{O}(N_1 kd)$ by Lemma 8.5.11.

The bulk of the contribution to the sample complexity and runtime comes from the S iterations of the outer loop, each of which consists of MT iterations of the inner loop (over t) and a call to CheckOutcomeHyperplanes. The complexity of these MT iterations is dominated by an invocation of CompareMinVariances. By our choice of parameters and Corollary 8.5.9, the sample complexity of one run of CompareMinVariances is

$$N = p_{\min}^{-4} k \cdot \mathrm{poly}\left(k^{3/5}, \ln(1/p_{\min}), \ln(1/\varepsilon)\right)^{O\left(k^{3/5}\ln(1/p_{\min})\right)}$$

and the runtime is bounded by $\widetilde{O}(N)$. Each iteration $i \in [S]$ involves $M \cdot T = \widetilde{O}\left(k^{3/5}e^{k^{3/5}}\ln(1/\varepsilon)\right)$ such iterations. Additionally, the i-th iteration runs CheckOutcomeHyperplanes, a run of which has time and sample complexity

$$N_2 = O\left(p_{\min}^{-2} \cdot \ln(2S/\delta)\right) = O\left(p_{\min}^{-2} \cdot k^{3/5}\ln(2/\delta)\right) .$$

We conclude that HyperplaneMomentDescent requires sample complexity

$$\widetilde{O}\left(N_1 + S \cdot \left(k^{3/5} e^{k^{3/5}} N + N_2\right)\right)$$

and runs in time

$$\widetilde{O}\left(dN_1 + S \cdot \left(k^{3/5} e^{k^{3/5}} N + N_2\right)\right).$$

□

Part IV

Data Science and the Sciences

Chapter 9

Mixture Models and the Diffraction Limit

9.1 Introduction

The final mixture model that we study in this thesis is inspired by the following classic question from optics. For more than a century and a half it has been widely believed (but was never rigorously shown) that the physics of diffraction imposes certain fundamental limits on the resolution of an optical system. In the standard physical setup, we observe incoherent illumination from far-away point sources through a perfectly circular aperture (see Figure 9-1). Each point source produces a two-dimensional image, computed explicitly by Sir George Biddell Airy in 1835 [Air35] and now called an *Airy disk*. For a point source of light whose angular displacement from the optical axis is $\mu \in \mathbb{R}^2$, recall from our discussion in Section 1.2.3 that the normalized intensity at a point $\mathbf{x}$ on the observation plane is given by

$$I(\mathbf{x}) = \frac{1}{\pi\sigma^2}\left(\frac{2J_1(\|\mathbf{x}-\mu\|_2/\sigma)}{\|\mathbf{x}-\mu\|_2/\sigma}\right)^2$$

where J_1 is a Bessel function of the first kind. Under Feynman's path integral formalism, $I(x)$ is precisely the pdf of the distribution over where the photon is detected (see Appendix 9.7). The physical properties of the optical system, namely its numerical aperture and the wavelength of light being observed, determine σ which governs the amount by which each point

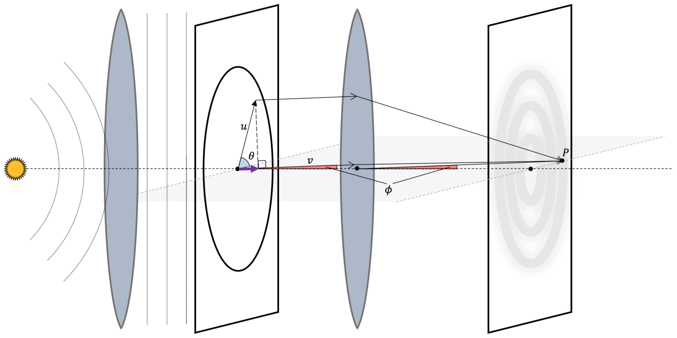


Figure 9-1: Fraunhofer diffraction of incoherent illumination from point source through aperture onto observation plane

source gets blurred.

Intuitively, when point sources are closer together it seems harder to resolve them. However, despite considerable interest over the years [Abb73,Ray79,Sch04,Spa16,Hou27,Bux37], our understanding of what exactly can and cannot be resolved has never risen above heuristic arguments. In 1879 Lord Rayleigh [Ray79] proposed a criterion for assessing the resolving power of an optical system, which is still widely-used today, of which he wrote:

> *"This rule is convenient on account of its simplicity and it is sufficiently accurate in view of the necessary uncertainty as to what exactly is meant by resolution."*

Over the years, many researchers have proposed alternative criteria and offered arguments about why some are more appropriate than others. For example, in 1916 Carroll Sparrow proposed a new criterion [Spa16] that bears his name, which he justified as follows:

> *"It is obvious that the undulation condition should set an upper limit to the resolving power . . . The effect is observable both in positives and in negatives, as well as by direct vision . . . My own observations on this point have been checked by a number of my friends and colleagues."*

Even more resolution criteria were proposed, both before and after, by Ernst Abbe [Abb73], Sir Arthur Schuster [Sch04], William Houston [Hou27], etc. Their popularity varies depending on the application area and research community. Many researchers have also pushed back on the idea that there is a fundamental diffraction limit at all. In his 1964 Lectures on Physics [FLS11, Section 30-4], Richard Feynman writes:

> *"... it seems a little pedantic to put such precision into the resolving power formula. This is because Rayleigh's criterion is a rough idea in the first place. It tells you where it begins to get very hard to tell whether the image was made by one or by two stars. Actually, if sufficiently careful measurements of the exact intensity distribution over the diffracted image spot can be made, the fact that two sources make the spot can be proved even if θ is less than λ/L."*

Or as Toraldo di Francia [DF55] puts it:

> *"Mathematics cannot set any lower limit for the distance of two resolvable points."*

Our goal in this chapter is to remedy this gap in the literature and place the notion of the diffraction limit on rigorous statistical foundations by drawing new connections to recent work in theoretical computer science on provably learning mixture models, as we will describe next. First we remark that the way the diffraction limit is traditionally studied is in fact a mixture model. In particular we assume that, experimentally, we can measure photons that are sampled from the true diffracted image. However we only observe a finite number of them because our experiment has finite exposure time, and indeed as we will see in some settings the number of samples needed to resolve closely-spaced objects can explode and be essentially impossible just from statistical considerations. Moreover we may only be able to record the location of observed photons up to some finite accuracy, which can also be thought of as being related to sampling error. The main question we will be interested in is:

> *How many samples (i.e. photons) are needed to accurately estimate the centers and relative intensities of a mixture (i.e. superposition) of two or more Airy disks, as a function of their separation and the parameters of the optical system?*

This is a central question in optics. Fortunately, there are many parallels between this question and the problem of provably learning mixture models that surprisingly seem to have gone undiscovered. In particular, let us revisit Sparrow's argument that resolution is impossible when the density function becomes unimodal. In fact there are already counterexamples to this claim, albeit not for mixtures of Airy disks. It is known that there are algorithms for learning mixtures of two Gaussians that take a polynomial number of samples and run in polynomial time. These algorithms work even when the density function is unimodal, and require just that the overlap between the components can be bounded away from one. Moreover when there are k components it is known that there is a critical separation above which it is possible to learn the parameters accurately with a polynomial number of samples, and below which accurate learning requires a superpolynomial number of samples information-theoretically [RV17]. Thus a natural way to formulate what the diffraction limit is, so that it can be studied rigorously, is to ask:

> *At what critical separation does the sample complexity of learning mixtures of k Airy disks go from polynomial to exponential?*

In this chapter we will give algorithms whose running time and sample complexity are polynomial in k above some critical separation, and prove that below some other critical separation the sample complexity is necessarily exponential in k. These bounds will be within a universal constant, and thus we approximately locate the true diffraction limit. There will also be some surprises along the way, such as the fact that the *Abbe limit*, which has long been postulated to be the true diffraction limit, is not actually the correct answer!

Before we proceed, we also want to emphasize that there is an important conceptual message in our work. First, for mixtures of Gaussians the model was only ever supposed to be an *approximation* to the true data generating process. For example, Karl Pearson introduced mixtures of Gaussians in order to model various physical measurements of the Naples crabs. However mixtures of Gaussians always have some chance of producing samples with negative values, but Naples crabs certainly do not have negative forehead lengths! In contrast, for mixtures of Airy disks the model is an extremely accurate approximation to the observations in many experimental setups *because it comes from first principles.* It is particularly accurate in astronomy where for all intents and purposes the lens is spherical and the star is so far away that it is a point source, and the question itself is highly relevant because it arises when we want to locate double-stars [Fal67] .

Furthermore we believe that there ought to be many more examples of inverse problems in science and engineering where tools and ideas from the literature on provably learning mixture models ought to be useful. Indeed both mixtures of Gaussians and mixtures of Airy disks can be thought of as inverse problems with respect to simple differential equations, for the heat equation and a modified Bessel equation respectively. While this is a well-studied topic in applied mathematics, usually one makes some sort of smoothness assumption on the initial data. What is crucial to both the literature on learning mixtures of Gaussians and our work is that we have a parametric assumption that there are few components. Thus we ask: Are there provable algorithms for other inverse problems, coming from differential equations, under parametric assumptions? Even better: Could techniques inspired by the method of moments play a key role in such a development?

9.1.1 Overview of Results

It is often the case that heuristic arguments, despite being quite far from a rigorous proof, predict the correct thresholds for a wide range of statistical problems. However here there will be a surprise. In a seminal work in 1873, Ernest Abbe formulated what is now called the *Abble limit*. Since then it has been widely accepted in the optics literature as the critical distance below which diffraction makes resolution impossible for classical optical systems. In the mixture model formalism outlined above, it corresponds to a separation of $\pi\sigma$ between any pair of Airy disk centers μ_i, μ_j. This distance arises naturally because it corresponds to the radius of the support of the Fourier transform of the Airy disk kernel $A_\sigma : \mathbf{x} \mapsto \frac{1}{\pi\sigma^2}\left(\frac{J_1(\mathbf{x}/\sigma)}{\mathbf{x}/\sigma}\right)^2$ (see Appendix 9.8.4 for further discussion).

One of the main results of our work in this chapter is to show that resolution is statistically hard even above the Abbe limit! Specifically, we show that even for mixtures of Airy disks whose centers have a pairwise separation that is a constant factor larger than the Abbe limit, the problem of recovering their locations can require $\exp(\Omega(\sqrt{k}))$ samples. The main challenge is that no configuration where the Airy disk centers are all on the same line can beat the Abbe limit. Instead we construct a new, natural lower bound instance.

Theorem 9.1.1 (Informal, see Theorem 9.5.1)**.** *Let* $\underline{\gamma} \triangleq \sqrt{4/3} \approx 1.155$. *For any* $0 < \varepsilon < 1$, *there exist two superpositions of* k *Airy disks* ρ, ρ' *which are both* $\underline{\gamma} \cdot (1-\varepsilon) \cdot \pi\sigma$*-separated and such that 1)* ρ *and* ρ' *have noticably different sets of centers, and yet 2) it would take at* *least* $\exp(\Omega(\varepsilon\sqrt{k}))$ *samples to distinguish whether the samples came from* ρ *or from* ρ'.

On the other hand, we also show that when the Airy disks have separation that is a small constant factor larger than this critical distance, there is an algorithm for recovering the centers that takes a polynomial number of samples and runs in polynomial time.

Theorem 9.1.2 (Informal, see Theorem 9.4.2)**.** *Define the absolute constant* $\overline{\gamma} = \frac{2j_{0,1}}{\pi} = 1.530\ldots$, *where* $j_{0,1}$ *is the first positive zero of the Bessel function* J_0. *Let* ρ *be a* $\overline{\gamma} \cdot \pi\sigma$*-separated superposition of* k *Airy disks where every disk has relative intensity at least* λ. *Then for any target error* $\varepsilon > 0$, *there is an algorithm with time and sample complexity* $N = \mathrm{poly}\,(k, 1/\Delta, 1/\lambda, 1/\varepsilon)$ *which outputs an estimate for the centers and relative intensities of* ρ *which incurs error* ε *with probability at least* $9/10$. *Furthermore, this holds even when there is granularity in the photon detector, as long as it is at most some inverse polynomial in all parameters.*

The main open question of our work is to prove matching upper and lower bounds that pin down the true diffraction limit. However, as we will discuss, this is a challenging problem in harmonic analysis, despite being connected to areas where there has been considerable recent progress. Moreover this phase transition for resolution is actually more dramatic than what happens for mixtures of Gaussians [RV17]. Even ignoring the issue of computational complexity, for spherical Gaussian mixtures it is known that at separation $o(\sqrt{\log k})$, super-polynomially many samples are needed, while at separation $\Omega(\sqrt{\log k})$, polynomially many suffice.

We now say a word about the techniques that go into proving Theorem 9.1.1 and Theorem 9.1.2. It turns out that both are closely related to proving a modified version of an *Ingham-type estimate* [KL05]:

Q6. *What is the smallest* Δ *for which the quantity*

$$\int_B \left| \sum_{j=1}^{k} \lambda_j e^{-2\pi i \langle \mu_j, \omega \rangle} \right|^2 d\omega \geq \frac{1}{\mathrm{poly}(k)} \|\lambda\|_2^2$$

for all vectors $\lambda \in \mathbb{R}^k$ *and all sets of centers* $\{\mu_j\}$ *for which* $\|\mu_i - \mu_j\|_2 > \Delta$ *for all* $j \neq j'$, *where the integration is over the origin-centered unit ball* $B \subset \mathbb{R}^2$*?*

In particular, the main technical step for showing Theorem 9.1.2 is to show that the critical Δ in Question 6 is at most $2j_{0,1}/\pi$. This can be obtained via the following extremal function. A *ball minorant*, is a function F satisfying the properties that

(1) $F(x) \le \mathbb{1}[x \in B]$ and

(2) $\widehat{F}$ is supported on the ball of radius Δ

In [HV$^+$96, CCLM17, Gon18] it was shown that such a ball minorant exists for $\Delta = 2j_{0,1}/\pi$ (interestingly, this paved the way to some recent progress on Montgomery's famous pair correlation conjecture for the Riemann zeta function [CCLM17]). One can use property (1) to pass from integrating against the function $\mathbb{1}[x \in B]$ to integrating against F. And because by property (2) F is localized in the frequency domain, the latter integral is large. In fact the one-dimensional analogue of Question 6 was resolved in [Moi15] using the univariate analogue of F, namely the *Beurling-Selberg minorant*. However the algorithmic approach only made sense in one-dimension. In our case, we employ the tensor generalization of the matrix pencil method, originally introduced in [HK15]. We defer the details of this to Section 9.4.3.

For the lower bound in Theorem 9.1.1, we need to answer a variant of Question 6.

Q7. *What is the smallest Δ for which*

$$\int_B \left| \sum_{j=1}^k \lambda_j e^{-2\pi i \langle \mu_j, \omega \rangle} - \sum_{j=1}^k \lambda'_j e^{-2\pi i \langle \mu'_j, \omega \rangle} \right|^2 d\omega \ge \frac{1}{\mathrm{poly}(k)} \tag{9.1}$$

for all nonnegative $\lambda, \lambda' \in \mathbb{R}^k$ *whose entries sum to one and* $\{\mu_j\}, \{\mu'_j\} \in \mathbb{R}^2$ *for which* $\|\mu_j - \mu_{j'}\|_2 > \Delta$ *and* $\|\mu'_j - \mu'_{j'}\|_2 > \Delta$ *for* $j \neq j'$, *where integration is over the origin-centered unit ball* B*?*

The connection to Theorem 9.1.1 is straightforward: By Plancherel's and smoothness properties of A_σ, one can upper bound the L_1 distance between the mixture of Airy disks given by parameters $\{\lambda_j\}, \{\mu_j\}$ and the mixture given by $\{\lambda'_j\}, \{\mu'_j\}$ in terms of the left-hand side of (9.1). So if one can construct a set of Δ-separated centers $\{\mu_j\}, \{\mu'_j\}$ for which (9.1) fails to hold but for which the collection $\{\mu_j\}$ is separated from $\{\mu'_j\}$ but the resulting mixtures of Airy disks are $o(1/\mathrm{poly}(k))$-close in total variation distance it implies that resolution is statistically impossible with a polynomial number of samples. This is the recipe used in known lower bounds [MV10,HP15,RV17] for learning mixtures of Gaussians.

For our purposes, it turns out that "tensoring" one-dimensional lower bounds does not work because it would not beat the Abbe limit [Moi15]. Morally, this is because tensoring the unit interval with itself would give us the unit square, which corresponds to separation in the L_∞ distance rather than the L_2 distance, and the L_2 distance is the right distance in optics because it is rotationally invariant. The main technical contribution in our lower bound is to give a more sophisticated construction given by interleaving two triangular lattices and placing the centers at points on these lattices (see Figure 9-5). The analysis is rather delicate, and we defer the details to Section 9.2 and Section 9.5.

To complete the picture, we show that there is no diffraction limit when the number of Airy disks is a constant. In particular we show that for any constant number of Airy disks there is an algorithm that takes a polynomial number of samples and runs in polynomial time that learns the parameters to any desired accuracy *regardless of the separation.*

Theorem 9.1.3 (Informal, see Theorem 9.4.1)**.** *Let ρ be a Δ-separated superposition of k Airy disks where every disk has relative intensity at least λ. Then for any target error $\varepsilon > 0$ and failure probability $\delta > 0$, there is an algorithm which draws $N = \mathrm{poly}\left((k\sigma/\Delta)^{k^2}, 1/\lambda, 1/\varepsilon, \log(1/\delta)\right)$ samples from ρ, runs in time $O(N)$, and outputs an estimate for the centers and relative intensities of ρ which incurs error ε with probability at least $1-\delta$. Furthermore, this holds even when there is granularity in the photon detector, as long as it is at most some inverse polynomial in N.*

This result turns out to be simple in retrospect, and comes from assembling a few standard tools from the literature on provably learning mixture models. Nevertheless it underscores an important point that existing tools can already have important implications for inverse problems the sciences. Our approach is to first estimate the Fourier transform $\widehat{\rho}$ from samples and then pointwise divide by $\widehat{A_\sigma}$. In this way we can simulate noisy access to the Fourier transform of the mixture of delta functions at $\mu_1, \ldots, \mu_k$. However $\widehat{A_\sigma}$ has compact support, so we can only access frequencies with bounded L_2 norm. Now we can reduce to the one-dimensional case [Moi15] by projecting ρ along two nearby directions, solving each resulting univariate problem, and then solving an appropriate linear system to recover the centers and relative intensities. This method is reminiscent of [KMV10, MV10], which gives algorithms for learning high-dimensional mixtures of Gaussians based on reducing to a series of one-dimensional problems and stitching together these estimates carefully.

9.1.2 Related Work

We have already mentioned that our work is closely related to the vast literature on learning mixture models and, in particular, on learning mixtures of Gaussians [Das99, DS00, AK01, VW02, AM05, BV08, KMV10, MV10, BS15, HP15, HK13, GHK15, RV17, HL18, KS17, DKS18b]. Here we mention some other connections to work on recovering spike trains from noisy, band-limited Fourier measurements.

Superresolution The seminal work of [DS89, Don92] was one of the first to put this question on rigorous footing. Donoho studied the modulus of continuity for this problem on a grid as the grid width goes to zero. Later Candes and Fernandez-Granda [CFG14] gave a practical algorithm based on L_1 minimization over a continuous domain. There has been a long line of work on this problem which it would also be impossible to survey fully, so we refer the reader to [CFG13, TBSR13, FG13, Lia15, Moi15, FG16, KPRvdO16, MC16] and references therein.

We remark that essentially all works on super-resolution in high dimensions focus on the case where measurements are L_∞-band-limited rather than L_2-band-limited. Given the prevalence of Airy disks and circular apertures in statistical optics, one upshot of our work is that, technical issues related to the so-called box (aka L_∞ ball) minorant problem notwithstanding, the L_2 setting may be the more practically relevant one to consider anyways.

Sparse Fourier Transform There are also connections to the extensive literature on the sparse Fourier transform, which can be interpreted in some sense as the "agnostic" version of the super-resolution problem where the goal is to compete with the error of the best k-sparse approximation to the discrete Fourier transform, even in the presence of noise, using few measurements [GGI$^+$02, GMS05, HIKP12, GIIS14, IKP14, Kap16]. When the k spikes need not be at discrete locations and the low-frequency measurements are randomly chosen, this

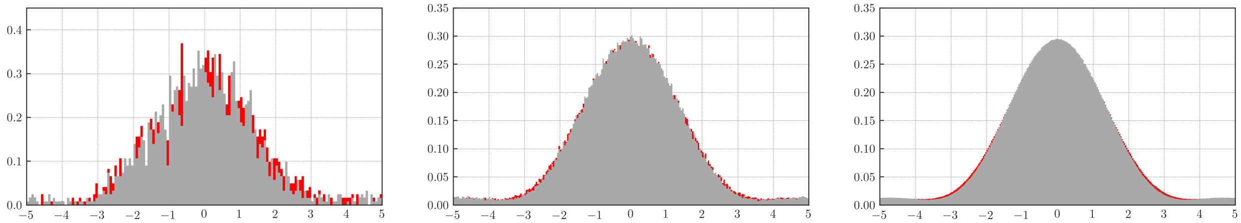


Figure 9-2: With enough samples, one can distinguish which of two superpositions the data comes from, even below the diffraction limit: In each plot, a histogram of x-axis positions of photons sampled from a superposition of two equal-intensity Airy disks (red) centered on the x-axis with separation a tenth of the Abbe limit is overlaid with a histogram of x-axis positions of photons sampled from a single Airy disk at the origin (gray). As number of samples increases (left to right), minute differences between the two intensity profiles become clear.

is the problem of *compressed sensing off the grid* introduced by [TBSR13], for which recovery is possible with far fewer measurements. This can be thought of as the one-dimensional case of the setting of [HK15]. To our knowledge, the only work that addresses the continuous, high-dimensional version of the sparse Fourier transform is the very recent work of [JLS20]. The emphasis in this literature is primarily on obtaining sample complexity near-linear in k, whereas our guarantees are only polynomial in k. Consequently the results in the sparse Fourier transform literature lose log factors in the level of separation they require, whereas in our setting the emphasis is primarily on the level of separation needed to get polynomial-time and -sample algorithms.

9.1.3 Visualizing the Diffraction Limit

In this short section we provide some figures to help conceptualize our results. Figure 9-2 illustrates the basic notion that separation is information-theoretically unnecessary for parameter learning of superpositions of Airy disks. We compare the discretized empirical distribution of samples from two diffraction patterns whose components have separation well below the diffraction limit and thus well below what conventional wisdom in optics suggests is resolvable. While the differences in the diffraction patterns are minute, they do indeed become statistically significant with enough samples. Eventually it becomes possible to conclude that the gray diffraction pattern is generated by one point source and the red diffraction pattern is generated by two.

Next, we present a striking visual representation of the statistical barrier imposed by

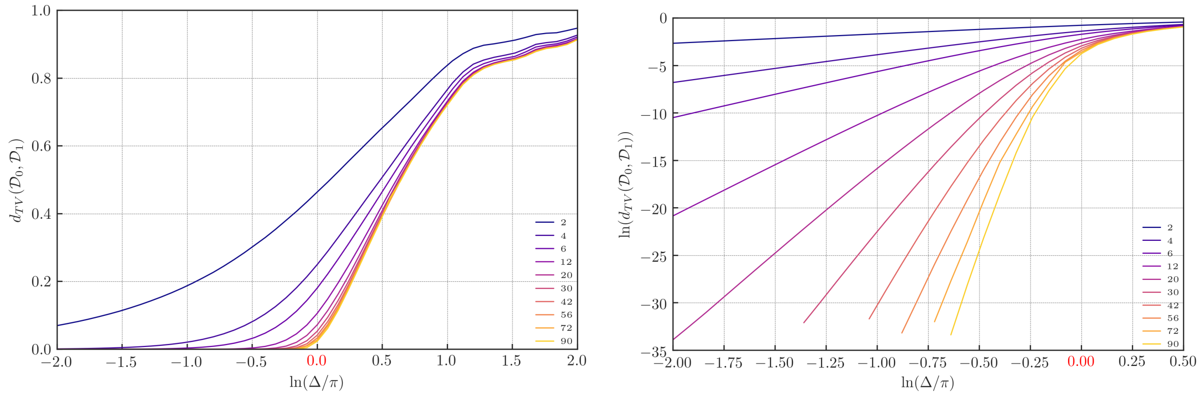


Figure 9-3: The Abbe limit as a statistical phase transition: For any level of separation Δ and number of disks k, we carefully construct a pair of hypotheses $\mathcal{D}_0(\Delta, k), \mathcal{D}_1(\Delta, k)$ which are each superpositions of $k/2$ Airy disks where the separation among its components is at least Δ. The left figure plots total variation distance $d_{\mathrm{TV}}(\mathcal{D}_0(\Delta, k), \mathcal{D}_1(\Delta, k))$ between the two distributions as a function of Δ, for various choices of k, with the Abbe limit highlighted in red. The right figure plots total variation distance on a log-scale.

the diffraction limit when the number of components is large. Recall that the upshot of Theorems 9.1.1 and 9.1.2 is that k plays a leading role in determining when resolution is and is not feasible: slightly above the Abbe limit, the sample (and computational) complexity is polynomial in k, and anywhere beneath the Abbe limit, the sample complexity becomes exponential in k. This helps clarify why in some domains like astronomy, where there are only ever a few tightly spaced point sources, there is evidently no diffraction limit. Yet in other domains like microscopy where there are a large number of tightly spaced objects, the diffraction limit is indeed a fundamental barrier, at least in the classical physical setup. This helps explain why different communities have settled on different beliefs about whether there is or is not a diffraction limit.

In Figure 9-3 we experimentally investigate this phenomenon and illustrate how the total variation distance scales as we vary the number of disks and the separation in our earlier constructions. It is evident from these plots that for any superposition of a few Airy disks, there is no sharp dividing line between what is and is not possible to resolve. But when the number of Airy disks becomes large, with any reasonable number of samples, it is feasible to resolve the superposition if and only if their separation is at least as large as the Abbe limit.

We emphasize that in the instance constructed for Figure 9-3 (as well as the instance we construct and analyze for Theorem 9.1.1), the centers are plotted on a line. For such instances, by projecting in the direction of the line and using our deconvolution techniques, one can actually reduce to the problem of one-dimensional super-resolution, for which polynomial-time algorithms exist for *any* separation strictly greater than the diffraction limit [Moi15], and by adapting the lower bound in [Moi15] to this specific instance, one can see that this is tight. In contrast, if the centers can be placed anywhere in $\mathbb{R}^2$, there is a constant factor gap ($\sqrt{4/3}$ versus $\frac{2j_{0,1}}{\pi}$) between the lower bound in Theorem 9.1.1 and the upper bound in Theorem 9.1.2.

9.1.4 Roadmap

In Section 9.2 we give a preview of our lower bound proof by providing a self-contained answer to Question 7. In Section 9.3, we give an overview of our probabilistic model, some notation, and other mathematical preliminaries. In Section 9.4, we prove the algorithmic results in Theorems 9.1.3 and 9.1.2. In Section 9.5 we complete the proof of our lower bound from Theorem 9.1.1. In Section 9.6 we conclude with some directions for future work. In Appendix 9.7, we overview previous attempts in the optics literature to put the diffraction limit on rigorous footing. In Appendix 9.8, we describe and motivate our model and also define the various resolution criteria which have appeared in the literature. In Appendix 9.9, we catalogue quotations from the literature that are representative of the points of view addressed in the introduction. In Appendix 9.10, we complete some deferred proofs. Lastly, in Appendix 9.11, we give details on how Figure 9-3 was generated.

9.2 Lower Bound Preview

In this section we give a self-contained proof of one of the main technical ingredients in the proof of our main result, Theorem 9.1.1. Before proceeding, it will be convenient to introduce a bit of notation; any outstanding notation we will present Section 9.3, e.g. our convention for the Fourier transform. Recalling that $\underline{\gamma} \triangleq \sqrt{4/3}$, define

$$m \triangleq \frac{2}{(1-\varepsilon)\underline{\gamma}\pi\sigma} \tag{9.2}$$

for any small constant $\varepsilon > 0$ so that the critical level of separation for which Theorem 9.1.1 applies is $\Delta \triangleq 2/m = \underline{\gamma} \cdot (1-\varepsilon) \cdot \pi\sigma$.[1] Additionally, let k be an odd square and define

$$\nu_{j_1,j_2} = \frac{\Delta}{2} \cdot (j_1, \sqrt{3} \cdot j_2), \quad j_1, j_2 \in \mathcal{J} \triangleq \left[-\frac{\sqrt{k}-1}{2}, \dots, \frac{\sqrt{k}-1}{2} \right].$$

This construction is illustrated in Figure 9-5: there, similarly colored points correspond to centers in the same mixture, and our choice of $\{\nu_{j_1,j_2}\}$ ensures that the level of separation between any two points in a particular mixture is Δ, which is slightly less than $\sqrt{4/3}$ times the Abbe limit of $\pi\sigma$. As such, the following tells us that the answer to Question 7 is surprisingly at least $\sqrt{4/3}$, rather than 1 as the Abbe limit would suggest:

Lemma 9.2.1. *There exists a vector $u \triangleq (u_{j_1,j_2})_{j_1,j_2 \in \mathcal{J}} \in \mathbb{R}^k$ for which*

$$\left| \sum_{j_1,j_2 \in \mathcal{J}} u_{j_1,j_2} e^{-2\pi i \langle \nu_{j_1,j_2}, \mathbf{x} \rangle} \right|^2 \le \exp\left(-\Omega(\varepsilon\sqrt{k}) \right)$$

for all $\|x\| \le 1/\pi\sigma$. Furthermore, $\mathrm{sgn}(u_{j_1,j_2}) = (-1)^{j_1+j_2}$, *and*

$$\sum_{j_1+j_2 \ \textit{even}} |u_{j_1,j_2}| = \sum_{j_1+j_2 \ \textit{odd}} |u_{j_1,j_2}| = 1. \tag{9.3}$$

We need the following ingredient from the proof of the one-dimensional lower bound in [Moi15].

Definition 9.2.2. *The Fejer kernel is given by*

$$K_\ell(x) = \frac{1}{\ell^2} \sum_{j=-\ell}^{\ell} (\ell - |j|) e(jx) = \frac{1}{\ell^2} \left(\frac{\sin \ell\pi x}{\sin \pi x} \right)^2. \tag{9.4}$$

We will denote the r-th power of $K_\ell(\cdot)$ by $K_\ell^r(\cdot)$.

Fact 9.2.3. *K_ℓ is even and periodic with period 1. For $x \in [-1/2, 1/2]$, $K_\ell(x) \le \frac{1}{4\ell^2 x^2}$.*

Proof. That K_ℓ is even and periodic follow from the second definition of K_ℓ in (9.4). For

[1] This $\pi\sigma$ scaling is not important to the result in this section but is the natural choice of scaling for Airy disks, so it will be convenient to work with this when we apply the results of this section to prove Theorem 9.1.1.

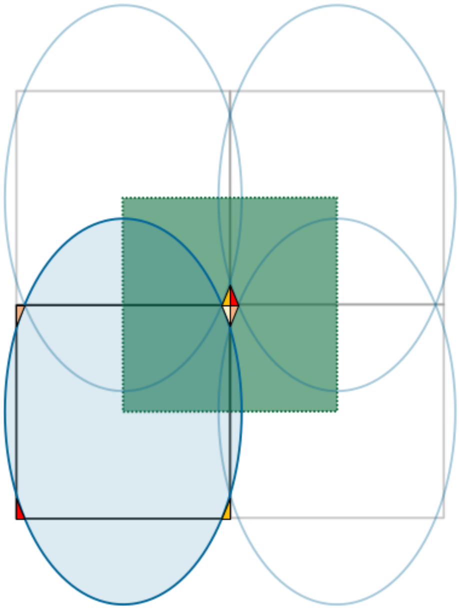


Figure 9-4: The squares correspond to periods of K_ℓ^r, while the ellipses have major and minor axes of length $\underline{\gamma}(1-\varepsilon)$ and $2(1-\varepsilon)$. The figure is centered around the origin, and the bottom-left ellipse K is the set of points $\left(\frac{x_1}{m}-\frac{1}{2}, \frac{x_2\sqrt{3}}{m}-\frac{1}{2}\right)$ as (x_1, x_2) ranges over the origin-centered L_2 ball of norm $1/\pi\sigma$. By appropriately translating the four quadrants of this ellipse by distances in $\mathbf{Z}^2$, we obtain overlapping regions whose union is given by $R\backslash S$, where $R = [-1/2, 1/2] \times [-1/2, 1/2]$ is given by the central square (green) and S is the multi-colored set in the middle given by tranlates of the four connected components of $([-1, 0] \times [-1, 0])\backslash K$.

the bound on K_ℓ, we can use the elementary bounds $\sin \pi x \geq 2x$ for $x \in [0, 1/2]$ and $(\sin \ell\pi x)^2 \leq 1$. □

Proof of Lemma 9.2.1. Let $\ell = 4/\varepsilon$ and $r = (\sqrt{k}-1)/2\ell = \Theta(\varepsilon\sqrt{k})$, and assume without loss of generality that ℓ is even. Consider the function $H : \mathbb{R}^2 \to \mathbb{R}$ given by

$$H(x_1, x_2) = K_\ell^r\left(\frac{x_1}{m} - \frac{1}{2}\right) \cdot K_\ell^r\left(\frac{x_2\sqrt{3}}{m} - \frac{1}{2}\right).$$

We know that $\widehat{K}_\ell[t] = \frac{1}{\ell^2}\sum_{j=-\ell}^{\ell}(\ell - |j|)\delta(t-j)$, so $\widehat{K_\ell^r}[t] = \sum_{j\in\mathcal{J}} \alpha_j \delta(t-j)$ for nonnegative α_j which sum to 1. Assuming without loss of generality suppose that m defined by (9.2) is an odd integer, we conclude that for $\mathbf{t} = (t_1, t_2) \in \mathbb{C}^2$,

$$\widehat{H}[\mathbf{t}] = \sum_{j_1, j_2 \in \mathcal{J}} h_{j_1,j_2} e^{-\pi i m(t_1 + t_2/\sqrt{3})}\delta(mt_1 - j_1) \cdot \delta(mt_2/\sqrt{3} - j_2)$$

$$= \sum_{j_1,j_2\in\mathcal{J}} h_{j_1,j_2}(-1)^{j_1+j_2}\delta(\mathbf{t}-\nu_{j_1,j_2}),$$

where $h_{j_1,j_2} = \alpha'_{j_1}\alpha'_{j_2} \geq 0$, where $\alpha'_j \triangleq \alpha_j \cdot \mathbb{1}[j=0] + m\alpha_j \cdot \mathbb{1}[j\neq 0]$. We will take

$$u_{j_1,j_2} \triangleq h_{j_1,j_2}(-1)^{j_1+j_2} \quad \forall j_1,j_2\in\mathcal{J}.$$

Observe that $\operatorname{sgn}(u_{j_1,j_2}) = (-1)^{j_1+j_2}$ as desired.

By taking the inverse Fourier transform of $\widehat{H}$, we get that

$$H(x_1,x_2) = \sum_{j_1,j_2} u_{j_1,j_2} e^{2\pi i\langle \nu_{j_1,j_2},\mathbf{x}\rangle}. \tag{9.5}$$

To complete our proof, it therefore suffices to show that $H(\mathbf{x}) \leq \exp(-\Omega(\varepsilon\sqrt{k}))$ for all $\|x\| \leq 1/\pi\sigma$.

Let $R \subseteq \mathbb{R}^2$ denote the region $[-1/2,1/2]\times[-1/2,1/2]$. But as x ranges over the L_2 ball of norm $1/\pi\sigma$, $\left(\frac{x_1}{m}-\frac{1}{2}, \frac{x_2\sqrt{3}}{m}-\frac{1}{2}\right)$ ranges over the interior of the ellipse centered at $(-1/2,1/2)$ with axes of length $\underline{\gamma}(1-\varepsilon)$ and $2(1-\varepsilon)$. For the subsequent discussion in this paragraph, we refer the reader to Figure 9-4. By periodicity of K^r_ℓ, the image of this ellipse under K^r_ℓ is identical to the region $T \triangleq R\backslash S$, where S is defined as follows. Denote the interior of the ellipse by B_1, and denote its translates along the vectors $(0,1)$, $(1,0)$, and $(1,1)$ by B_2,B_3,B_4. Define S to be the set of points in R that belong to none of B_1,B_2,B_3,B_4.

We claim that S contains the origin-centered L_∞ ball of radius $\varepsilon/2\sqrt{2}$. Note that S is given by translating the four connected components of $([-1,0]\times[-1,0])\backslash B_1$, which is nonempty because B_1 consists of points (x_1,x_2) satisfying

$$\frac{4}{\underline{\gamma}^2(1-\varepsilon)^2}(x_1-1/2)^2 + \frac{1}{(1-\varepsilon)^2}(x_2-1/2)^2 \leq 1. \tag{9.6}$$

In particular, for $x_1,x_2\in[-1,0]$ satisfying $|x_1-1/2|,|x_2-1/2| > (1-\varepsilon)/2$, observe that the left-hand quantity in (9.6) satisfies

$$\frac{4}{\underline{\gamma}^2(1-\varepsilon)^2}(x_1-1/2)^2 + \frac{1}{(1-\varepsilon)^2}(x_2-1/2)^2 > \left(\frac{4}{\underline{\gamma}^2(1-\varepsilon)^2}+\frac{1}{(1-\varepsilon)^2}\right)\cdot\frac{(1-\varepsilon)^2}{4} = 1,$$

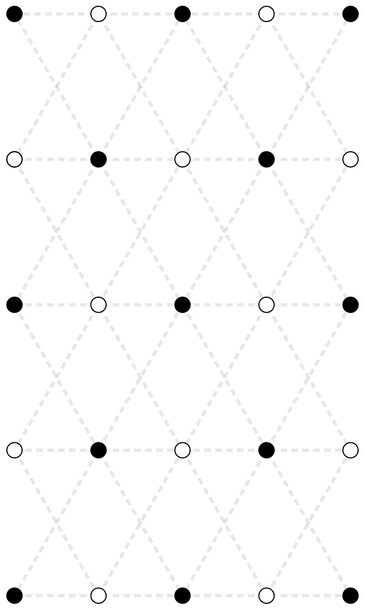


Figure 9-5: Locations of centers of Airy disks for the two mixtures in the lower bound instance of Theorem 9.5.1 when $k = 25$. Black (resp. white) points correspond to centers for ρ (resp. ρ'). The separation between any adjacent pair of identically colored points is $2/m = \Delta$, and the points of any particular color form a triangular lattice.

where the last step follows by our choice of $\underline{\gamma} = \sqrt{4/3}$. We conclude that S contains the origin-centered L_∞ ball of radius $\varepsilon/2$ as claimed.

Now by Fact 9.2.3, for any $(x_1, x_2) \in R$ we have that $K^r_\ell(x_1), K^r_\ell(x_2) \le \frac{1}{4^{2r}\ell^{4r}x^{4r}}$. So because S contains the origin-centered L_∞ ball of radius $\varepsilon/2$, for $(x_1, x_2) \in T$ we conclude that $K^r_\ell(x_1), K^r_\ell(x_2) \le 1/4^{4r}$. We conclude that for $\|\mathbf{x}\| \le 1/\pi\sigma$, $H(\mathbf{x}) \le \exp(-\Omega(r)) = \exp(-\Omega(\varepsilon\sqrt{k}))$.

The last step is just to scale u so that (9.3) holds. First note that by substituting $x = 0$ into (9.5), we have that

$$\sum_{j_1, j_2 \in \mathcal{J}} u_{j_1,j_2} = H(0,0) = K^{2r}_\ell(-1/2) = \frac{1}{\ell^{4r}} \sin^{4r}(\ell\pi/2).$$

In particular, because we assumed at the outset that ℓ is even, $H(0,0) = 0$. Together with the fact that $\mathrm{sgn}(u_{j_1,j_2}) = (-1)^{j_1+j_2}$, we get the first equality in (9.3). Finally, note that $\sum |u_{j_1,j_2}| > 1$ because $\sum \alpha_j = 1$ and $h_{j_1,j_2} \ge \alpha_{j_1}\alpha_{j_2}$ for all j_1, j_2. Thus, by multiplying u by a factor of at most 2, we get the second equality in (9.3). □

9.3 Preliminaries

In this section we explain the terminology and notation that we will adopt in this chapter and also provide some technical preliminaries that will be useful later.

Generative Model We first restate the family of distributions we study in this chapter, originally introduced in Definition 1.2.24.

Definition 9.3.1. *[Superpositions of Airy Disks] A* superposition of k Airy disks ρ *is a distribution over* $\mathbb{R}^2$ *specified by relative intensities* $\lambda_1, ..., \lambda_k \geq 0$ *summing to 1, centers* $\mu_1, ..., \mu_k \in \mathbb{R}^2$*, and an* a priori *known "spread parameter"* $\sigma > 0$*. Its density is given by*

$$\rho(\mathbf{x}) = \sum_{i=1}^{k} \lambda_i \cdot A_\sigma\left(\mathbf{x} - \mu_i\right) \quad \textit{for} \quad A_\sigma(\mathbf{z}) = \frac{1}{\pi\sigma^2}\left(\frac{J_1(\|\mathbf{z}\|_2/\sigma)}{\|\mathbf{z}\|_2/\sigma}\right)^2.$$

Note that the factor of $\frac{1}{\pi\sigma^2}$ *in the definition of* A_σ *is to ensure that* $A_\sigma(\cdot)$ *is a probability density.*

Also define

$$\Delta \triangleq \min_{i \neq j} \|\mu_i - \mu_j\|_2 \quad \textit{and} \quad \mathcal{R} \triangleq \max_{i \in [k]} \|\mu_i\|_2.$$

It will be straightforward to extend the above model to take into account error stemming from the fact that the photon detector itself only has finite precision.

Definition 9.3.2 (Discretization Error)**.** *Given discretization parameter* $\varsigma > 0$*, we say* $\mathbf{x}$ *is a* ς-granular sample *from* ρ *if it is produced via the following generative process: 1) a point* $\mathbf{x}'$ *is sampled from* ρ*, 2)* $\mathbf{x}$ *is obtained by moving* $\mathbf{x}'$ *an arbitrary distance of at most* ς*.*

Optical Transfer Function The following is a standard calculation.

Fact 9.3.3. $\widehat{A}_\sigma[\omega] = \frac{2}{\pi}(\arccos(\pi\sigma\|\omega\|) - \pi\sigma\|\omega\|\sqrt{1 - \pi^2\sigma^2\|\omega\|^2}$.

Proof. It is enough to show this for $\sigma = 1$. Let $G(\mathbf{x}) \triangleq J_1(\|\mathbf{x}\|)/\|\mathbf{x}\|$. It is a standard fact that the zeroeth-order Hankel transform of the function $r \mapsto J_1(r)/r$ is the indicator function of the interval $[0, 1]$. Using our convention for the Fourier transform (see (1.13)), this implies that $\widehat{G}[\omega] = 2\pi \cdot \mathbb{1}[\|\omega\| \in [0, 1/2\pi]]$. Because $A_1 = G^2/\pi$, by the convolution theorem we conclude that $\widehat{A}_1$ is $\frac{1}{\pi}$ times the convolution of $\widehat{G}$ with itself, which is just $4\pi^2$ times the convolution of the indicator function of the unit disk of radius $1/2\pi$ with itself. By elementary Euclidean geometry one can compute this latter function to be $\omega \mapsto \frac{1}{2\pi^2} \cdot \left(\arccos(\pi\|\omega\|) - \pi\|\omega\|\sqrt{1-\pi^2\|\omega\|^2}\right)$, from which the claim follows. □

In optics, the two-dimensional Fourier transform of the point-spread function is called the *optical transfer function*, a term we will occasionally use in the sequel.

Now note that by Fact 9.3.3, $\widehat{A}_\sigma$ is supported only over the disk of radius $\frac{1}{\pi\sigma}$ centered at the origin in the frequency domain. In the spatial domain, this corresponds to a separation of $\pi\sigma$; this is the definition of the *Abbe limit*. We will need the following elementary estimate for $\widehat{A}[\omega]$:

Fact 9.3.4. *For all $\|\omega\|_2 \le 1$, $\widehat{A}[\omega] \ge (1 - \|\omega\|_2)^2$.*

Scaling As the algorithms we give will be scale-invariant, we will assume that $\sigma = 1/\pi$ in the rest of this chapter and refer to $A_{1/\pi}$ as A.

Parameter Estimation Accuracy The following terminology formalizes what it means for an algorithm to return an accurate estimate for the parameters of a superposition of Airy disks.

Definition 9.3.5. $\left(\{\lambda^*_i\}_{i\in[k]}, \{\mu^*_i\}_{i\in[k]}\right)$ *is an $(\varepsilon_1, \varepsilon_2)$-accurate estimate for the parameters of a superposition of k Airy disks ρ with centers $\{\mu_i\}$ and relative intensities $\{\lambda_i\}$ if there exists a permutation τ for which*

$$\|\mu_i - \widetilde{\mu}_{\tau(i)}\|_2 \le \varepsilon_1 \ \ \textit{and} \ \ |\lambda_i - \widetilde{\lambda}_{\tau(i)}| \le \varepsilon_2$$

for all $i \in [k]$.

Generalized Eigenvalue Problems Given matrices M, N, we will denote by (M, N) the *generalized eigenvalue problem $Mx = \lambda Nx$*. In any solution (λ, x) to this, λ is called a *generalized eigenvalue* and x is called a *generalized eigenvector*.

Bessel Function Estimates In Section 9.5, we need the following estimate for $J_\nu(z)$:

Theorem 9.3.6 ([Lan00]). *For some absolute constant* $c_{28} = 0.7857...$, *we have for all* $\nu \geq 0$ *and* $r \in \mathbb{R}$ *that* $|J_\nu(r)| \leq c_{28}|r|^{-1/3}$.

Matrices, Tensors, and Flattenings In this chapter, given a matrix $M \in \mathbb{C}^{a\times b}$, we will denote its i-th row vector by M_i, its j-th column vector by M^j, and its (i,j)-th entry by $M_{i,j}$.

Given a tensor $\mathbf{T} \in \mathbb{C}^{m_1\times m_2\times m_3}$ and matrices $M_1 \in \mathbb{C}^{m_1\times m'_1}$, $M_2 \in \mathbb{C}^{m_2\times m'_2}$, and $M_3 \in \mathbb{C}^{m_3\times m'_3}$, define the flattening $\mathbf{T}(M_1, M_2, M_3) \in \mathbb{C}^{m'_1\times m'_2\times m'_3}$ by

$$\mathbf{T}(M_1, M_2, M_3)_{i_1,i_2,i_3} = \sum_{(j_1,j_2,j_3)\in[m_1]\times[m_2]\times[m_3]} \mathbf{T}_{m_1,m_2,m_3} \cdot (M_1)_{j_1,i_1}(M_2)_{j_2,i_2}(M_3)_{j_3,i_3}$$

for all $(i_1, i_2, i_3) \in [m'_1] \times [m'_2] \times [m'_3]$.

Miscellaneous Notation Let $\mathbb{S}^{d-1}$ denote the Euclidean unit sphere. Given $r > 0$, let $B^d(r)$ denote the Euclidean ball of radius r centered at the origin in $\mathbb{R}^d$.

9.4 Learning Superpositions of Airy Disks

In this section we present the technical details of our algorithmic results. In Sections 9.4.2 and 9.4.4, we prove the following formal version of Theorem 9.1.3.

Theorem 9.4.1. *Let* ρ *be a* Δ*-separated superposition of* k *Airy disks with minimum mixing weight* $\lambda_{\min}$ *and such that* $\|\mu_i\| \leq \mathcal{R}$ *for all* $i \in [k]$.

For any $\varepsilon_1, \varepsilon_2 > 0$, *there is some* $\alpha = \mathrm{poly}\left(\log 1/\delta, 1/\lambda_{\min}, 1/\varepsilon_1, 1/\varepsilon_2, \mathcal{R}, (k\sigma/\Delta)^{k^2}\right)^{-1}$ *for which there exists an algorithm with time and sample complexity* $\mathrm{poly}(1/\alpha)$ *which, given* $\varsigma = \mathrm{poly}(\alpha)$*-granular sample access to* ρ, *outputs an* $(\varepsilon_1, \varepsilon_2)$*-accurate estimate for the parameters of* ρ *with probability at least* $1 - \delta$.

Specifically, in Section 9.4.2, we show how one can use the matrix pencil method to recover the parameters for ρ given oracle access to the *optical transfer function*, i.e. the two-dimensional Fourier transform of ρ, up to some small additive error. In Section 9.4.4, we show how to implement this approximate oracle.

In Section 9.4.3, we also use the oracle of Section 9.4.4 to prove the following formal version of Theorem 9.1.2.

Theorem 9.4.2. *Let ρ be a Δ-separated superposition of k Airy disks with minimum mixing weight $\lambda_{\min}$ and such that $\|\mu_i\| \leq \mathcal{R}$ for all $i \in [k]$. Let*

$$\overline{\gamma} = \frac{2j_{0,1}}{\pi} = 1.530\ldots, \tag{9.7}$$

where $j_{0,1}$ is the first positive zero of the Bessel function of the first kind J_0. For any $\Delta > \overline{\gamma} \cdot \pi \cdot \sigma$, the following holds:

For any $\varepsilon_1, \varepsilon_2 > 0$, there is some $\alpha = 1/\text{poly}\,(k, \mathcal{R}, \sigma/\Delta, 1/\lambda_{\min}, 1/\varepsilon_1, 1/\varepsilon_2, 1/(\Delta - \overline{\gamma}))$ for which there exists an algorithm with time and sample complexity $\text{poly}(1/\alpha)$ *which, given* $\text{poly}(\alpha)$*-granular sample access to ρ, outputs an $(\varepsilon_1, \varepsilon_2)$-accurate estimate for the parameters of ρ with probability at least 4/5.*

9.4.1 Reduction to 2D Superresolution

In this section we reduce the problem of learning superpositions of Airy disks to the problem of learning a convex combination of Dirac deltas given the ability to make noisy, band-limited Fourier measurements.

Formally, suppose we had access to the following oracle:

Definition 9.4.3. *An m*-query, η-approximate OTF oracle *$\mathcal{O}$ takes as input a frequency $\omega \in \mathbb{R}^2$ and, given frequencies $\omega_1, ..., \omega_m$, outputs numbers $u_1, ..., u_m \in \mathbb{R}$ for which $|u_j - \widehat{\rho}[\omega_j]| \leq \eta$ for all $j \in [m]$.*

Remark 9.4.4. *As we will see in Section 9.4.4, $\mathcal{O}$ will be constructed by sampling some number of points from ρ and computing empirical averages. The number m and accuracy η of queries that $\mathcal{O}$ can answer dictates the sample complexity of this procedure. As we will see in the proofs of Lemma 9.4.14 and Lemma 9.4.21 below, the m that we need to take will be small, so the reader can ignore m and pretend it is unbounded for most of this section.*

Given $\omega \in \mathbb{R}^2$, the Fourier transform of ρ evaluated at frequency ω is given by

$$\widehat{\rho}[\omega] = \sum_{j=1}^{k} \lambda_j \widehat{A}[\omega] e^{-2\pi i \langle \mu_j, \omega \rangle}, \tag{9.8}$$

where for $\omega = (r\cos\theta, r\sin\theta)$, we have by Fact 9.3.3 that

$$\widehat{A}[\omega] = \frac{2}{\pi}(\arccos(r) - r\sqrt{1-r^2}).$$

In particular, $\widehat{A}[\omega]$ only depends on $r = \|\omega\|$ (because $A(\cdot)$ is radially symmetric), so henceforth regard $\widehat{A}$ as a function merely of r.

Define

$$F(\omega) = \sum_{j=1}^{k} \lambda_j e^{-2\pi i \langle \mu_j, \omega \rangle}.$$

This is a trigonometric polynomial to which we have noisy pointwise access using $\mathcal{O}$:

Lemma 9.4.5. *Let $0 < r < 1$. With an η-approximate OTF oracle $\mathcal{O}$, on input $\omega \in B^2(r)$ we can produce an estimate of $F(\omega)$ to within $\eta/\widehat{A}[r]$ additive error.*

Proof. By dividing by $\widehat{A}[\omega]$ on both sides of (9.8), we get that

$$\frac{\widehat{\rho}[\omega]}{\widehat{A}[\|\omega\|]} = \sum_{j=1}^{k} \lambda_j e^{-2\pi i \langle \mu_j, \omega \rangle},$$

so given that $\mathcal{O}$, on input ω, outputs $u \in \mathbb{R}$ satisfying $|u - \widehat{\rho}[\omega]| \le \eta$, we have that

$$\left| \frac{u}{\widehat{A}[\|\omega\|]} - F(\omega) \right| \le \frac{\eta}{\min_{0 \le r' \le r} \widehat{A}[r']} = \frac{\eta}{\widehat{A}[r]},$$

where the last step uses the fact that $\widehat{A}[\cdot]$ is decreasing on the interval $[0, 1]$. □

So concretely, given an η-approximate OTF oracle, we have reduced the problem of learning superpositions of Airy disks to that of recovering the locations of $\{\mu_j\}$ given the ability to query $F(\omega)$ at arbitrary frequencies ω for which $\|\omega\|_2 < 1$ to witin additive accuracy $\eta/\widehat{A}[\|\omega\|_2]$.

Lastly, for reasons that will become clear in subsequent sections (see e.g. (9.18)), it will be convenient to assume that $\mathcal{R} \leq 1/3$. This is without loss of generality, as otherwise, we can scale the data down by a factor of $3\mathcal{R}$ so that they are now i.i.d. samples from the superposition of Airy disks with density $\rho'(\mathbf{x}) \triangleq \sum_{j=1}^k \lambda_j \cdot A_{1/\mathcal{R}}(\mathbf{x} - \mu_j/\mathcal{R})$. Define the rescaled centers $\mu'_j \triangleq \mu_j/\mathcal{R}$ and note that by assumption, $\|\mu'_j\|_2 \leq 1/3$ for all $j \in [k]$.

The Fourier transform of ρ' is then given by $\widehat{\rho}'(\omega) = \widehat{A}_{1/\mathcal{R}}[\omega] \sum_{j=1}^k \lambda_j e^{-2\pi i \langle \mu'_j, \omega \rangle}$, so by the proof of Lemma 9.4.5 we conclude that with an η-approximate OTF oracle for ρ, for any $0 < r < 1$ on input $\omega \in B^2(r \cdot \mathcal{R})$ we can produce an estimate of $\sum_{j=1}^k \lambda_j e^{-2\pi i \langle \mu'_j, \omega \rangle}$ to within $\eta/\widehat{A}[r]$ additive error. Recovering the centers $\{\mu'_j\}$ to within additive error ε then translates to recovering the centers $\{\mu_j\}$ to within additive error $3\mathcal{R}\varepsilon$. For this reason, we will henceforth assume that $\mathcal{R} \leq 1/3$.

9.4.2 Learning via the Optical Transfer Function

Our basic approach is as follows. To solve the superresolution problem of Section 9.4.1, we will project in two random, correlated directions $\omega_1, \omega_2 \in \mathbb{R}^2$ and solve the resulting one-dimensional superresolution problems via matrix pencil method (see ModifiedMPM) to recover the projections of $\mu_1, ..., \mu_k$ in the directions ω_1 and ω_2, as well as the relative intensities $\lambda_1, ..., \lambda_k$. From these projections we can then recover the actual centers for ρ by solving a linear system (PreConsolidate). Such an approach already achieves constant success probability, and we can amplify this by repeating and running a simple clustering algorithm (see Select). The full specification of the algorithm is given as LearnAiryDisks.

Learning in a Random Direction

Fix a unit vector $v \in \mathbb{S}^1$. We first show how to leverage Lemma 9.4.5 and the matrix pencil method to approximate the projection of $\mu_1, ..., \mu_k$ along v.

By the discussion at the end of Section 9.4.1, we may assume $\|\mu_i\|_2 \leq 1/2$ for all $i \in [k]$, so $\|\mu_i - \mu_j\|_2 \leq 1$ for all $i \neq j$. For $j \in [k]$, let $m_j = \langle \mu_j, v \rangle$ and $\alpha_j = e^{2\pi i \cdot (m_j/4k)}$. In this section we will assume that $m_j \neq 0$ for all $j \in [k]$

For $\ell \in \mathbf{Z}_{\geq 0}$, let

$$v_\ell = F\left(\frac{\ell v}{4k}\right) = \sum_{j=1}^{k} \lambda_j \alpha_j^\ell.$$

Note that $v_0 = F(\mathbf{0}) = \sum_j \lambda_j = 1$. Also note that we do not have access to $\alpha_1, ..., \alpha_k$ and would like to recover m_1, m_2 given (noisy) access to $\{v_\ell\}_{0\leq \ell \leq 2k-1}$.

Consider the generalized eigenvalue problem $(VD_\lambda V^\top, VD_\lambda D_\alpha V^\top)$ where

$$V = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_k \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{k-1} & \alpha_2^{k-1} & \cdots & \alpha_k^{k-1} \end{pmatrix}, \qquad D_\lambda = \mathrm{diag}(\lambda), \qquad D_\alpha = \mathrm{diag}(\alpha).$$

The following standard facts are key to the matrix pencil method:

Observation 9.4.6. *The generalized eigenvalues of* $(VD_\lambda V^\top, VD_\lambda D_\alpha V^\top)$ *are exactly* $\alpha_1, ..., \alpha_k$.

Observation 9.4.7.

$$VD_\lambda V^\top = \begin{pmatrix} v_0 & v_1 & \cdots & v_{k-1} \\ v_1 & v_2 & \cdots & v_k \\ \vdots & \vdots & \ddots & \vdots \\ v_{k-1} & v_k & \cdots & v_{2k-2} \end{pmatrix} \qquad VD_\lambda D_\alpha V^\top = \begin{pmatrix} v_1 & v_2 & \cdots & v_k \\ v_2 & v_3 & \cdots & v_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ v_k & v_{k+1} & \cdots & v_{2k-1} \end{pmatrix}.$$

By Lemma 9.4.5, in reality we only have η'_ℓ-approximate access to each v_ℓ, where

$$\eta'_\ell \leq \frac{\eta}{\widehat{A}[\ell/4k]}, \tag{9.9}$$

so we must instead work with the generalized eigenvalue problem $(VD_\lambda V^\top + E, VD_\lambda D_\alpha V^\top + F)$, where the (i,j)-th entry of E (resp. F) is the noise η'_{i+j-2} (resp. η'_{i+j-1}) in the observation of v_{i+j-2} (resp. v_{i+j-1}).

If V is well-conditioned, one can apply standard perturbation bounds to argue that the solutions to this generalized eigenvalue problem are close to those of the original $(VD_\lambda V^\top, VD_\lambda D_\alpha V^\top)$. Moreover, given approximations $\widehat{\alpha}_1, ..., \widehat{\alpha}_k$ to these generalized eigenvalues, we can find approximations $\widehat{\lambda}_1, ..., \widehat{\lambda}_k$ to $\lambda_1, ..., \lambda_k$ by solving the system of equations $\mathbf{v} = \widehat{V}\lambda$, where

$\mathbf{v} = (v_0, ..., v_{k-1})$, $\lambda = (\widehat{\lambda}_1, ..., \widehat{\lambda}_k)$, and

$$\widehat{V} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \widehat{\alpha}_1 & \widehat{\alpha}_2 & \cdots & \widehat{\alpha}_k \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\alpha}_1^{k-1} & \widehat{\alpha}_2^{k-1} & \cdots & \widehat{\alpha}_k^{k-1} \end{pmatrix}.$$

The formal specification of the matrix pencil method algorithm ModifiedMPM that we use is given in Algorithm 40.

```
Algorithm 40: ModifiedMPM($\omega, \mathcal{O}$)
  Input: $\omega \in \mathbb{S}^1$, $\eta$-approximate OTF oracle $\mathcal{O}$
  Output: Estimates $(\widehat{\lambda}_1, ..., \widehat{\lambda}_k)$ for the mixing weights and $(\widehat{m}_1, ..., \widehat{m}_k)$ for the
          centers of $\rho$ projected in direction $\omega$
  $\widehat{v}_0 \leftarrow 1$.
  for $0 \le \ell \le 2k-1$ do
      Invoke $\mathcal{O}$ on input $\frac{\ell\omega}{4k}$ to produce $u_\ell \in \mathbb{R}$.
      $\widehat{v}_\ell \leftarrow \frac{u_\ell}{\widehat{A}[\ell/4k]}$.
  Form the matrices
      $X \triangleq \begin{pmatrix} \widehat{v}_0 & \cdots & \widehat{v}_{k-1} \\ \vdots & \ddots & \vdots \\ \widehat{v}_{k-1} & \cdots & \widehat{v}_{2k-2} \end{pmatrix} \qquad Y \triangleq \begin{pmatrix} \widehat{v}_1 & \cdots & \widehat{v}_k \\ \vdots & \ddots & \vdots \\ \widehat{v}_k & \cdots & \widehat{v}_{2k-1} \end{pmatrix}$
  Solve the generalized eigenvalue problem $(X, Y)$ to produce generalized eigenvalues
    $\widehat{\alpha}_1, \widehat{\alpha}_2$.
  For $i = 1, 2$, let $\widehat{m}_i$ be the argument of the projection of $\widehat{\alpha}_i$ onto the complex unit
    disk.
  Form the matrix
      $\widehat{V} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \widehat{\alpha}_1 & \widehat{\alpha}_2 & \cdots & \widehat{\alpha}_k \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\alpha}_1^{k-1} & \widehat{\alpha}_2^{k-1} & \cdots & \widehat{\alpha}_k^{k-1} \end{pmatrix}.$
  Solve for $\widehat{\lambda} = (\widehat{\lambda}_1, ..., \widehat{\lambda}_k)$ such that $\widehat{V}\widehat{\lambda} = (\widehat{v}_0, ..., \widehat{v}_{k-1})$.
  return $\{\widehat{\lambda}_i\}_{i \in [k]}$ and $\{\widehat{m}_i\}_{i \in [k]}$.
```

The following theorem, implicit in the proof of Theorem 2.8 in [Moi15], makes the above reasoning precise. Henceforth, let $\kappa(\Delta')$ and $\sigma_{\min}(\Delta')$ respectively denote the condition number and minimum singular value of V when $\frac{m_i}{4k}, \frac{m_j}{4k}$ have minimum separation Δ' for all $i \neq j$, and define $\lambda_{\min} = \min_i \lambda_i$, $\lambda_{\max} = \max_i \lambda_i$.

Theorem 9.4.8 ([Moi15]). *Suppose $\frac{m_1}{4k}, \frac{m_2}{4k} \in [-1/4, 1/4]$ have separation at least Δ' and we are given η'_ℓ-close estimates to v_ℓ for $0 \leq \ell \leq 2k-1$.*

Define

$$\gamma = \frac{2\|\eta'\|_2}{\lambda_{\min}}\left(4\kappa(\Delta')^2 \cdot \frac{\lambda_{\max}}{\lambda_{\min}} + \frac{1}{\sigma_{\min}(\Delta')^2}\right) \quad \text{and} \quad \zeta = O\left(\frac{2\gamma\lambda_{\max} + \|\eta'\|_2}{\sigma_{\min}(\Delta' - 2\gamma)}\right)$$

Then if $\|E\| + \|F\| < \sigma_{\min}(\Delta')^2\lambda_{\min}$ and $\gamma < \Delta'/4$, ModifiedMPM *produces estimates $\{\widehat{\lambda}_i\}$ for the mixing weights and estimates $\{\widehat{m}_i\}$ for the projected centers such that for some permutation τ:*

$$|m_i - \widehat{m}_{\tau(i)}| \leq 8\gamma \quad \text{and} \quad |\lambda_i - \widehat{\lambda}_i| \leq \zeta.$$

for all $i \in [k]$.

Note that the guarantees of Theorem 9.4.8 are stated in [Moi15] in terms of wraparound distance on the interval $[-1/2, 1/2]$, but because $\frac{m_i}{4k} \in [-1/4, 1/4]$ for all $j \in [k]$, $\frac{m_1}{4k}, \ldots., \frac{m_k}{4k}$ have pairwise separation Δ' both in absolute and wraparound distance.

In other words, the output of ModifiedMPM converges to the true values for $\{\langle \mu_1, v\rangle\}_{j\in[k]}$ and $\{\lambda_j\}_{j\in[k]}$ at a rate polynomial in the noise rate, condition number of V, and relative intensity of the Airy disks, provided $\sigma_{\min}(\Delta')$ is inverse polynomially large and $\kappa(\Delta')$ is polynomially small in those parameters.

To complete the argument, we must establish these bounds on $\sigma_{\min}$ and κ. Henceforth, let

$$\Delta' = \min_{i\neq j} \frac{m_i - m_j}{4k}.$$

Lemma 9.4.9. *For any $k \geq 2$, we have that*

$$\sigma_{\min}(\Delta')^2 \geq (\Delta'^k/k^2)^{k-1} \quad \text{and} \quad \kappa(\Delta')^2 \leq k^{2k-1}/\Delta'^{k(k-1)}$$

Proof. First note that $\sigma_{\max}(\Delta')^2 \leq k^2$. Indeed because the entries of V all have absolute value at most 1, we conclude that for any $v \in \mathbb{S}^{k-1}$ and any row index $j \in [k]$,

$$\langle V_j, v\rangle^2 \le \left(\sum_{i=1}^k |v_j|\right)^2 \le k.$$

On the other hand, we also have that

$$\prod_{i=1}^k \sigma_i(V) = |\det(V)| = \prod_{1\le i<j\le k} |\alpha_i - \alpha_j| \le \left|e^{2\pi i\Delta'} - 1\right|^{\binom{k}{2}} = (2 - 2\cos(\Delta'))^{\binom{k}{2}/2} \ge \Delta'^{k(k-1)/2},$$

where in the first step we used the standard fact that the absolute value of the determinant of a square matrix is equal to the product of its singular values, in the second step we used the standard identity for the determinant of a Vandermonde matrix, in the third step we used the angular separation of the α_i's, and in the final step we used the elementary inequality $\cos(\Delta') \le 1 - \Delta'^2/2$. We may thus naively lower bound $\sigma_{\min}(V)$ by $\frac{\Delta'^{k(k-1)/2}}{k^{k-1}}$, from which the lemma follows. □

This yields the following consequence for ModifiedMPM.

Corollary 9.4.10. *Given $\omega \in \mathbb{S}^1$ and access to an η-approximate OTF oracle $\mathcal{O}$, if the projected centers $m_j = \langle \mu_j, v\rangle$ satisfy $|m_i - m_j| \le 4k\cdot\Delta'$ for all $i \ne j$ for some $0 < \Delta' \le 1/16$, then there exists a constant $c_{29} > 0$ such that provided that*

$$\eta \le c_{29}\lambda_{\min}^2 \Delta'^{k^2} \cdot k^{-2k-1/2}, \tag{9.10}$$

then ModifiedMPM *produces estimates $\{\widehat{\lambda}_i\}$ for the mixing weights and estimates $\{\widehat{m}_i\}$ for the projected centers such that for some permutation τ:*

$$|m_i - \widehat{m}_{\tau(i)}| \le O\left(\frac{k^{2k+1/2}\cdot\eta}{\lambda_{\min}^2\Delta'^{k(k-1)}}\right) \quad \text{and} \quad |\lambda_i - \widehat{\lambda}_{\tau(i)}| \le O\left(\frac{k^{3k-1/2}\cdot\eta}{\lambda_{\min}^2\Delta'^{3k(k-1)/2}}\right)$$

for all $i \in [k]$.

Proof. From Lemma 9.4.9, we have that $\sigma_{\min}(\Delta')^2 \ge (\Delta'^k/k^2)^{k-1}$. Then because $\kappa(\Delta')^2 \le$

$k^2/\sigma_{\min}(\Delta')^2$, we would like to conclude by Theorem 9.4.8 that $|m_i - \widehat{m}_{\tau(i)}| \le 8\gamma$, where

$$\gamma = O\left(\frac{\|\eta'\|_2 \cdot k^2}{\lambda_{\min}^2 \cdot \sigma_{\min}(\Delta')^2}\right) = O\left(\frac{\|\eta'\|_2 \cdot k^2}{\lambda_{\min}^2 \cdot (\Delta'^k/k^2)^{k-1}}\right) = O\left(\frac{k^{2k+1/2} \cdot \eta}{\lambda_{\min}^2 \Delta'^{k(k-1)}}\right),$$

where in the last step we use that the vector η' has length $O(k)$ and satisfies $\|\eta'\|_\infty \le O(\eta)$ by (9.9). To do so, we just need to verify that $\|E\| + \|F\| < \sigma_{\min}(\Delta')^2\lambda_{\min}$ and $\gamma < \Delta'/4$. The latter clearly follows from the bound (9.10) for sufficiently small c_{29}. For the former, note that

$$\|E\|_2 \le \|E\|_F \le \sqrt{k} \cdot \sqrt{\eta_1'^2 + \cdots + \eta_{2k-1}'^2} \le \eta\sqrt{k} \cdot \sqrt{\sum_{\ell=1}^{2k-1} \frac{1}{\widehat{A}[\ell/4k]}} \le O(\eta \cdot k),$$

where the last step follows by the fact that $\widehat{A}[\ell/4k] \ge \widehat{A}[1/2] \ge \Omega(1)$. The same bound holds for $\|F\|_2$. Recalling that $\sigma_{\min}(\Delta')^2 \ge (\Delta'^k/k^2)^{k-1}$, it is enough for $\eta \le O(\Delta'^k/k^2)^{k-1}\lambda_{\min}/k$, which certainly holds for η satisfying (9.10), for c_{29} sufficiently small.

Finally, Theorem 9.4.8 also implies that $|\lambda_i - \widehat{\lambda}_i| \le \zeta$, where

$$\zeta \le O\left(\frac{\gamma + k^{1/2}\eta}{\sigma_{\min}(\Delta' - 2\gamma)}\right) \le O\left(\frac{k^{2k+1/2} \cdot \eta}{\lambda_{\min}^2 \Delta'^{k(k-1)} \cdot (\Delta'^{k(k-1)/2}/k^{k-1})}\right) = O\left(\frac{k^{3k-1/2} \cdot \eta}{\lambda_{\min}^2 \Delta^{3k(k-1)/2}}\right)$$

as claimed. □

Combining Directions

We can run ModifiedMPM to approximately recover $\{\langle \mu_j, \omega_1\rangle\}_{j\in[k]}$ and $\{\langle \mu_j, \omega_2\rangle\}_{j\in[k]}$ for two randomly chosen directions $\omega_1, \omega_2 \in \mathbb{S}^1$. As these directions are random, with high probability we can combine these estimates to obtain an accurate estimate of $\{\mu_j\}_{j\in[k]}$. One subtlety is that the estimates $\{\widehat{m}_j\}$ and $\{\widehat{m}'_j\}$ output by ModifiedMPM for the centers projected in directions ω_1 and ω_2 respectively need not be aligned, that is we only know that there exists some permutation τ for which $\widehat{m}_j = \widehat{m}'_{\tau(j)}$ for $j \in [k]$.

We first show a "pairing lemma" stating that if ω_1 is chosen randomly and ω_2 is chosen to be close to ω_1, then if one sorts the centers $\mu_1, ..., \mu_k$, first in terms of their projections in the ω_1 direction, and then in terms of their projections in the ω_2 direction, the corresponding elements in these two sorted sequences will correspond to the same centers.

We require the following elementary fact.

Lemma 9.4.11. *For $\mu \in \mathbb{R}^2$ a unit vector and $\omega \in \mathbb{R}^2$ a random unit vector, $Pr_\omega[|\langle \mu, \omega \rangle| \le \sin\theta] = 2\theta/\pi$ for all $0 \le \theta \le \pi/2$.*

Lemma 9.4.12. *Fix an arbitrary $0 < \theta \le \pi/2$ and let $\upsilon = \frac{\Delta \sin\theta}{8}$. Let $\omega_1 \in \mathbb{R}^2$ be a random unit vector, and let $\omega_2 \in \mathbb{R}^2$ be either of the two unit vectors for which $\|\omega_1 - \omega_2\|_2 = \upsilon$. For every $i \in [k]$, define $m_i \triangleq \langle \mu_i, \omega_1 \rangle$ and $m'_i \triangleq \langle \mu_i, \omega_2 \rangle$, and let $\widehat{m}_i, \widehat{m}'_i \in \mathbb{R}$ be any numbers for which $\|\widehat{m}_i - m_i\|_2, \|\widehat{m}'_i - m'_i\|_2 \le 2\upsilon$.*

Then with probability at least $1 - \frac{k(k-1)\theta}{\pi}$, for every $i \neq j$ the following are equivalent: I) $m_i > m_j$, II) $m'_i > m'_j$, III) $\widehat{m}_i > \widehat{m}_j$, and IV) $\widehat{m}'_i > \widehat{m}'_j$.

Proof. By Lemma 9.4.11 and a union bound we have that with probability $1 - \frac{k(k-1)\theta}{\pi}$, $|m_i - m_j| > \Delta \sin\theta$ for all $i \neq j$. Fix any $i \neq j$ and suppose that $m_i > m_j$. Then by triangle inequality and Cauchy-Schwarz, we have that

$$m'_i - m'_j = \langle \mu_1 - \mu_2, \omega_2 \rangle = \langle \mu_1 - \mu_2, \omega_1 \rangle + \langle \mu_1 - \mu_2, \omega_2 - \omega_1 \rangle \ge \Delta \sin\theta - 4\upsilon > 0,$$

where the final inequality follows by the definition of υ. So I) implies II) and by symmetry we can show II) implies I). We also have that

$$\widehat{m}_i - \widehat{m}_j \ge (m_i - m_j) - 4\upsilon > 0,$$

so I) implies III) and by symmetry we can show II) implies IV).

It is enough to show that III) implies I). Suppose $\widehat{m}_i > \widehat{m}_j$. Then

$$m_i - m_j \ge (\widehat{m}_i - \widehat{m}_j) - 4\upsilon > -\frac{1}{2}\Delta \sin\theta > -\Delta \sin\theta,$$

so it must be the case that $m_i - m_j > 0$ given that $|m_i - m_j| > \Delta \sin\theta$. □

We now show that we can combine these projected center estimates to approximately recover the two-dimensional centers by solving a linear system. The specification of this algorithm, which we call PreConsolidate, is given in Algorithm 41.

Algorithm 41: $\textsc{PreConsolidate}(\omega_1, \omega_2, \{\widehat{\lambda}_i, \widehat{\lambda}'_i\}, \{\widehat{m}_i, \widehat{m}'_i\})$

```
Input: Directions $\omega_1, \omega_2 \in \mathbb{S}^1$ and estimates $\{\widehat{\lambda}_i\}, \{\widehat{m}_i\}$ and $\{\widehat{\lambda}'_i\}, \{\widehat{m}'_i\}$ for the
       parameters of $\rho$ projected in the directions $\omega_1$ and $\omega_2$ respectively
Output: An estimate of the form $(\{\widetilde{\lambda}_i\}, \{\widetilde{\mu}_i\})$ for the parameters of $\rho$
for $i \in [k]$ do
    Let $\ell, \ell' \in [k]$ be the indices for which $\widehat{m}_\ell$ and $\widehat{m}'_{\ell'}$ are the $i$-th largest in $\{\widehat{m}_j\}_{j\in[k]}$
     and $\{\widehat{m}'_j\}_{j\in[k]}$ respectively.
    Define a formal vector-valued variable $\mathbf{v}^{(i)} \in \mathbb{R}^2$ and solve the linear system
        $\langle \omega_1, \mathbf{v}^{(i)} \rangle = \widehat{m}_\ell$
        $\langle \omega_2, \mathbf{v}^{(i)} \rangle = \widehat{m}'_{\ell'}.$
return $\left(\{\widehat{\lambda}_i\}_{i\in[k]}, \{\mathbf{v}^{(i)}\}_{i\in[k]}\right)$.
```

Lemma 9.4.13. *Let $\xi > 0$. Let the parameters θ, υ and the random vectors ω_1, ω_2 be as in Lemma 9.4.12. Suppose $\{\widehat{m}_i\}_{i\in[k]}$ and $\{\widehat{m}'_i\}_{i\in[k]}$ are collections of numbers for which there exist permutations $\tau_1, \tau_2 \in \mathbb{S}_k$ for which*

$$\left|\langle \omega_1, \mu_i \rangle - \widehat{m}_{\tau_1(i)}\right| \le \xi \quad \text{and} \quad \left|\langle \omega_2, \mu_i \rangle - \widehat{m}'_{\tau_2(i)}\right| \le \xi$$

for all $i \in [k]$.

Then for any estimates $\{\widehat{\lambda}_i\}_{i\in[k]}$ and $\{\widehat{\lambda}'_i\}_{i\in[k]}$, with probability at least $1 - \frac{k(k-1)\theta}{\pi}$ we have that the output $\left(\{\widetilde{\lambda}_i\}, \{\widetilde{\mu}_i\}\right)$ of $\textsc{PreConsolidate}$*(ω_1,ω_2, $\{\widehat{\lambda}_i\}$, $\{\widehat{m}_i\}$, $\{\widehat{\lambda}'_i\}$, $\{\widehat{m}'_i\}$) satisfies*

$$\|\mu_i - \widetilde{\mu}_{\tau(i)}\|_2 \le \frac{\xi}{\upsilon\sqrt{1 - \upsilon^2/4}}$$

for some permutation $\tau \in \mathbb{S}_k$.

Proof. Condition on the event of Lemma 9.4.12 occurring, which happens with probability at least $1 - \frac{k(k-1)\theta}{\pi}$. This event implies that there is a permutation $\tau \in \mathbb{S}_k$ such that for every $i \in [k]$ in the loop of $\textsc{PreConsolidate}$, the indices ℓ, ℓ' in that iteration are such that $\widehat{m}_\ell$ and $\widehat{m}'_{\ell'}$ are ξ-close estimates for the projections of $\mu_{\tau(i)}$ in the directions ω_1 and ω_2 respectively. In other words, $\tau_1(\tau(i)) = \ell$ and $\tau_2(\tau(i)) = \ell'$.

Let $A \in \mathbb{R}^{2\times 2}$ be the matrix with rows consisting of ω_1 and ω_2. We conclude that

$$\|\mu_{\tau(i)} - \mathbf{v}^{(i)}\|_2 = \|A^{-1} \cdot \big((\widehat{m}_\ell, \widehat{m}'_{\ell'}) - \big(\langle \omega_1, \mu_{\tau(i)}\rangle, \langle \omega_2, \mu_{\tau(i)}\rangle\big)\big)\|_2 \le \sigma_{\min}(A) \cdot \xi,$$

so it remains to bound $\sigma_{\min}(A)$. Without loss of generality we may assume $\omega_1 = (1, 0)$ and $\omega_2 = (x, \sqrt{1-x^2})$ for $x \triangleq 1 - \upsilon^2/2$, in which case $\sigma_{\min}(A) = \upsilon\sqrt{1-\upsilon^2/4}$, and the claim follows. □

Finally, we show how to boost the success probability via the following naive clustering-based algorithm SELECT (Algorithm 42), whose guarantees we establish below.

```
Algorithm 42: SELECT($\varepsilon'_1, \varepsilon'_2, \mathcal{L}$)
  Input: Accuracy parameters $\varepsilon'_1, \varepsilon'_2$, and list $\mathcal{L}$ consisting of $T$ candidate estimates
         for the parameters of $\rho$, each of the form $\big(\{\widetilde{\lambda}^t_i\}_{i\in[k]}, \{\widetilde{\mu}^t_i\}_{i\in[k]}\big)$ for $t \in [T]$,
         such that for at least $1 - \frac{1}{2k}$ fraction of all $t \in [T]$, $\big(\{\widetilde{\lambda}^t_i\}_{i\in[k]}, \{\widetilde{\mu}^t_i\}_{i\in[k]}\big)$ is an
         $(\varepsilon'_1, \varepsilon'_2)$-accurate estimate of the parameters of $\rho$
  Output: A $(3\varepsilon'_1, \varepsilon'_2)$-accurate estimate of the parameters of $\rho$
  $\mathcal{S} \leftarrow T \times [k]$.
  Form the graph $G = (V, E)$ whose vertices consist of all $(t, i)$ for which $\widetilde{\mu}^t_i \in \mathcal{S}$ is
    $2\varepsilon'_1$-close to at least $2T/3$ other points in $\mathcal{S}$, with edges between any $(t, i), (t', i')$ for
    which $\|\widetilde{\mu}^t_i - \widetilde{\mu}^{t'}_{i'}\| > 6\varepsilon'_1$.
  $G$ is $k$-partite. Denote the parts by $V^{(1)}, ..., V^{(k)} \subset V$.
  for $j \in [k]$ do
      Form the set $\{\widetilde{\lambda}^t_i\}_{(t,i)\in V^{(j)}}$ and let $\lambda^*_j$ be the median of this set, corresponding to
        some $(t_j, i_j) \in \mathcal{S}$.
      $\mu^*_j \leftarrow \widetilde{\mu}^{t_j}_{i_j}$.
  return $\big(\{\lambda^*_j\}_{j\in[k]}, \{\mu^*_j\}_{j\in[k]}\big)$.
```

We give the full specification of our algorithm LEARNAIRYDISKS in Algorithm 43.

Lemma 9.4.14. *Let ρ be a Δ-separated superposition of k Airy disks. For any $\varepsilon_1, \varepsilon_2, \delta > 0$, let*

$$\eta = O\left(\left(\frac{\Delta}{4k}\right)^{O(k^2)} \cdot \lambda^2_{\min}\right) \cdot \min\{\varepsilon_1/M, \varepsilon_2\}. \tag{9.11}$$

Without loss of generality suppose $\varepsilon_1 < 3\Delta/8$. Then the output $(\lambda^*_1, \lambda^*_2, \mu^*_1, \mu^*_2)$ of LEARNAIRYDISKS*, given $\varepsilon_1, \varepsilon_2, \delta$ and access to an η-approximate, $O(\log(1/\delta))$-query OTF oracle*

```
Algorithm 43: LearnAiryDisks($\varepsilon_1, \varepsilon_2, \delta, \mathcal{O}$)
  Input: Error parameters $\varepsilon_1, \varepsilon_2$, confidence parameter $\delta$, access to $\eta$-approximate,
         $O(\log 1/\delta)$-query OTF oracle $\mathcal{O}$
  Output: With probability at least $1-\delta$, an $(\varepsilon_1, \varepsilon_2)$-accurate estimate $(\{\widetilde{\lambda}_i\}, \{\widetilde{\mu}_i\})$
          for the parameters of $\rho$.
1 $\mathcal{L} \leftarrow \emptyset$.
2 $\theta \leftarrow \frac{\pi}{3k^2(k-1)}$.
3 Set $\eta$ according to (9.11).
4 for $T = 1, ..., \Omega(\log(1/\delta))$ do
5     Sample a random unit vector $\omega_1 \in \mathbb{S}^1$ and let $\omega_2$ be either of the two unit vectors
       for which $\|\omega_1 - \omega_2\|_2 = \Delta \sin\theta/8$.                      // Lemma 9.4.12
6     Run ModifiedMPM($\omega_1, \mathcal{O}$) and ModifiedMPM($\omega_2, \mathcal{O}$) to obtain estimates
       $\{\widehat{\lambda}_i\}, \{\widehat{m}_i\}$ and $\{\widehat{\lambda}'_i\}, \{\widehat{m}'_i\}$ for the parameters of $\rho$ projected in the directions
       $\omega_1, \omega_2$ respectively.
7     Let $\{\widetilde{\lambda}^t_i\}, \{\widetilde{\mu}^t_i\}$ be the estimates output by
       PreConsolidate($\omega_1, \omega_2, \{\widehat{\lambda}_i\}, \{\widehat{m}_i\}, \{\widehat{\lambda}'_i\}, \{\widehat{m}'_i\}$). Append these to $\mathcal{L}$.
8 return Select($\mathcal{L}, \varepsilon_1/3, \varepsilon_2$)).
```

$\mathcal{O}$ for ρ, satisfies

$$\|\mu_i - \mu^*_{\tau(i)}\|_2 \le \varepsilon_1 \;\; \textit{and} \;\; |\lambda_i - \lambda_{\tau(i)}| \le \varepsilon_2$$

for some permutation τ with probability at least $1-\delta$. Furthermore, the runtime of LearnAiryDisks *is dominated by the time it takes to invoke the OTF oracle $O(\log(1/\delta))$ times.*

Proof. Suppose we are given a valid η-approximate OTF oracle $\mathcal{O}$. By taking $\theta = \frac{\pi}{3k^2(k-1)}$ and invoking Lemmas 9.4.10 and 9.4.13, we ensure that a single run of PreConsolidate in an iteration of the loop in Step 4 of LearnAiryDisks will yield, with probability at least $1 - \frac{1}{3k}$, an $(\varepsilon'_1, \varepsilon'_2)$-accurate estimate, where

$$\varepsilon'_1 = \frac{8}{\Delta \sin\theta} \cdot O\left(\frac{k^{2k+1/2} \cdot \eta}{\lambda^2_{\min}\left(\frac{\Delta \sin\theta}{4k}\right)^{k(k-1)}}\right) \quad \text{and} \quad \varepsilon'_2 = O\left(\frac{k^{3k-1/2} \cdot \eta}{\lambda^2_{\min}\left(\frac{\Delta \sin\theta}{4k}\right)^{3k(k-1)/2}}\right).$$

In this case we say that such an iteration of the loop in LearnAiryDisks "succeeds." Note that if we take

$$\eta = O\left(\min\left\{\varepsilon_1 \cdot \frac{\Delta \sin\theta}{8} \cdot \frac{\lambda^2_{\min}\left(\frac{\Delta \sin\theta}{4k}\right)^{k(k-1)}}{k^{2k+1/2}}, \varepsilon_2 \cdot \frac{\lambda^2_{\min}\left(\frac{\Delta \sin\theta}{4k}\right)^{3k(k-1)/2}}{k^{3k-1/2}}\right\}\right),$$

then we can ensure that $\varepsilon_1' = \varepsilon_1/3$ and $\varepsilon_2' = \varepsilon_2$. The bound in (9.11) then follows from the elementary inequality $\sin\theta \geq \theta/2$ for $0 \leq \theta \leq 1$, together with our choice of $\theta = \frac{\pi}{3k^2(k-1)}$.

Each iteration of the loop in Step 4 of LEARNAIRYDISKS individually succeeds with probability at least $1 - \frac{1}{3k}$. So by a Chernoff bound, by taking $T = \Omega(\log(1/\delta))$, we conclude that with probability at least $1-\delta$, at least $1 - \frac{1}{2k}$ fraction of these iterations will succeed. So of the $k \cdot T$ elements in $\mathcal{S}$, at most $T/2$ correspond to failed iterations.

Now note that all (t,i) for which t corresponds to a successful iteration will be $2\varepsilon_1'$-close to at least $k\cdot T - T/2 > 2T/3$ points. In particular, any such (t,i) will be among the vertices V of G in Algorithm 42. Conversely, for any $(t,i) \in V$, $\widetilde{\mu}_i^t$ is by definition $2\varepsilon_1'$-close to at least $2T/3$ points and there are at most $T/2 < 2T/3$ points which do not correspond to successful iterations. In particular, at least one of the points that $\widetilde{\mu}_i^t$ is close to will correspond to a successful iteration, so by the triangle inequality $\|\widetilde{\mu}_i^t - \mu_j\| \leq 3\varepsilon_1'$ for some choice of $j \in [k]$.

Observe that G is k-partite because every vertex in V is $3\varepsilon_1'$-close to some center of ρ, but two vertices which are $3\varepsilon_1'$-close to μ_i and μ_j respectively for $i \neq j$ must be distance at least $\Delta - 6\varepsilon_1' > 2\varepsilon_1'$ apart. We conclude that with high probability, SELECT will output $3\varepsilon_1' = \varepsilon_1$-accurate estimates for the centers of ρ.

It remains to show that λ_1^*, λ_2^* are ε_2-accurate estimates for the mixing weights. We know the estimates $\widetilde{\lambda}_i^t$ corresponding to successful iterations t and center μ_i lie in $\{\widetilde{\lambda}_i^t\}_{(t,i)\in V^{(\ell)}}$ for some ℓ. Then $\{\widetilde{\lambda}_i^t\}_{(t,i)\in V^{(\ell)}}$ contains at least $\left(1 - \frac{1}{2k}\right)T > 2T/3$ values that are ε_2'-close to λ_1, and at most $T/2 < 2T/3$ other values. Call these values "good" and "bad" respectively. Either the median is good, in which case we are done, or the median is bad, in which case because there are strictly more good values than bad values, the median must be upper and lower bounded by good values, in which case we are still done.

Finally, note that in each iteration of the main loop of LEARNAIRYDISKS, $\mathcal{O}$ is invoked exactly six times. Furthermore, other than these invocations of $\mathcal{O}$, the remaining steps of LEARNAIRYDISKS all require constant time. So the runtime of LEARNAIRYDISKS is indeed dominated by the $O(\log(1/\delta))$ calls to $\mathcal{O}$. □

9.4.3 Learning Airy Disks Above the Diffraction Limit

In this subsection we present the proof of Theorem 9.4.2. Recall that we are assuming that $\sigma = 1/\pi$ and $\Delta > \overline{\gamma}$, where $\overline{\gamma}$ is defined in (9.7). Let $c \triangleq \frac{1}{2}(\Delta + \overline{\gamma})$ and define $R \triangleq \frac{\overline{\gamma}}{2c}$ and $r \triangleq 1/2 - R$.

We will use the following Algorithm 44 that we call TensorResolve. While this is only a slight modification of the tensor decomposition algorithm of [HK15] for high-dimensional superresolution, our analysis is novel and obtains sharper results in low dimensions by using certain extremal functions [Gon18, HV+96, CCLM17] arising in the study of de Branges spaces (see Theorem 9.4.19.

```
Algorithm 44: TensorResolve($\varepsilon_1, \varepsilon_2, \delta, \mathcal{O}$)
Input: Error parameters $\varepsilon_1, \varepsilon_2$, confidence parameter $\delta$, access to $\eta$-approximate,
       $\Theta\left(\frac{k^2 \log(k/\delta)}{(\Delta - \overline{\gamma}) \wedge 1}\right)$-query OTF oracle
Output: With probability at least $1 - \delta$, an $(\varepsilon_1, \varepsilon_2)$-accurate estimate $(\{\widetilde{\lambda}_i\}, \{\widetilde{\mu}_i\})$
        for the parameters of $\rho$, provided the separation is sufficiently above the
        diffraction limit (see Lemma 9.4.21)
$R \leftarrow \overline{\gamma}/2c$ and $r \leftarrow 1/2 - R$.
Sample $\omega^{(1)}, ..., \omega^{(m)}$ i.i.d. from the uniform distribution over $B^2(R)$. Also define
  $\omega^{(m+1)} = (1, 0)$, $\omega^{(m+2)} = (0, 1)$, and $\omega^{(m+3)} = (0, 0)$.
$m' \leftarrow m + 3$.
Sample $v$ uniformly from $\mathbb{S}^1$ and define $v^{(1)} \leftarrow r \cdot v$, $v^{(2)} \leftarrow 2r \cdot v$, and $v^{(3)} \leftarrow 0$.
$\xi_{a,b,i} \leftarrow \omega^{(a)} + \omega^{(b)}) + v^{(i)}$ for every $a, b \in [m'], i \in [3]$.
Query $\mathcal{O}$ at $\{\xi_{a,b,i}\}$ to obtain numbers $\{u_{a,b,i}\}$.
Construct the tensor $\widetilde{\mathbf{T}} \in \mathbb{C}^{m' \times m' \times 3}$ given by $\widetilde{\mathbf{T}}_{a,b,i} = u_{a,b,i}/\widehat{A}[\xi_{a,b,i}]$.
$\widehat{V} \in \mathbb{R}^{m' \times k} \leftarrow$ Jennrich($\widetilde{\mathbf{T}}$).
Divide each column $\widehat{V}^j$ by a factor of $\widehat{V}_{m,j}$.
For each $j \in [k]$, $i \in [2]$, let $\widehat{\mu}_j \in \mathbb{R}^2$ have $i$-th entry equal to the argument of the
  projection of $\widehat{V}_{m+i,j}$ onto the complex disk.
Query $\mathcal{O}$ at frequencies $\{\omega^{(a)}\}_{a \in [m']}$ to get numbers $\{u'_a\}_{a \in [m']}$ and form the vector
  $\widehat{b} \in \mathbb{R}^{m'}$ whose $a$-th entry is $u'_a/\widehat{A}[\omega^{(a)}]$ for every $a \in [m']$.
$\widehat{\lambda} \leftarrow \mathrm{argmin}_\lambda \|\widehat{V}\lambda - \widehat{b}\|_2$.
return $(\widehat{\lambda}_1, ..., \widehat{\lambda}_k)$ and $(\widehat{\mu}_1, ..., \widehat{\mu}_k)$.
```

Using the notation of TensorResolve, define the tensor $\mathbf{T} \in \mathbb{C}^{m' \times m' \times 3}$ given by

$$\mathbf{T}_{a,b,i} = \sum_{j=1}^{k} \lambda_j e^{-2\pi i \langle \mu_j, \omega^{(a)} + \omega^{(b)} + v^{(i)} \rangle}$$

and note that it admits a low-rank decomposition as

$$\mathbf{T} = \sum_{j=1}^{k} V^j \otimes V^j \otimes (W^j D_\lambda), \tag{9.12}$$

where D_λ is the diagonal matrix whose entries consist of the mixing weights $\{\lambda_j\}$ and, for every $j \in [k]$, $W^j = (e^{-2\pi i \langle \mu_j, v^{(1)} \rangle}, e^{-2\pi i \langle \mu_j, v^{(2)} \rangle}, e^{-2\pi i \langle \mu_j, v^{(3)} \rangle})$ and $V^j = (e^{-2\pi i \langle \mu_j, \omega^{(1)} \rangle}, \cdots, e^{-2\pi i \langle \mu_j, \omega^{(m')} \rangle})$. Let $V \in \mathbb{R}^{m' \times k}$ denote the matrix whose j-th column is V^j.

Note that by our choice of r, R and triangle inequality, we have that $\|\omega^{(a)} + \omega^{(b)} + v^{(i)}\|_2 \leq r + 2R = 1 - \frac{c - \overline{\gamma}}{2c} < 1$ for any entry index a, b, i. So if $\{u_{a,b,i}\}$ are the numbers obtained from an η-approximate, $(m+3)$-query OTF oracle as in Algorithm 44, and $\widetilde{\mathbf{T}}$ is constructed as in Step 7 of TensorResolve, then by Lemma 9.4.5 we have that

$$|\mathbf{T}_{a,b,i} - \widetilde{\mathbf{T}}_{a,b,i}| \leq \frac{\eta}{\widehat{A}[1 - \frac{c - \overline{\gamma}}{2c}]} \leq \eta \cdot \left(\frac{c - \overline{\gamma}}{2c}\right)^2,$$

where the last step follows by Fact 9.3.4.

The following is a consequence of the stability of Jennrich's algorithm.

Lemma 9.4.15. *[e.g. [HK15], Lemma 3.5] For any $\varepsilon, \delta > 0$, suppose $|\mathbf{T}_{a,b,i} - \widetilde{\mathbf{T}}_{a,b,i}| \leq \eta'$ for $\eta' \triangleq O\left(\frac{(c - \overline{\gamma})\delta \Delta \lambda_{\min}^2}{k^{5/2} m^{3/2} \kappa(V)^5} \cdot \varepsilon\right)$, and let $\widehat{V} = \text{Jennrich}(\widetilde{\mathbf{T}})$ (Algorithm 46). Then with probability at least $1 - \delta$ over the randomness of $v^{(1)}$, there exists permutation matrix Π such that $\|\widehat{V} - V\Pi\|_F \leq \varepsilon$ for all $j \in [k]$.*

The setting of parameters in [HK15] is slightly different from ours, so we provide a self-contained proof of Lemma 9.4.15 in Appendix 9.10.

We will also need the following basic lemma about the stability of solving for $\widehat{\lambda}$ in Step 12 in TensorResolve.

Lemma 9.4.16. *For any $\varepsilon, \varepsilon' > 0$, if $\lambda \in \mathbb{R}^k$ satisfies $V\lambda = b$ for some $V \in \mathbb{R}^{m' \times k}$ and $b \in \mathbb{R}^{m'}$, and furthermore $\widehat{V}, \widehat{b}$ satisfy $\|V - \widehat{V}\|_2 \leq \varepsilon$ and $\|b - \widehat{b}\|_2 \leq \varepsilon'$, then $\widehat{\lambda} \triangleq \text{argmin}_{\widehat{\lambda}} \|\widehat{V}\widehat{\lambda} - \widehat{b}\|_2$*

satisfies $\|\lambda - \widehat{\lambda}\|_2 \le \frac{2\varepsilon\|\lambda\|_2 + 2\varepsilon'}{\sigma_{\min}(V) - \varepsilon}$.

Proof. Note that

$$\|\widehat{V}\lambda - \widehat{b}\|_2 \le \|(\widehat{V} - V)\lambda\|_2 + \|\widehat{b} - b\|_2 \le \varepsilon\|\lambda\|_2 + \varepsilon'.$$

By triangle inequality and definition of $\widehat{\lambda}$, $\|\widehat{V}(\widehat{\lambda} - \lambda)\|_2 \le 2\varepsilon\|\lambda\|_2 + 2\varepsilon'$, so $\|\widehat{\lambda} - \lambda\|_2 \le \frac{2\varepsilon\|\lambda\|_2 + 2\varepsilon'}{\sigma_{\min}(\widehat{V})}$. The lemma follows because $\sigma_{\min}(V') \ge \sigma_{\min}(V) - \varepsilon$. □

It remains to show the following condition number bound.

Lemma 9.4.17. *For any* $\delta > 0$, *if* $m = \Theta\left(\frac{k^2 \log(/\delta)}{(\Delta - \overline{\gamma}) \wedge 1}\right)$, *then* $\kappa(V) \le O\left(k \vee \frac{k}{\sqrt{\Delta - \overline{\overline{\gamma}}}}\right)$ *and* $\sigma_{\min}(V) \ge \Omega\left(k^2 \log(/\delta)\right)$ *with probability at least* $1 - \delta$.

Proof. Let $V^* \in \mathbb{R}^{m \times k}$ denote the submatrix given by the first m rows of V. We will need the following basic lemma from [HK15] relating the condition number of V^* to that of V:

Lemma 9.4.18 ([HK15], Lemma 3.8)**.** $\kappa(V) \le \sqrt{2k} \cdot \kappa(V^*)$.

The primary technical component of this section is to upper bound $\kappa(V^*)$. First, note that given any $\lambda \in \mathbb{C}^{k-1}$, we have that

$$\lambda^\dagger V^{*\dagger} V^* \lambda = \sum_{i=1}^m |\langle \lambda, V_i^* \rangle|^2 = \sum_{i=1}^m \left| \sum_{j=1}^k \lambda_j e^{-2\pi i \langle \mu_j, \omega^{(i)} \rangle} \right|^2.$$

As each $\omega^{(i)}$ is an independent draw from the uniform distribution over $\mathbb{S}^1$, we have that

$$\mathbb{E}_{\omega^{(1)}, \ldots, \omega^{(m)}}[\lambda^\dagger V^{*\dagger} V^* \lambda] = m \int_{B^2(R)} \left| \sum_{j=1}^k \lambda_j e^{-2\pi i \langle \mu_j, \omega \rangle} \right|^2 \mathrm{d}\psi(\omega),$$

where $\mathrm{d}\psi(\omega)$ is the uniform measure over $B^2(R)$. Furthermore, for any $\omega \in B^2(R)$ and $i \in [m]$, we have that

$$0 \le |\langle \lambda, V_i^* \rangle|^2 \le \|\lambda\|_1^2 \le k \cdot \|\lambda\|_2^2. \tag{9.13}$$

So by matrix Hoeffding applied to the random variables $m \cdot {V^*}_1^\dagger V_1^*, \ldots, m \cdot {V^*}_m^\dagger V_m^*$, each of which is upper bounded in spectral norm by $m \cdot k$ based on (9.13), we conclude that

$$\Pr\left[\|V^{*\dagger}V^* - \mathbb{E}_{\omega^{(1)},\ldots,\omega^{(m)}}[V^{*\dagger}V^*]\|_2 > \sqrt{m}kt\right] \le k \cdot e^{-\Omega(t^2)} \ \forall\ t > 0. \tag{9.14}$$

Lemma 9.4.20 below allows us to bound the quadratic form given by the expectation term evaluated at any λ. Taking $t = O(\sqrt{\log k/\delta})$ and $m = \Theta\left(\frac{k^2\log(k/\delta)}{(\Delta-\overline{\gamma})\wedge 1}\right)$ in (9.14) and applying Lemma 9.4.20, we conclude that with probability at least $1-\delta$,

$$\Omega(m)\cdot\{(\Delta-\overline{\gamma})\wedge 1\}\cdot\|\lambda\|_2^2 \le \lambda^\dagger V^{*\dagger}V^*\lambda \le O(m)\cdot(k+\{(\Delta-\overline{\gamma})\wedge 1\})\cdot\|\lambda\|_2^2,$$

from which it follows that with this probability, $\kappa(V^*) \le O\left(\frac{k}{(\Delta-\overline{\gamma})\wedge 1}\right)^{1/2}$, from which the lemma follows by Lemma 9.4.18. □

It remains to show Lemma 9.4.20 below, the key technical ingredient of this section. We will require the following special case of a result of [Gon18], which essentially follows from results of [CCLM17, HV+96]. This can be thought of as the high-dimensional generalization of the well-known Beurling-Selberg minorant (see, e.g., [Vaa85] for a discussion of the one-dimensional case).

Theorem 9.4.19 ([Gon18], Theorem 1)**.** *For any $d \in \mathbb{N}$ and $\frac{j_{d/2-1,1}}{\pi} < r < \frac{j_{d/2,1}}{\pi}$, there exists a function $M \in L^1(\mathbb{R}^d)$ whose Fourier transform is supported in $B^d(r)$, and which satisfies $M(x) \le \mathbb{1}[x \in B^d(1)]$ for all $x \in \mathbb{R}^d$ and $\widehat{M}[0] = \frac{(2/r)^d}{|\mathbb{S}^{d-1}|}\cdot\frac{C(d,r)}{1+C(d,r)/d}$, where $|\mathbb{S}^{d-1}|$ denotes the surface area of $\mathbb{S}^{d-1}$ and $C(r,d) \triangleq -\frac{\pi r J_{d/2-1}(\pi r)}{J_{d/2}(\pi r)} > 0$.*

Lemma 9.4.20.

$$\Omega((\Delta-\overline{\gamma})\wedge 1)\cdot\|\lambda\|_2^2 \le \int_{B^2(R)}\left|\sum_{j=1}^k \lambda_j e^{-2\pi i\langle\mu_j,\omega\rangle}\right|^2 \mathrm{d}\psi(\omega) \le k\|\lambda\|_2^2 \tag{9.15}$$

where $\mathrm{d}\psi(\omega)$ denotes the uniform probability measure over $B^2(R)$ for $R = \frac{\overline{\gamma}}{2\Delta}$.[2]

Proof. The upper bound follows by (9.13). We now show the lower bound. By Theorem 9.4.19 applied to $d = 2$, for any $\overline{\gamma}/2 < r < \frac{j_{1,1}}{\pi}$ there is a function M which mi-

[2] In fact, one can improve the upper bound in (9.15) by using a suitable majorant for the indicator of the ball, but because we are only after polynomial time and sample complexity, this is not needed.

norizes the indicator function of $B^2(1)$ and has Fourier transform supported in $B^2(r)$. Take $r = \{\Delta R \wedge \frac{\overline{\gamma}/2 + j_{1,1}/\pi}{2}\}$ which satisfies $\overline{\gamma}/2 < r < \frac{j_{1,1}}{\pi}$. This implies that the function $M'(\omega) \triangleq \frac{1}{\pi R^2} \cdot \frac{1}{R} \cdot M(\omega/R)$ minorizes the density $\psi(\omega)$, has Fourier transform supported in $B^2(r) \subseteq B^2(\Delta)$, and satisfies

$$\widehat{M'}[0] = \frac{1}{\pi R^2}\frac{(2/r)^2}{|\mathbb{S}^1|} \cdot \frac{C(2,r)}{1 + C(2,r)/2} = \frac{4C(2,r)}{\pi^2 r^3 R^2 \cdot (2 + C(2,r))} \geq \frac{r - \overline{\gamma}/2}{R^2} \geq 4r - 2\overline{\gamma}, \quad (9.16)$$

where in the last step we used that $R < 1/2$. We can lower bound (9.15) by

$$\begin{aligned}\int \left| \sum_{j=1}^{k} \lambda_j e^{-2\pi i \langle \mu_j, \omega \rangle} \right|^2 \cdot M'(\omega) \mathrm{d}\omega &= \sum_{j,j'=1}^{k} \lambda_j \lambda_{j'}^{\dagger} \int e^{-2\pi i \langle \mu_j - \mu_{j'}, \omega \rangle} \cdot M'(\omega) \mathrm{d}\omega \\ &= \sum_{j,j'=1}^{k} \lambda_j \lambda_{j'}^{\dagger} \widehat{M'}[\mu_j - \mu_{j'}] \geq (4r - 2\overline{\gamma}) \|\lambda\|_2^2,\end{aligned}$$

where the last step follows by (9.16) and the fact that $\widehat{M'}[\mu_j - \mu_{j'}] = 0$ for all $j \neq j'$. The lemma follows from noting that $4r - 2\overline{\gamma} > \{2\overline{\gamma}(\Delta/c - 1)\} \wedge \left\{\frac{2j_{1,1}}{\pi} - \overline{\gamma}\right\} \geq O(\Delta - \overline{\gamma} \wedge 1)$. □

Putting everything together, we have the following guarantee:

Lemma 9.4.21. *Let ρ be a Δ-separated superposition of k Airy disks. For any $\varepsilon_1, \varepsilon_2, \delta > 0$, let*

$$m = \Theta\left(\frac{k^2 \log(k/\delta)}{(\Delta - \overline{\gamma}) \wedge 1}\right) \qquad \text{and} \qquad \eta = O\left(\frac{4\Delta^3 \delta \lambda_{\min}^2}{(\Delta - \overline{\gamma}) k^{5/2} m^{3/2} \kappa(V)^5} \cdot \varepsilon_1\right). \quad (9.17)$$

Without loss of generality suppose $\varepsilon_1 < 1/6$. Then the output $(\lambda_1^*, \lambda_2^*, \mu_1^*, \mu_2^*)$ of TensorResolve*, given $\varepsilon_1, \varepsilon_2, \delta$ and access to an η-approximate, m-query OTF oracle $\mathcal{O}$ for ρ, satisfies*

$$\|\mu_i - \mu_{\tau(i)}^*\|_2 \leq \varepsilon_1 \;\; \textit{and} \;\; |\lambda_i - \lambda_{\tau(i)}| \leq \varepsilon_2$$

for some permutation τ with probability at least $1 - \delta$. Furthermore, the runtime of LearnAiryDisks *is polynomial in k, the number of OTF oracle queries, and the time it takes to make those queries.*

Proof. By Lemma 9.4.15, if we take $m = \Theta\left(\frac{k^2 \log(k/\delta)}{(\Delta - \overline{\gamma}) \wedge 1}\right)$ and $\eta' = O\left(\frac{(c - \overline{\gamma}) \delta \Delta \lambda_{\min}^2}{k^{5/2} m^{3/2} \kappa(V)^5} \cdot \varepsilon_1\right)$, then the output $\widehat{V}$ of JENNRICH($\widetilde{\mathbf{T}}$) satisfies $\|\widehat{V} - V\Pi\|_F \le \varepsilon_1$ for some permutation matrix Π. Assume without loss of generality that $\Pi = \mathrm{Id}$. Then we get that for all $j \in [k]$ and $\ell \in [m']$,

$$|\widehat{V}_{\ell,j} - V_{\ell,j}| = \left| e^{-2\pi i \langle \widehat{\mu}_j - \mu_j, \omega^{(\ell)} \rangle} - 1 \right| \le \varepsilon_1,$$

and because of the elementary inequality $|e^{-2\pi i x} - 1| \ge 2|x|$ for any $|x| \le 2/3$ and the fact that

$$\langle \widehat{\mu}_j - \mu_j, \omega^{(\ell)} \rangle \le \|\widehat{\mu}_j - \mu_j\|_2 \|\omega^{(\ell)}\|_2 \le 2\mathcal{R} \le 2/3, \tag{9.18}$$

we conclude that $|\langle \widehat{\mu}_j - \mu_j, \omega^{(\ell)} \rangle| \le \varepsilon_1/2$ for all $j \in [k]$, $\ell \in [m']$. In particular, this holds for all $\ell = m+1$ and $\ell = m+2$, so $\|\widehat{\mu}_j - \mu_j\|_\infty \le \varepsilon_1$. By dividing ε_1 by $\sqrt{2}$ and absorbing constants, we get that the estimates $\{\widehat{\mu}_j\}$ for the centers are ε_1-close to the true centers.

To show that the mixing weights are ε_2-close to the true mixing weights, we can apply Lemma 9.4.16 to conclude that

$$\|\lambda - \widehat{\lambda}\|_2 \le O\left(\frac{\varepsilon_1 + \eta'}{k^2 \log(1/\delta) - \varepsilon_1}\right) = O\left(\frac{\varepsilon_1}{k^2 \log(k/\delta)}\right),$$

so, possibly by modifying ε_1 to be $\frac{\varepsilon_2}{k^2 \log(1/\delta)}$, we get that the estimates $\{\widehat{\lambda}_j\}$ for the mixing weights are ε_2-close to the true mixing weights. □

Note that we can also amplify the success probability of TENSORRESOLVE by running SELECT from Section 9.4.2, but we do not belabor this point here.

9.4.4 Approximating the Optical Transfer Function

In this section, we show that the following algorithm DFT is a valid implementation of an approximate OTF oracle. We begin by showing that when the samples have granularity $\varsigma = 0$, DFT can achieve arbitrarily small error with polynomially many samples.

Lemma 9.4.22. *For any $0 < \beta < 1$, $\eta > 0$, and frequencies $\omega_1, ..., \omega_m \in \mathbb{R}^2$, DFT($\{\omega_i\}_{i\in[m]}$) draws $N = O(\log(m/\beta)/\eta^2)$ samples and in time $T = O(N \cdot m)$ outputs numbers $u_1, ..., u_m$ for which $|u_j - \widehat{\rho}[\omega_j]| \le \eta$.*

Proof. By a union bound, it suffices to show that for any single $j \in [m]$, $|u_j - \widehat{\rho}[\omega_j]| \le \eta$

```
Algorithm 45: DFT($\eta, \rho, \beta, \{\omega_i\}$)
  Input: Error tolerance $\eta > 0$, sample access to $\rho$, confidence parameter $\beta > 0$,
         frequencies $\omega_1, ..., \omega_m$
  Output: With probability at least $1 - \beta$, numbers $u_1, ..., u_m$ such that for each
          $j \in [m]$, $|u_j - \widehat{\rho}[\omega_j]| \le \eta$
  $N \leftarrow O(\log(m/\beta)/\eta^2)$. Draw samples $\mathbf{x}_1, ..., \mathbf{x}_N$ from $\rho$. For each $j \in [m]$, compute
    the average $u_j \leftarrow \frac{1}{N}\sum_{i=1}^N \cos(2\pi \cdot \langle \omega_j, \mathbf{x}_i \rangle)$. return $u_1, ..., u_m$.
```

with probability at least $1 - \beta/m$. Note that

$$\mathbb{E}[u_j] = \mathbb{E}_{\mathbf{x}\sim\rho}[\cos(2\pi \cdot \langle \omega_j, \mathbf{x}\rangle)] = \mathbb{E}\left[\operatorname{Re}\widehat{\rho}[\omega_j]\right] = \widehat{\rho}[\omega_j],$$

where the last step follows by the fact that $\widehat{\rho}$ is real-valued (by circular symmetry of A). Furthermore, the summands in $\sum_{i=1}^N \cos(2\pi \cdot \langle \omega_j, \mathbf{x}_i\rangle)$ are $[-1, 1]$-valued, so by Chernoff,

$$\Pr\left[|u_j - \mathbb{E}[u_j]| > \eta\right] \le \exp(-\Omega(N\eta^2)),$$

from which the lemma follows by our choice of N. □

We now show that for general granularity $\varsigma > 0$, the output of DFT still achieves error $\eta + O(\varsigma)$.

Corollary 9.4.23. *For any $0 < \beta < 1$, $\eta, \varsigma > 0$, and frequencies $\omega_1, ..., \omega_m \in \mathbb{R}^2$, if* DFT*($\{\omega_i\}_{i\in[m]}$) draws $N = O(\log(m/\beta)/\eta^2)$ samples of granularity ς, then in time $T = O(N \cdot m)$ it outputs numbers $u_1, ..., u_m$ for which $|u_j - \widehat{\rho}[\omega_j]| \le \eta + O(\varsigma \cdot \|\omega_j\|_2)$.*

Proof. Note that $\cos(\cdot)$ is α-Lipschitz for some $\alpha < 3/4$. This implies that for any $\omega \in \mathbb{R}^2$, the function $\mathbf{x} \mapsto \cos(2\pi\langle \mathbf{x}, \omega\rangle)$ is at most $O(\|\omega\|_2)$-Lipschitz with respect to ℓ_2.

Take any collection of 0-granular samples $\mathbf{x}'_1, ..., \mathbf{x}'_N$ for which the averages $u'_1, ..., u'_m$ computed by DFT would be η-accurate. If DFT were instead passed ς-granular samples $\mathbf{x}_1, ..., \mathbf{x}_N$ for which $\|\mathbf{x}'_i - \mathbf{x}_i\|_2 \le \varsigma$ for each $i \in [N]$, then by triangle inequality, the averages $u_1, ..., u_m$ computed by DFT with these samples would satisfy $|u_j - u'_j| \le \eta + O(\varsigma \cdot \|\omega_j\|_2)$ for each $j \in [m]$, as claimed. □

Finally, with Lemma 9.4.14 and Lemma 9.4.22, we can complete the proof of Theorem 9.4.1.

Proof of Theorem 9.4.1. By Lemma 9.4.14, it suffices to produce an η-approximate, m-query OTF oracle for η defined in (9.11) and $m = O(\log 1/\delta)$. By Corollary 9.4.23, this can be done using

$$\log(m/\delta)/\eta^2 = \widetilde{O}\left(\log(1/\delta) \cdot \mathrm{poly}(1/\lambda_{\min}, 1/\varepsilon_1, 1/\varepsilon_2, (4k/\Delta)^{k^2})\right)$$

samples of granularity $\eta/2$ with probability at least $1-\delta$. Theorem 9.4.1 then follows by a union bound over the failure probabilities of LearnAiryDisks and DFT, and replacing 2δ with δ and absorbing constant factors. Finally, note that the dependence on $\mathcal{R}$ follows by the discussion at the end of Section 9.4.1. □

Proof of Theorem 9.4.2. By Lemma 9.4.21, it suffices to produce an η-approximate, m-query OTF oracle for η defined in (9.17) and $m = \Theta\left(\frac{k^2}{(\Delta-\overline{\gamma})\wedge 1}\right)$. By Corollary 9.4.23, this can be done with probability $9/10$ using

$$\log(10m)/\eta^2 = \widetilde{O}\left(\mathrm{poly}(k, 1/\Delta, 1/\lambda_{\min}, 1/\varepsilon_1, 1/\varepsilon_2, k, (\Delta - \overline{\gamma}) \wedge 1)\right)$$

samples of granularity $\eta/2$ with probability at least $1-\delta$. Theorem 9.4.1 then follows by a union bound over the failure probabilities of TensorResolveCorrect and DFT. As in the proof of Theorem 9.4.1, the dependence on $\mathcal{R}$ follows by the discussion at the end of Section 9.4.1. □

9.5 Information Theoretic Lower Bound

In this section we will exhibit two superpositions of Airy disks, both with minimum separation below the diffraction limit, which are close in statistical distance. Let ρ and ρ' respectively have mixing weights $\{\lambda_i\}$ and $\{\lambda'_i\}$, and centers $\{\mu_i\}$ and $\{\mu'_i\}$, where for each i, $\mu_i \triangleq (a_i, b_i)$ and $\mu'_i \triangleq (a'_i, b'_i)$ for some $a_i, a'_i, b_i, b'_i \in \mathbb{R}$. Concretely,

$$\rho(\mathbf{x}) = \sum_{i=1}^{\lceil k/2 \rceil} \lambda_i \cdot A\left(\frac{\|\mathbf{x}-\mu_i\|}{\sigma}\right) \quad \text{and} \quad \rho'(\mathbf{x}) = \sum_{i=1}^{\lfloor k/2 \rfloor} \lambda'_i \cdot A\left(\frac{\|\mathbf{x}-\mu'_i\|}{\sigma}\right)$$

for some $0 < \sigma < 1$. Note that under this setting of parameters, the Abbe limit corresponds to separation $\pi\sigma$.

Theorem 9.5.1. *Let $\underline{\gamma} \triangleq \sqrt{4/3}$. There exists a choice of $\{\mu_i\}$, $\{\mu'_i\}$, $\{\lambda_i\}$, $\{\lambda'_i\}$ such that the minimum separation among centers of ρ and among centers of ρ' is $\Delta = (1-\varepsilon)\underline{\gamma}\pi\sigma$, and $d_{TV}(\rho, \rho') \le \exp(-\Omega(\varepsilon\sqrt{k}))$.*

We first bound $\|\rho - \rho'\|_{L^2}$. By Plancherel's,

$$\begin{aligned}
\|\rho - \rho'\|^2_{L^2} &= \|\widehat{\rho} - \widehat{\rho'}\|^2_{L^2} \\
&= \sigma^2 \int_{\mathbb{R}^2} \widehat{A}[\sigma\omega]^2 \left|\sum_i \lambda_i e^{-2\pi i\langle\mu_i,\omega\rangle} - \sum_i \lambda'_i e^{-2\pi i\langle\mu'_i,\omega\rangle}\right|^2 d\omega \\
&\le \sigma^2 \int_{B_{1/\pi\sigma}(0)} (1 - \|\pi\sigma\omega\|)^2 \cdot \left|\sum_i \lambda_i e^{-2\pi i\langle\mu_i,\omega\rangle} - \sum_i \lambda'_i e^{-2\pi i\langle\mu'_i,\omega\rangle}\right|^2 d\omega \quad (9.19)
\end{aligned}$$

where the inequality follows by the elementary bound

$$\frac{2}{\pi}(\arccos(\pi r) - \pi r\sqrt{1 - \pi^2 r^2}) \le 1 - \pi r.$$

Recall now the construction in Lemma 9.2.1 (see the beginning of Section 9.2). As the entries of the vector u constructed in Lemma 9.2.1 satisfy $\mathrm{sgn}(u_{j_1,j_2}) = (-1)^{j_1,j_2}$, let I_0 (resp. I_1) denote the elements $\mathbf{j} = (j_1, j_2) \in \mathcal{J} \times \mathcal{J}$ for which $j_1 + j_2$ is even (resp. odd), and for every $\mathbf{j} \in I_0$ (resp. $\mathbf{j} \in I_1$), define $\lambda_{\mathbf{j}}$ and $\mu_{\mathbf{j}}$ (resp. $\lambda'_{\mathbf{j}}$ and $\mu'_{\mathbf{j}}$) by $u_{\mathbf{j}}$ and $\left(\frac{j_1}{m}, \frac{\sqrt{3}j_2}{m}\right)$. This construction is illustrated for $k = 25$ in Figure 9-5. By design, $\{\lambda_{\mathbf{j}}\}_{\mathbf{j}\in I_0}$ and $\{\lambda'_{\mathbf{j}}\}_{\mathbf{j}\in I_1}$ consist solely of nonnegative scalars and respectively sum to 1, so ρ, ρ' are valid superpositions of Airy disks. Furthermore, by design,

$$\sum u_{j_1,j_2} e^{-2\pi i\cdot(j_1x_1+\sqrt{3}j_2x_2)/m} = \sum_{\mathbf{j}\in I_0} \lambda_{\mathbf{j}} e^{-2\pi i\cdot\langle\mu_{\mathbf{j}},x\rangle} - \sum_{\mathbf{j}\in I_1} \lambda'_{\mathbf{j}} e^{-2\pi i\cdot\langle\mu'_{\mathbf{j}},x\rangle},$$

so we may now bound (9.19) to get the desired L_2 bound

$$\|\rho - \rho'\|^2_{L^2} \le \exp(-\Omega(\varepsilon\sqrt{k}))\sigma^2 \int_{B_{1/\pi\sigma}(0)} (1 - \|\pi\sigma\omega\|)^2 = \exp(-\Omega(\varepsilon\sqrt{k}))/6\pi = \exp(-\Omega(\varepsilon\sqrt{k})).$$

We are now ready to show that $d_{\mathrm{TV}}(\rho,\rho')$ is small. The following is a generic L^1 bound for functions whose univariate restrictions have bounded L^2 mass, whose derivatives inside some region Ω are bounded, and which decay sufficiently quickly outside of Ω.

Lemma 9.5.2. *Suppose for an $f \in L^1(\mathbb{R}^2)$, there exists some $T \geq 0$ such that for $\Omega = [-T,T]^2$ the following are satisfied:*

1. *For all $y \in [-T,T]$, $\max_{x\in[-T,T]} |f'(x,y)| \leq C$,*
2. *$\int_{\Omega^c} |f| \leq \eta$,*
3. *$f(-T,y) \leq \delta$ for all $y \in [-T,T]$.*

Then we have that

$$\|f\|_{L^1} \leq (2T)^{5/3} \cdot (3C\|f\|_{L^2}^2 + 2T\delta^3)^{1/3} + \eta.$$

Proof. By the triangle inequality and condition 3, it is enough to verify that

$$\int_\Omega |f| \leq (2T)^{5/3} \cdot (3C\|f\|_{L^2}^2 + 2T\delta^3)^{1/3}.$$

Note that for a fixed $y \in [-T,T]$, we have by the fundamental theorem of calculus and conditions 2 and 3 that for any $x \in [-T,T]$,

$$\frac{1}{3}|f(x,y)^3| \leq \frac{1}{3}|f(-T,y)^3| + \left(\int_{-T}^{x} f(t,y)^2 dt\right) \cdot \max_{t\in[-T,x]} |f'(t,y)| \leq \frac{1}{3}\delta^3 + C\int_{-T}^{T} f(t,y)^2 dt.$$

Define $g(y) \triangleq \int_{-T}^{T} f(t,y)^2 dt$ and note that $\int_{-T}^{T} g(y)dy \leq \|f\|_{L^2}^2$. Then

$$\begin{aligned}
\int_\Omega |f| &= \int_{-T}^{T}\int_{-T}^{T} |f(x,y)| dx\, dy \\
&\leq \int_{-T}^{T}\int_{-T}^{T} (3C \cdot g(y) + \delta^3)^{1/3} dx\, dy \\
&= 2T\int_{-T}^{T} (3C \cdot g(y) + \delta^3)^{1/3} dy \\
&\leq (2T)^2 \left(\int_{-T}^{T} \frac{1}{2T}(3C \cdot g(y) + \delta^3) dy\right)^{1/3}
\end{aligned}$$

$$= (2T)^{5/3}\left(\int_{-T}^{T}(3C\cdot g(y)+\delta^3)dy\right)^{1/3}$$
$$\leq (2T)^{5/3}\cdot(3C\|f\|_{L^2}^2+2T\delta^3)^{1/3},$$

where the penultimate inequality follows from the measure-theoretic generalization of Jensen's inequality. □

We will show that for an appropriate choice of T, the function $f:\mathbb{R}^2\to\mathbb{R}$ given by

$$f(\mathbf{x})\triangleq\rho(\mathbf{x})-\rho'(\mathbf{x})=\sum_{\mathbf{j}\in J_0}\lambda_{\mathbf{j}}\cdot A\left(\frac{\|\mathbf{x}-\mu_{\mathbf{j}}\|}{\sigma}\right)-\sum_{\mathbf{j}}\lambda'_{\mathbf{j}}\cdot A\left(\frac{\|\mathbf{x}-\mu'_{\mathbf{j}}\|}{\sigma}\right) \tag{9.20}$$

satisfies the conditions of Lemma 9.5.2.

For $\mathbf{a}=(a_1,a_2)\in\mathbb{R}^2, y\in\mathbb{R}$, define $A_{\mathbf{a},y}:\mathbb{R}\to\mathbb{R}$ by

$$A_{\mathbf{a},y}(x)=A\left(\frac{\sqrt{(x-a_1)^2+(y-a_2)^2}}{\sigma}\right).$$

As we will see below, by triangle inequality it will suffice to verify certain properties of $A_{\mathbf{a},y}$.

Lemma 9.5.3. *For any $y\in\mathbb{R}$, $\max_{x\in\mathbb{R}}|f'(x,y)|=O(1/\sigma)$.*

Proof. By triangle inequality, it suffices to show that for any $a,y\in\mathbb{R}$, $\max_{x\in\mathbb{R}}\left|\frac{\partial A_{\mathbf{a},y}(x)}{\partial x}\right| = O(1/\sigma)$. Of course we may as well assume $a_1=0$, in which case by a change of variable in x, we have that

$$\begin{aligned}
&\max_{x\in\mathbb{R}}\left|\frac{\partial A_{\mathbf{a},y}(x)}{\partial x}\right|\\
&\quad=\frac{1}{\pi\sigma}\max_x\left|\frac{\partial}{\partial x}\frac{J_1(\sqrt{x^2+((y-a_2)/\sigma)^2})^2}{x^2+((y-a_2)/\sigma)^2}\right|\\
&\quad=\frac{1}{\pi\sigma}\max_x\left|\frac{2xJ_1(\sqrt{x^2+((y-a_2)/\sigma)^2})\cdot J_2(\sqrt{x^2+((y-a_2)/\sigma)^2})}{(x^2+((y-a_2)/\sigma)^2)^{3/2}}\right|\\
&\quad\leq\frac{1}{\pi\sigma}\max_x\left|\frac{2\sqrt{x^2+((y-a_2)/\sigma)^2}J_1(\sqrt{x^2+((y-a_2)/\sigma)^2})\cdot J_2(\sqrt{x^2+((y-a_2)/\sigma)^2})}{(x^2+((y-a_2)/\sigma)^2)^{3/2}}\right|\\
&\quad=\frac{1}{\pi\sigma}\max_z\left|\frac{2J_1(z)J_2(z)}{z^2}\right|
\end{aligned}$$

$$= \frac{1}{\pi\sigma} \max_z \left| \frac{\partial A(z)}{\partial z} \right|$$
$$\leq O(1/\sigma)$$

as claimed. □

Lemma 9.5.4. *For $T > \Delta(\sqrt{k}-1)/4$, we have that $f(-T, y) \leq \Omega\left((T/\sigma)^{-8/3}\right)$ for all $y \in [-T, T]$.*

Proof. By linearity, it suffices to show that for any $y \in [-T, T]$ and any $j_1, j_2 \in \mathcal{J}$, the claimed bound holds for $A_{\nu_{j_1,j_2},y}(-T)$. By Theorem 9.3.6, we know that

$$A_{\nu_{j_1,j_2},y}(-T) \leq \frac{1}{\pi} c_{28}^2 \cdot |r|^{-8/3},$$

where

$$r \triangleq \frac{1}{\sigma}\left(\left(-T - \frac{j_1}{m}\right)^2 + \left(y - \frac{\sqrt{3}j_2}{m}\right)^2\right)^{1/2} \geq \frac{-T - j_1/m}{\sigma} > -2T/\sigma,$$

where in the last step we used the fact that $j_1/m \leq \Delta(\sqrt{k}-1)/4 < T$. □

Lemma 9.5.5. *For $T > \Delta(\sqrt{k}-1)/2$, we have that $\int_{\Omega^c} |f| \leq O(T^{-2/3}\sigma^{8/3})$, where $\Omega = [-T, T]^2$.*

Proof. By linearity and the fact that $\|\mathbf{x} - (j_1/m, \sqrt{3}j_2/m)\|_2 \geq T - j_1/m \geq T/2$ for every $\mathbf{x} \notin \Omega$, it suffices to show that for any $j_1, j_2 \in \mathcal{J}$, the claimed bound holds for $\int_{B_0(T/2)^c} |A(x, y)| dx\, dy$. Expressing this as a polar integral, we have

$$\begin{aligned}\int_{\Omega^c} \left|A_{\nu_{j_1,j_2},y}(x)\right| dx\, dy &= \int_0^{2\pi} \int_{T/2}^{\infty} r \cdot |A(r/\sigma)| dr\, d\theta \\ &\leq 2 \cdot \int_{T/2}^{\infty} r \cdot \left(c_{28}^2 \cdot (r/\sigma)^{-8/3}\right) \\ &\leq O(T^{-2/3}\sigma^{8/3}),\end{aligned}$$

as desired. □

Proof of Theorem 9.5.1. Take $T = \Theta(\|f\|_{L^2}^{-1/5}) \geq \exp(-\Omega(\varepsilon\sqrt{k}))$ which we may assume without loss of generality, by scaling σ, Δ appropriately, is greater than $\Delta\sqrt{k}$. By Lemma 9.5.2 and Lemmas 9.5.3, 9.5.4, and 9.5.5, we have that for f defined by (9.20),

$$\begin{aligned}\int_{\mathbb{R}^2} |f| &\leq (2T)^{5/3} \cdot \left(O(1/\sigma) \cdot \|f\|_{L^2}^2 + O(T \cdot (T/\sigma)^{-8})\right)^{1/3} + O\left(T^{-2/3}\sigma^{8/3}\right) \\ &\leq O\left(\|f\|_{L^2}^{2/15}\sigma^{-1/3}\right) \\ &\leq O\left(\exp\left(-\Omega(\varepsilon\sqrt{k})\right)\sigma^{-1/3}\right),\end{aligned}$$

so as soon as $k \geq C\log(1/\sigma))$ for sufficiently large $C > 0$, we have that $d_{\mathrm{TV}}(\rho, \rho') \leq \exp(-\Omega(\varepsilon\sqrt{k}))$. □

9.6 Conclusion and Open Problems

We hope that our work will be a stepping-stone towards developing a rigorous theory of resolution limits in more sophisticated optical systems. The setting that we study, namely diffraction through a perfectly circular aperture under incoherent illumination, is arguably the most basic model one can study in Fourier optics. As a natural next step, one can ask whether the techniques developed in this chapter can be pushed to answer questions about the following more challenging setting:

Coherent illumination As described in Appendix 9.8, in the presence of light emanating from a single point source, the (complex-valud) amplitude of the electric field at a point P on the observation plane is proportional to $e^{i\omega} \cdot J_1(z/\sigma)/(z/\sigma)$, where $e^{i\omega}$ is some phase factor, z is the angular displacement of the point P from the optical axis, and σ is the spread parameter which depends on the wavelength of the light and the radius of the aperture. This means that the actual probability distribution over where on the observation plane a photon gets detected is proportional to the squared modulus of this, i.e. $J_1(z/\sigma)^2/(z/\sigma)^2$. Throughout this work, we worked under the assumption that in the presence of many point sources, the light emanating from the various point sources is *incoherent*. In other words, there is no interference introduced by the extra phase factors, and mathematically this translates to a probability distribution given by a *nonnegative* linear combination of the probability densities coming from the individual sources of light, and this is what gives rise to the mixture model we studied.

The *coherent* setting is quite different. Suppose that for point source j, the extra phase factor in the electric field at any given point is $e^{i\omega_j}$ for some complex number ω_j. Under coherent illumination from multiple point sources, it is the *electric field* which is a linear combination, namely of the electric fields associated to each indivdiual point source. This gives rise to the following natural probabilistic model:

Definition 9.6.1 (Coherent superpositions of Airy disks)**.** *A* coherent superposition of k Airy disks ρ *is a distribution over* $\mathbb{R}^2$ *specified by phases* $\omega_1, \dots, \omega_k \in C_3$*, relative intensities* $\lambda_1, \dots, \lambda_k \geq 0$ *summing to 1, centers* $\mu_1, \dots, \mu_k \in \mathbb{R}^2$*, and an* a priori *known "spread parameter"* $\sigma > 0$*. Its density at* x *is proportional to*

$$\rho(\mathbf{x}) \propto \left| \sum_{i=1}^{k} \lambda_i \cdot e^{i\omega_j} \cdot \frac{J_1(\|\mathbf{x} - \mu_i\|/\sigma)}{\|\mathbf{x} - \mu_i\|/\sigma} \right|^2 .$$

One can ask analogues of all of the questions considered in the present work for this probabilistic model. This seems to be both mathematically natural and a physically well-motivated setting which now departs from the mixture model setup usually studied within theoretical computer science.

9.7 Appendix: Related Work In the Sciences

In this section, we survey previous approaches to understanding diffraction limits in the optics literature, as well as recent practical works on the need and methodologies to rigorously assess claims of achieving super-resolution.

9.7.1 Previous Approaches in Optics

In this section we will survey the many previous attempts to rigorously understand diffraction limits in the optics literature. There, the focus has been squarely on the semiclassical detection model (SDM). After describing this line of work, we explain the ways in which it falls short.

The SDM was originally proposed by [Man59] and has been the *de facto* generative model in essentially all subsequent works on the statistical foundations of resolution. We note that there are some minor differences in the definition of our model and that of the SDM, which we will discuss formally in Appendix 9.8.3.

Arguably the first significant work to study the SDM was that of Helstrom [Hel64], who considered it from the perspective of parameter estimation and hypothesis testing, initiating the study of the following two problems which remarkably have almost exclusively occupied this line of work. For normalized point spread function $A(\cdot)$ and separation parameter d, define

$$\rho_0(\mathbf{x}) = A(\mathbf{x}), \quad \rho_1(\mathbf{x}) = \frac{1}{2}\cdot A(\mathbf{x}-\mu) + \frac{1}{2}\cdot A(\mathbf{x}+\mu), \quad \mu = \begin{cases} d/2 & D=1 \\ (0,d/2) & D=2. \end{cases} \tag{9.21}$$

Problem 1 (Parameter Estimation)**.** *Given samples from ρ_1, estimate d.*

Problem 2 (Hypothesis Testing)**.** *Suppose we know the parameter d, and we know that either $\rho = \rho_0$ or $\rho = \rho_1$. Given samples from ρ, decide whether $\rho = \rho_0$ or $\rho = \rho_1$.*

For Problem 1, Helstrom [Hel64, Hel69, Hel70] studied the maximum likelihood estimator and computed Cramer-Rao lower bounds for a host of point-spread functions including the Airy PSF, both for the SDM and for progressively more physically sophisticated (though less practically relevant) models. The conceptual insights and problem formulation of [Hel64] were refined, or often rediscovered, numerous times [TD79, BVDDD$^+$99, SM04, SM06, RWO06, Far66, CWO16], and the primary thrust of this line of work has been centered on Cramer-Rao-style calculations for assorted point-spread functions and, to a lesser extent, analysis of the optimization landscape of the log-likelihood from the perspective of singularity theory [VDB01, VdBDD01, BVDDD$^+$99, DD96].

For Problem 2, Helstrom [Hel64] computed the reliability of the likelihood ratio test for various PSFs, under a CLT appoximation to the log-likelihood ratio. Similar calculations for the log-likelihood ratio for other PSFs followed in [Har64, AH97, SM04, SM06, Far66].

We emphasize that, with the exception of [SM04, SM06], all works giving rigorous guarantees have made the assumption implicit in (9.21) that the two point sources defining ρ_1 are located at *known* points μ and $-\mu$ centered about the origin. [SM04, SM06] study Problems 1 and 2 when the locations of the point sources are unknown and study the (locally optimal) generalized likelihood ratio test.

With regards to applications, Problems 1 and 2 have gained popularity in optical astronomy [Fal67, Zmu03, FB12, Luc92a, Luc92b] as well as fluorescence microscopy [MCSF10, SS14b, DZM$^+$14, vDSM17]. Cramer-Rao bounds as a "modern" proxy for assessing the limits of imaging systems have gained such popularity with practicioners that a number of review articles and surveys on the topic have appeared in the recent single-molecule microscopy literature [SS14b, DZM$^+$14, CWO16], most of which focus on the related parameter estimation problem of *localization*, that is, estimating the location of a *single* test object given its noisy image.

One other interesting line of work has focused on the generalizations of Problems 1 and 2 to the quantum setting. Elaborating on this literature would take us too far afield, so we mention only the comprehensive recent survey [Tsa19] and the references therein.

9.7.2 Comparison with Our Approach

Most crucially, all works on the SDM focus exclusively on *two*-point resolution. In the context of hypothesis testing, as we note above, these works even assume the two points lie on the x-axis at the same *known* distance $d/2$ from the origin, with the exception of [SM04, SM06]. That such a strong assumption is made and such focus is placed on $k = 2$ is evidently not just for aesthetics. From the standpoint of hypothesis testing, as noted in [SM04, SM06], any deviation from this idealized model would induce a composite hypothesis testing problem, for which the (generalized) likelihood ratio test has no global optimality guarantees. In the context of parameter estimation, because of the focus on $k = 2$, the conclusion in the literature has repeatedly been that the classical resolution criteria (Abbe, Rayleigh, etc.) are not meaningful in a statistical sense, and that the only true limitation comes from the number of samples. We view this as one of the primary reasons that a result like Theorem 9.1.1 has gone overlooked for so long.

Another drawback of the literature is that because of the focus on Cramer-Rao bounds, which only provide guarantees for the maximum likelihood estimate in the infinite-sample limit, none of these works actually give non-asymptotic algorithmic guarantees. Additionally, Cramer-Rao bounds only apply to unbiased estimators, and to the best of our knowledge, the only paper that addresses biased estimators is [Tsa18], which only derives Bayesian Cramer-Rao bounds for the already well-studied setting of a mixture of two Gaussians. From a technical standpoint, another disadvantage of existing works is that they work either with the Gaussian point-spread function or invoke Taylor approximations of the Airy point-spread function. And because the log-likelihood here is analytically cumbersome, it is common to invoke a central limit theorem-style approximation.

One last shortcoming arises from the definition of the SDM itself (see Definition 9.8.2): it models photon detection as a Poisson process when in reality this need not be the case. As Goodman (Chapter 9.2 of [Goo15]) notes, "in most problems of real interest, however, the light wave incident on the photosurface has stochastic attributes ... For this reason, it is necessary to regard the Poisson distribution as a *conditional* probability distribution ... the statistics are in general *not* Poisson when the classical intensity has random fluctuations of its own." The increased generality of not assuming Poissonanity allows our model to smoothly handle such stochastic fluctuations.

9.7.3 Super-Resolution and the Practical Need to Understand Diffraction Limits

In the past half century, a host of techniques of increasing sophistication have been developed to shift or fundamentally surpass the diffraction limit. As these techniques change the underlying physical setup of the imaging system, they are not relevant to the theoretical setting we consider, though we believe that placing the classical setting of Fraunhofer diffraction on a rigorous statistical footing can pave the way towards better understanding notions of resolution in these modern techniques.Here we very briefly describe some these techniques, deferring to the comprehensive overviews on the matter found in [HG09, Hel07, Hel09, HBZ10, JSZB08, Lau12, LSM09, MW17, Ric07, WS15]. The earliest attempts at going below the diffraction limit involved modifying the aperture, e.g. via *apodization* as pioneered by Toraldo di Francia [DF52]. Among even more elementary approaches, an annular aperture can be used to distinguish a pair of points sources slightly better than a circular one, a fact that [MW17] notes was known even to Rayleigh. Other approaches for circumventing the diffraction limit include near-field optics [AN72, PDL84, Syn28], TIRF [Axe81, Tem81], confocal microscopy [Min61], two-lens techniques [HS92, HSLC94], structured illumination [Gus99], UV/X-ray/electron microscopy [BEZ+97, KJH95, Rus34].

Betzig, Hell, and Moerner were awarded the 2014 Nobel Prize in Chemistry for their pioneering work on super-resolution microscopy, which now includes technologies such as STED [HW94, KH99], RESOLFT [HK95, Hel04, BEH07], PALM [BPS+06], STORM [RBZ06], and FPALM [HGM06]. These fundamentally break the diffraction limit by leveraging the ability to switch fluorescent markers between a bright and a dark state via photophysical effects like stimulated emission and ground-state depletion. In light of such advancements, rigorously characterizing the resolving power of imaging systems remains a challenge of practical as much as theoretical interest. [DWSD15] revisited what resolution means given these new technologies technologies and proposed approaches for comparing resolution between different super-resolution methods. [HHP+16] pushed back on some claims of super-resolution in nonfluorescent microscopy, advocating for the Siemens star as an imaging benchmark and for the adoption of certain standards when documenting such claims. Sheppard [She17] was similarly motivated to clarify such claims and calculates the images of various test object geometries and suggests "these results can be used as a reference . . . to determine if super-resolution has indeed been attained."

9.8 Appendix: Physical Basis for Our Model

In this paper we focus on the idealized setting of Fraunhofer diffraction of incoherent illumination by a circular aperture, originally studied in the pioneering work of Airy [Air35]. In this section, we first give a brief overview of this setting in Appendix 9.8.1, deferring the details to any of a number of excellent expository texts on the subject [Ken08, Hec15, Goo05, Goo15, JW37, Fow89]. Then in Appendix 9.8.2, we demonstrate how our probabilistic model arises naturally from the preceeding setup. Finally, in Appendix 9.8.4, we catalogue the various resolution criteria that have appeared in the literature and instantiate them in our framework.

9.8.1 A Review of Fraunhofer Diffraction

Consider a scenario in which plane waves of monochromatic, incoherent light emanate from a far-away point source in the image plane, pass through a circular aperture, and form a diffraction pattern on a far-away observation plane. This is the standard setting of *Fraunhofer diffraction*. As depicted in Figure 9-1, the far-field assumption on the observation plane is captured in practice by placing a lens behind the aperture and placing the observation plane at the focal plane of the lens.

Under the Huygens-Fresnel-Kirchhoff theory, the aperture induces a diffraction pattern, a so-called *Airy disk*, on the observation plane because the secondary spherical wavelets emanating from different points of the aperture are off by phase factors. Concretely, suppose the plane waves are parallel to the optical axis, and take a point P on the observation plane at angular distance θ from the optical axis, and a point $\mathbf{u}$ on the circular aperture A, say of radius r. Letting $\mathbf{v}$ be the unit vector from the center of the aperture to P, we see that the propagation path of the wavelet from the center of the aperture to P and that of the wavelet from $\mathbf{u}$ to P differ in length by $\langle \mathbf{u}, \mathbf{v}\rangle$, corresponding to a phase delay of $\frac{2\pi}{\lambda}\langle \mathbf{u}, \mathbf{v}\rangle$ where λ is the wavelength of light. So by integrating over the contributions to the amplitude of the electric field at P by the points $\mathbf{u}$ in A, we conclude that the amplitude at P is

$$E = E_0 \int_A e^{2\pi i\cdot\langle \mathbf{u},\mathbf{v}\rangle/\lambda} d\mathbf{u},$$

where E_0 is, up to phase factors, a constant capturing the contribution to the field per unit area of the aperture. In other words, the amplitude at P is proportional to the 2D Fourier transform of the pupil function $F(\mathbf{u}) = \mathbb{1}[\mathbf{u} \in A]$ at frequency $\mathbf{v}/\lambda$. This can be computed explicitly as

$$E = 2\pi r^2 E_0 \cdot \frac{J_1(\kappa r \sin\theta)}{\kappa r \sin\theta},$$

where $\kappa \triangleq \frac{2\pi}{\lambda}$ is the wavenumber of the light. In particular, the intensity $I(\theta)$ of the diffraction pattern at P is the squared modulus of E. We conclude that

$$I(\theta) = I(0) \cdot \left(\frac{2J_1(\kappa r \sin\theta)}{\kappa r \sin\theta} \right)^2, \tag{9.22}$$

where $J_1(\cdot)$ is the Bessel function of the first kind. We will typically regard $I(\cdot)$ as a function $\mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ which takes in a point $(x, y) \in \mathbb{R}^2$ and outputs $I(\theta)$, where θ is the angular distance between (x, y) and the optical axis. The function $I(x, y)$ is the so-called *Airy point spread function.*

Remark 9.8.1. *In general, if the plane waves of the point source travel at an angle ψ to the optical axis, they will be focused not at the focal point but at some other point on the observation plane at an angular distance of ψ with respect to the optical axis. In this case the resulting Airy point spread function will be shifted to be centered at that point.*

9.8.2 Photon Statistics and Our Model

First suppose there is a single point source of light. In a sense which can be made rigorous via Feynman's path integral formalism (see e.g. Section 4.11 of [Hec15]), the intensity $I(x, y)$ of the diffraction pattern at a point (x, y) on the observation plane is proportional to the (infinitesimal) probability of detecting a photon at P. That is, the point spread function $I(x, y)$ can be identified with a probability density

$$\rho(x, y) \triangleq \frac{1}{Z} \cdot I(x, y), \quad \text{where} \quad Z \triangleq \int_{\mathbb{R}^2} I(x, y)\, \mathrm{d}x\mathrm{d}y$$

over the two-dimensional observation plane. Concretely, for any measurable subset S of the observation plane, if one were to count photons arriving over time and compute the fraction that land inside the region S, this fraction would tend towards $\int_S \rho(x, y)\, \mathrm{d}x\mathrm{d}y$.

In the presence of k *incoherent* point sources of light, the absence of interference means that the contributions from each point source to the intensities of the resulting diffraction pattern simply add. In other words, if $I_1(\cdot), ..., I_k(\cdot)$ are the corresponding point spread functions, which by Remark 9.8.1 are merely shifted versions of (9.22), the resulting probability density ρ over the observation plane is simply proportional to $\sum_{i=1}^k I_i(\cdot)$.

For every $i \in [k]$, let $Z_i \triangleq \int_{\mathbb{R}^2} I_i(x,y)\ \mathrm{d}x\mathrm{d}y$ be the normalizing constant for the i-th density $\rho_i(\cdot) \triangleq \frac{1}{Z_i} I_i(\cdot)$. Let $\lambda_i \triangleq \frac{Z_i}{\sum_{j=1}^m Z_j}$. Then we see that

$$\rho(x,y) = \sum_{i=1}^m \lambda_i \cdot \rho_i(x,y).$$

In the jargon of statistics, this is an example of a *mixture model*, i.e. a convex combination of structured distributions, and one can think of sampling from ρ by first sampling an index $i \in [m]$ with probability λ_i and then sampling a point (x,y) in the observation plane according to the probability density associated to the i-th point source. This brings us to the generative model that we study in this work, the definition of which we restate here for the reader's convenience.

Definition 9.3.1. *[Superpositions of Airy Disks] A* superposition of k Airy disks ρ *is a distribution over* $\mathbb{R}^2$ *specified by relative intensities* $\lambda_1, ..., \lambda_k \geq 0$ *summing to 1, centers* $\mu_1, ..., \mu_k \in \mathbb{R}^2$*, and an* a priori *known "spread parameter"* $\sigma > 0$*. Its density is given by*

$$\rho(\mathbf{x}) = \sum_{i=1}^k \lambda_i \cdot A_\sigma(\mathbf{x} - \mu_i) \quad \textit{for} \quad A_\sigma(\mathbf{z}) = \frac{1}{\pi\sigma^2}\left(\frac{J_1(\|\mathbf{z}\|_2/\sigma)}{\|\mathbf{z}\|_2/\sigma}\right)^2.$$

Note that the factor of $\frac{1}{\pi\sigma^2}$ *in the definition of* A_σ *is to ensure that* $A_\sigma(\cdot)$ *is a probability density.*

Also define

$$\Delta \triangleq \min_{i\neq j}\|\mu_i - \mu_j\|_2 \quad \textit{and} \quad \mathcal{R} \triangleq \max_{i\in[k]}\|\mu_i\|_2.$$

We now describe briefly how the parameters in Definition 9.3.1 translate to the setting of Fraunhofer diffraction by a circular aperture that we have outlined thus far. One should think of the spread parameter σ as $(\kappa r)^{-1}$. As σ in practice depends on known quantities pertaining to the underlying optical system, we assume henceforth that it is known *a priori*. The norm of the argument in $A_\sigma(\|\mathbf{x} - \mu_i\|_2)$ corresponds to the quantity $\sin\theta$, where θ is the angle of displacement between the line from the center of the aperture to the center μ_i of the i-th Airy disk, and the line between the center of the aperture and the point $\mathbf{x}$ on the observation plane. Lastly, by Remark 9.8.1, angular separation of ψ between two point sources translates to angular separation of ψ between the centers of their Airy disks on the observation plane. The parameters Δ and $\mathcal{R}$ can thus be interpreted respectively as the minimum angular separation among the point sources, and the maximum angular distance of any of the point sources to the optical axis.

9.8.3 Comparison to Semiclassical Detection Model

In this section we clarify the distinctions between the model we study and the semiclassical detection model. We begin by formally defining the latter.

Definition 9.8.2 (Semiclassical Detection Model)**.** *For $D = 1, 2$, let $S_1, ..., S_m \subset \mathbb{R}^D$ be disjoint subsets corresponding to different regions of a photon detector, and suppose the detector receives some number N' of photons, where $N' \sim Poi(N)$. We observe photon counts $N_1, ..., N_m$ corresponding to the number of photons that interact with each region of the detector, where for each $i \in [m]$,*

$$N_i \triangleq N_i' + \gamma_i, \quad N_i' \sim Poi(\lambda_i \cdot N), \gamma_i \sim \mathbb{N}(0, \sigma^2),$$

where $N_1', ..., N_m', \gamma_1, ..., \gamma_m$ are independent, γ_i represents white detector noise[3]*, and*

$$\lambda_i \triangleq \int_{S_i} \rho(\mathbf{x}) \, \mathrm{d}x,$$

where $\rho(\cdot)$, as in our model, is the idealized, normalized intensity profile of the optical signal.

To see how this relates to our model, first consider the idealized case where $\sigma = 0$ and that the different regions S_i of the detector form a partition of the entire ambient space. To get quantitative guarantees, existing works assume that each of these regions S_i is, e.g., a segment or box of fixed length ς. In this case, the semiclassical detection model is a special case of our model. Indeed, if one samples $\mathrm{Poi}(N)$ points from ρ and moves each of them by distance $O(\varsigma)$ to the center of the region S_i of the photon detector to which they

[3] While these white noise terms $\{\gamma_i\}$ were not present in [Man59, Hel64], they are considered in some later treatments of this model, so we include them here for completeness.

respectively belong, this collection of $O(\varsigma)$-granular samples from ρ is identical in information and distribution to a sample of photon counts $\{N_i\}$ from the semiclassical detection model.

Our model can also capture the case where the regions S_i only partition a *subset* of the ambient space $\mathbb{R}^D$. In this case, we only get access to samples from the density $\rho_{\text{trunc}}(\mathbf{x}) \propto \mathbb{1}[\mathbf{x} \in \cup S_i] \cdot \rho(\mathbf{x})$, but this has known Fourier transform, given up to a universal multiplicative factor by the convolution of $\widehat{\rho}$ with the indicator function of $\cup S_i$. So our techniques still apply in a straightforward fashion. In addition, by standard estimates on the tails of J_1, for $\cup S_i$ of radius polynomially large in the relevant parameters, with high probability none of the samples used by our algorithms will fall outside of $\cup S_i$ to begin with. For these reasons, we will not belabor this point in this work and will assume $\cup S_i = \mathbb{R}^D$ throughout.

Lastly, while our model does not incorporate white detector noise σ, we note that our algorithms can nevertheless handle the semiclassical detection model with $\sigma > 0$: from a set of photon counts $N_1, ..., N_m$, we can still estimate the Fourier transform of ρ to accuracy depending polynomially on N and inverse polynomially on σ and the sizes of the detector regions, so our techniques based on the matrix pencil method still apply.

9.8.4 A Menagerie of Diffraction Limits

In this section we give a precise characterization of the various limits that have appeared in the literature as candidates for the threshold at which resolution becomes impossible in diffraction-limited optical systems.

Abbe Limit The Abbe limit first arose in Abbe's studies [Abb73] of the following setup in microscopy: light illuminates an idealized object, namely an diffraction grating consisting of infinitely many closely spaced slits corresponding to the fine features of the object being imaged, and passes through the slits, behind which is an aperture stop placed in the back focal plane of the lens. Abbe observed that the angle at which the light gets diffracted by the slits increases as the grating gets finer, and he calculated the point at which the angle is too wide to enter the aperture. This threshold is now called the *Abbe limit*, and in the modern language of Fourier optics, the Abbe limit corresponds to the point at which the Fourier transform of the corresponding point spread function (see Fact 9.3.3) vanishes. In the remainder of this section, we will refer to the Abbe limit as τ.

Remark 9.8.3 (Scaling and Numerical Aperture)**.** *The argument z in $A_\sigma(z)$ corresponds to the more familiar-looking quantity*

$$z = \frac{2\pi}{\lambda} \cdot a \sin\theta, \tag{9.23}$$

where λ is the average wavelength of illumination, a is the radius of the aperture, and θ is the angle of observation.

As noted above, $\widehat{A}_\sigma[\omega]$ is only supported on ω for which $\|\omega\| \leq \frac{1}{\pi}$. Equating this threshold $\frac{1}{\pi}$ with $1/z$, where z is given by (9.23), and rearranging, we conclude that $\sin\theta = \frac{\lambda}{2a}$. We may write $\sin\theta$ as q/R for q the distance between the observation point and the optical axis and R the distance between the observation point and the center of the aperture. It then follows that $q = \frac{\lambda R}{2a} \approx \frac{\lambda}{2NA}$, where NA is the numerical aperture. This recovers the usual formulation of the Abbe limit*.*

In the literature on super-resolution microscopy, the Abbe limit is the definition of diffraction limit that is usually given. Indeed, Lauterbach notes in his survey [Lau12] that "Abbe is perhaps the one who is most often cited for the notion that the resolution in microscopes would always be limited to half the wavelength of blue light."

Rayleigh Criterion The Rayleigh criterion is the point at which the point spread function first vanishes. For $\sigma = 1$, this is precisely the smallest positive value of r for which $J_1(r) = 0$, which can be numerically computed to be $r \approx 3.83 \approx 1.22 \cdot \pi$. So for general σ, we conclude that the Rayleigh criterion is $\approx 1.22\tau$.

This is typically touted in standard references as the most common definition of resolution limit. Indeed, Weisenburger and Sandoghdar remark in their survey [WS15] that "Although Abbe's resolution criterion is more rigorous, a more commonly known formulation...is the Rayleigh criterion." Kenyon [Ken08] calls it the "standard definition of the limit of the resolving power of a lens system." In his classic text, Hecht [Hec15] refers to it as the "ideal theoretical angular resolution" Rayleigh himself [Ray79] emphasized however that "This rule is convenient on account of its simplicity and it is sufficiently accurate in view of the necessary uncertainty as to what exactly is meant by resolution." We refer to Appendix 9.9 for further quotations regarding the Rayleigh criterion.

Sparrow Criterion The Sparrow criterion, put forth in [Spa16], is the smallest Δ for which a superposition of two Δ-separated Airy disks becomes unimodal. Numerically, this threshold is $\approx 0.94\tau$.

The Sparrow limit is often cited as the most mathematically rigorous resolution criteria (in den Dekker and van de Bos' survey [DDVdB97], they even call it "the natural resolution limit that is due to diffraction...even a hypothetical perfect measurement instrument would not be able to detect a central dip in the composite intensity distribution, simply because there is no such dip anymore."). It is less relevant in practical settings as it requires perfect knowledge of the functional form of the point spread function. Again, we refer to Appendix 9.9 for further quotations regarding the Sparrow criterion.

Houston Criterion The Houston criterion is twice the radius at which the value of the density is half of its value at zero, i.e. the "full width at half maximum" (FWHM). This threshold is $\approx 1.03\tau$.

This measure is one of the most popular in practice where one does not have fine-grained knowledge of the point spread function, in particular because it can apply even when the point spread function in question does not fall exactly to zero, either due to noise or aberrations in the lens. In [DWSD15] where the authors explore alternative means of assessing resolution in light of new super-resolution microscopy technologies, they remark in their conclusion that "the best approach to compare between techniques is still to perform the simple and robust fitting of a Gaussian to a sub-resolution object and then to extract the FWHM."

Miscellaneous Additional Criteria The *Buxton limit* is nearly the same as Houtson, except it is the FWHM for the *amplitude* rather than the intensity, which yields a threshold of $\approx 1.46\tau$ [Bux37]. The *Schuster criterion* is defined to be twice the Rayleigh limit [Sch04], that is, two Airy disks are separated only when their central bands are disjoint, which yields a threshold of $\approx 2.44\tau$. The *Dawes limit*, which is $\approx 1.02\tau$, is a threshold proposed by

Dawes [Daw67]; its definition is purely empirical, as it was derived by direct observation by Dawes.

9.9 Appendix: Debate Over the Diffraction Limit: A Historical Overview

In this section, we catalogue quotations from the literature relevant to the challenge of identifying the right resolution criterion, as well as to the need to take noise into account when formulating such definitions.

9.9.1 Identifying a Criterion

Since its introduction, the Rayleigh criterion has repeatedly been both touted as a practically helpful proxy by which to roughly assess the resolving power of diffraction-limited imaging systems, and characterized as somewhat arbitrary.

Rayleigh himself in his original 1879 work [Ray79]:

> *"This rule is convenient on account of its simplicity and it is sufficiently accurate in view of the necessary uncertainty as to what exactly is meant by resolution."*

Williams [Wil50, p. 79] in 1950:

> *"Although with the development of registering microphotomers such as the Moll, dips much smaller than [the one exhibited by a superposition of two Airy disks at the Rayleigh limit] can be accurately measured, it is convenient for the purpose of comparison with gratings and echelons to keep to this standard."*

Born and Wolf [BW13, p. 418] in 1960:

> *"The conventional theory of resolving power...is appropriate to direct visual observations. With other methods of detection (e.g. photometric) the presence of two objects of much smaller angular separation than indicated by Rayleigh's criterion may often be revealed."*

Feynman [FLS11, Section 30-4] in his Lectures on Physics from 1964:

> *"...it seems a little pedantic to put such precision into the resolving power formula. This is because Rayleigh's criterion is a rough idea in the first place. It tells you where it begins to get very hard to tell whether the image was made by one or by two stars. Actually, if sufficiently careful measurements of the exact intensity distribution over the diffracted image spot can be made, the fact that two sources make the spot can be proved even if θ is less than λ/L."*

Hecht in his standard text [Hec15, p.431,492] from 1987:

> *"We can certainly do a bit better than this, but Rayleigh's criterion, however arbitrary, has the virtue of being particularly uncomplicated."*
> *"Lord Rayleigh's criterion for resolving two equal-irradiance overlapping slit images is well-accepted, even if somewhat arbitrarily in the present application."*

In fact, as early as 1904, Schuster [Sch04, p. 158] made the same point and on the same page advocated for an alternative criterion, corresponding to twice the separation posited by Rayleigh:

> *"There is something arbitrary in (the Rayleigh criterion) as the dip in intensity necessary to indicate resolution is a physiological phenomenon, and there are other forms of spectroscopic investigation besides that of eye observation... It would therefore have been better not to have called a double line "resolved" until the two images stand so far apart, that no portion of the centeral band of one overlaps the central band of the other, as this is a condition which applies equally to all methods of observation. This would diminish to one half the at present recognized definition of resolving power."*

Ever since, the question of identifying the "right" notion of a resolution criterion has been periodically revisited in the literature.

Ramsay et al. [RCK41, p. 26] in 1941, on this problem's theoretical and practical importance:

> *“Before the theory itself can be developed in full, and applied to the assignment of numerical values, it is necessary to consider the persistently vexing problem of criteria for a limit of resolution.”*

Three decades after Ramsay’s work, Thompson [Tho69, p. 171]:

> *“The specification of the quality of an optical image is still a major problem in the field of image evaluation and assessment. This statement is true even when considering purely incoherent image formation.”*

The Sparrow criterion is often regarded as the most mathematically rigorous resolution criterion.

Sparrow [Spa16, p. 80] in 1916 on its mathematical and physiological justification:

> *“It is obvious that the undulation condition should set an upper limit to the resolving power. The surprising fact is that this limit is* apparently actually attained*, and that the doublet still appears resolved, the effect of contrast so intensifying the edges that the eye supplies a minimum where none exists. The effect is observable both in positives and in negatives, as well as by direct vision...My own observations on this point have been checked by a number of my friends and colleagues.”*

In the survey of den Dekker and van den Bos [DDVdB97, p. 548] eighty years later:

> *“Since Rayleigh’s days, technical progress has provided us with more and more refined sensors. Therefore, when visual inspection is replaced by intensity measurement, the natural resolution limit that is due to diffraction would be [the Sparrow limit]...even a hypothetical perfect measurement instrument would not be able to detect a central dip in the composite intensity distribution, simply because there is no such dip anymore.”*

In light of advancements in super-resolution microscopy, rigorously characterizing the resolving power of imaging systems remains as pressing a challenge as ever.

In 2017, Demmerle et al. [DWSD15] revisited what resolution means in light of these new technologies technologies and propose approaches for comparing resolution between different super-resolution methods. As they note in their introduction [DWSD15, p. 3]:

> *"The recent introduction of a range of commercial super-resolution instruments means that resolution has once again become a battleground between different microscope technologies and rival companies."*

Notably, in the conclusion, they remark that a classical Houston criterion-style approach is still the best for comparing different methods [DWSD15, p. 9].

> *"Given the above points, the best approach to compare between techniques is still to perform the simple and robust fitting of a Gaussian to a sub-resolution object and then to extract the FWHM."*

9.9.2 The Importance of Noise

An idea that has been repeated one way or another in the literature is that if one has perfect access to the *exact* intensity profile of the diffraction image of two point sources, then one could brute-force search over the space of possible parameters to find a hypothesis that fits the point spread function arbitrarily well, thereby learning the positions of the point sources regardless of their separation. As such, for any notion of diffraction limit to have practical meaning, it must take into account factors like aberrations and measurement noise that preclude getting perfect access to the intensity profile.

This perspective was distilled emphatically by di Francia [DF55, p. 497] in 1955:

> *"Moreover it is only too obvious that from the mathematical standpoint, the image of two points, however close to one another, is different from that of one point. It is not at all absurd to assume that technical progress may provide us with more and more refined kinds of receptors, detecting the difference between the image of a single point and the image of two points located closer and closer to another. This means that at present there is only a* practical *limit (if any) and not a* theoretical *limit for two-point resolving power."*

Contemporaneously, in discussions at the 1955 Meeting of the German Society of Applied Optics culminating in [Ron61, p. 459], Ronchi made the following distinction:

> *"Nowadays it seems imperative to differentiate three kinds of images, i.e., (1) the ethereal image, (2) the calculated image, and (3) the detected image.*
>
> *The nature of the ethereal image should be physical, but in reality it is only a hypothesis. It is said that the radiant flux emitted by the object...is concentrated and distributed in the so-called image by means of a number of processes. But actually this is only a hypothesis...attempts have been made to give a mathematical representation of the phenomenon, both geometrically and algebraically...The images which have been calculated in this way...should therefore be called calculated images.*
>
> *If we now consider the field of experience, we find the detected images. They are the figures either perceived by the eye when looking through the instrument, or obtained by means of a photosensitive emulsion, or through a photoelectric device.*

den Dekker and van den Bos [DDVdB97, p. 547] in their 1997 survey:

> *"Since Ronchi's paper, further research on resolution— concerning detected images instead of calculated ones— has shown that in the end, resolution is limited by systematic and random errors resulting in an inadequacy of the description fo the observations by the mathematical model chosen. This important conclusion was independently drawn by many researchers who were approaching the concept of resolution from different points of view."*

den Dekker and van den Bos summarize the state of affairs as follows [DDVdB97, p. 547]:

> *"If calculated images were to exist, the known two-component model could be fitted numerically to the observations with respect to the component locations and amplitudes. Then the solutions for these locations and amplitudes would be exact, a perfect fit would result, and in spite of diffraction there would be no*

> *limit to resolution no matter how closely located the two point sources; this would mean that no limit to resolution for calculated images would exist. However, imaging systems constructed without any aberration or irregularity are an ideal that is never reached in practice.... Therefore one should consider the resolution of detected images instead of calculated images."*

Goodman [Goo15, p. 326-7] in 2000:

> *"...the question of when two closely spaced point sources are barely resolved is a complex one and lends itself to a variety of rather subjective answers...An alternative definition is the so-called Sparrow criterion...In fact, the ability to resolve two point sources depends fundamentally on the signal-to-noise ratio associated with the detected image intensity pattern, and for this reason criteria that do not take account of noise are subjective."*

Maznev and Wright [MW17, p. 3] in 2016 on the earlier quote by Born and Wolf:

> *"Indeed, if any number of photons is available for the measurement, there is no fundamental limit to how well one can resolve two point sources, since it is possible to make use of curve fitting to arbitrary precision (however, there are obvious practical limitations related to the finite measurement time and other factors such as imperfections in the optical system, atmospheric turbulence, etc)."*

Demmerle et al. in the work mentioned in the previous section [DWSD15, P. 9]:

> *"If one, a priori, knows that there are two point sources, then measuring their separation, and hence calculating the system's resolution is purely limited by Signal-to-Noise Ratio."*

A related point that has been made repeatedly in the literature is that the original setting in which Abbe introduced his diffraction limit should not be conflated with the setting of resolving two point sources of light.

In the work of di Francia cited above [DF55, p. 498], he notes that the classic impossibility result for resolving a lattice of alternatively dark and bright points with separation below the Abbe limit says nothing about the impossibility of resolving a pair of points sources:

> *"[The impossibility result at the Abbe limit] has often been given a wrong interpretation and it has too hastily been extended to the case of two points. The [Abbe limit] applies only* when we want the available information uniformly distributed *over the whole image. Mathematics cannot set any lower limit for the distance of two resolvable points."*

Indeed, he argues informally, by way of the Nyqist sampling theorem, that when there is a prior on the number of components in a superposition of Airy disks being upper bounded by a known constant, then in theory, there is no diffraction limit. Rather, he posits, it is the entropy of the prior that dictates the limits of resolution [DF55, p. 498]:

> *"The fundamental question of how many independent data are contained in an image formed by a given optical instrument. This seems to be the modern substitute for the theory of resolving power."*

Sheppard [She17, p. 597] in 2017, sixty years after di Francia's work, clarifies again that the abovementioned impossibility result should not be misinterpreted as saying anything about the impossibility of resolving two point sources:

> "The Abbe resolution limit is a sharp limit to the imaging of a periodic object such as a grating. Super-resolution refers to overcoming this resolution limit. The Rayleigh resolution criterion refers to imaging of a two-point object. It is based on an arbitrary criterion, and does not define a sharp transition between structures being resolved or not resolved."

9.10 Appendix: Proof of Lemma 9.4.15

TensorResolve (Algorithm 44) uses the standard subroutine given in Algorithm 46. We remark that this algorithm appears to be deterministic unlike usual treatments of Jennrich's algorithm simply because we have absorbed the usual randomness of the choice of flattening into the construction of the tensor $\mathbf{T}$ on which TensorResolve calls Jennrich.

```
Algorithm 46: Jennrich($\widetilde{\mathbf{T}}$)
  Input: Tensor $\widetilde{\mathbf{T}} \in \mathbb{C}^{m\times m\times 3}$ which is close to a rank-$k$ tensor $\mathbf{T}$ of the form (9.12)
  Output: $\widehat{V} \in \mathbb{C}^{m\times k}$ close to $V$ up to column permutation (see Lemma 9.4.15)
  Compute the $k$-SVD $\widehat{P}\widehat{\Lambda}\widehat{P}^{\dagger}$ of the flattening $\widetilde{\mathbf{T}}(\mathrm{Id}, \mathrm{Id}, e_1)$.
  Define the whitened tensor $\widehat{\mathbf{E}} = \widetilde{\mathbf{T}}(\widehat{P}, \widehat{P}, \mathrm{Id})$ and its flattenings $\widehat{E}_i \triangleq \widehat{\mathbf{E}}(\mathrm{Id}, \mathrm{Id}, e_i)$ for
    $i \in [2]$.
  $\widehat{M} \leftarrow \widehat{E}_1 \widehat{E}_2^{-1}$.
  Form the matrix $\widehat{U}$ whose columns are equal to the eigenvectors, scaled to have
    norm $\sqrt{m}$, for the $k$ eigenvalues of $\widehat{M}$ that are largest in absolute value.
  return $\widehat{V} \triangleq \widehat{P}\widehat{U}$.
```

We restate Lemma 9.4.15 here for the reader's convenience:

Lemma 9.4.15. *[e.g. [HK15], Lemma 3.5] For any $\varepsilon, \delta > 0$, suppose $|\mathbf{T}_{a,b,i} - \widetilde{\mathbf{T}}_{a,b,i}| \le \eta'$ for $\eta' \triangleq O\left(\frac{(c-\overline{\gamma})\delta\Delta\lambda_{\min}^2}{k^{5/2}m^{3/2}\kappa(V)^5} \cdot \varepsilon\right)$, and let $\widehat{V} = $ Jennrich$(\widetilde{\mathbf{T}})$ (Algorithm 46). Then with probability at least $1-\delta$ over the randomness of $v^{(1)}$, there exists permutation matrix Π such that $\|\widehat{V} - V\Pi\|_F \le \varepsilon$ for all $j \in [k]$.*

This proof closely follows that of [HBZ10], though we must make some modifications because the scaling of the frequencies $v^{(i)}$ for $i \in [3]$ defined in Step 4 of TensorResolve is different.

Proof. We first define the noiseless versions of the objects $\widehat{P}, \widehat{\Lambda}, \widehat{E}, \widehat{E}_1, \widehat{E}_2, \widehat{M}, \widehat{U}$ introduced in Jennrich. Note that for $i \in [2]$,

$$\mathbf{T}(\mathrm{Id}, \mathrm{Id}, e_i) = VD_iV^{\dagger} \tag{9.24}$$

for D_i the diagonal matrix whose diagonal entries are given by $\{\lambda_j e^{-2\pi i\langle \mu_j, v^{(i)}\rangle}\}_{j\in[k]}$. Denote the k-SVD of $\mathbf{T}(\mathrm{Id}, \mathrm{Id}, e_1)$ by $P\Lambda P^{\dagger}$. Define the whitened tensor $\mathbf{E} \triangleq \mathbf{T}(P, P, \mathrm{Id})$ and its flattenings $E_i = \mathbf{E}(\mathrm{Id}, \mathrm{Id}, e_i)$ for $i \in [2]$. Finally, define $U \triangleq P^{\dagger}V$ so that

$$\mathbf{E} = \sum_{j=1}^{k} \lambda_j U^j \otimes U^j \otimes W^j$$

and $E_i = UD_iU^\dagger$ for $i \in [2]$. Note that U also satisfies $M \triangleq E_1E_2^{-1} = UDU^\dagger$ for diagonal matrix $D \triangleq D_1D_2^{-1}$, and for every $j \in [k]$, $\|U^j\|_2 = \|V^j\|_2 = \sqrt{m}$, so U is indeed the noiseless analogue of $\widehat{U}$.

For any $j \in [k]$, we have that

$$D_{j,j} = e^{-2\pi i\langle \mu_j, v^{(1)} - v^{(2)}\rangle}.$$

Define $\Delta_D \triangleq \min_{j\neq j'} |D_{j,j} - D_{j',j'}|$.

For every $j, j' \in [k]$, by triangle inequality and the fact that $V^j = PU^j$ and $\widehat{V}^{=}\widehat{P}\widehat{U}^j$, we have

$$\|\widehat{V}^j - V^{j'}\|_2 \le \|\widehat{P} - P\|_2\|\widehat{U}^j\|_2 + \|P\|_2\|\widehat{U}^j - U^{j'}\|_2 \le \sqrt{m}\|\widehat{P} - P\|_2 + \|\widehat{U}^j - U^{j'}\|_2. \quad (9.25)$$

We proceed to upper bound $\|\widehat{P} - P\|_2$ and $\|\widehat{U}^j - U^{j'}\|_2$.

Lemma 9.10.1. $\|\widehat{P} - P\|_2 \le \frac{\eta'\sqrt{m}}{\lambda_{\min}\sigma_{\min}(V)^2}$.

Proof. By Wedin's theorem,

$$\|\widehat{P} - P\|_2 \le \frac{\|\widetilde{\mathbf{T}}(\mathrm{Id}, \mathrm{Id}, e_1) - \mathbf{T}(\mathrm{Id}, \mathrm{Id}, e_1)\|_2}{\sigma_{\min}(\mathbf{T}(\mathrm{Id}, \mathrm{Id}, e_1))}.$$

By (9.24), $\sigma_{\min}(\mathbf{T}(\mathrm{Id}, \mathrm{Id}, e_1)) \ge \lambda_{\min}\sigma_{\min}(V)^2$. Additionally, $\|\widetilde{\mathbf{T}}(\mathrm{Id}, \mathrm{Id}, e_1) - \mathbf{T}(\mathrm{Id}, \mathrm{Id}, e_1)\|_F \le \eta'\sqrt{m}$, from which the claim follows. □

Lemma 9.10.2. *If $\|M - \widehat{M}\|_2 \le \frac{\Delta_D}{2\sqrt{k}\kappa(U)}$, then the eigenvalues of $\widehat{M}$ are distinct, and there exists a permutation τ for which*

$$\|\widehat{U}^j - U^{\tau(j)}\|_2 \le \frac{3m\|M - \widehat{M}\|_2}{\Delta_D\sigma_{\min}(U)} \quad \forall\ j \in [k].$$

Proof. Consider the matrix $U^{-1}\widehat{M}U = D - U^{-1}(M - \widehat{M})U$. Because $\|U^{-1}(M - \widehat{M})U\|_2 \le \Delta_D/2\sqrt{k}$ by assumption, we conclude by Gershgorin's that the eigenvalues of $U^{-1}\widehat{M}U$, and thus of $\widehat{M}$, are distinct and each lies within $\Delta_D/2$ of a unique eigenvalue of M. Let τ be the permutation matching eigenvalues $\{\widehat{\beta}_j\}$ of $\widehat{M}$ to eigenvalues $\{\beta_j\}$ of M which are closest, and without loss of generality let τ be the identity permutation.

For fixed $j \in [k]$, let $\{c_{j'}\}$ be coefficients for which $\widehat{U}^j = \sum c_{j'} U^j$ and $\sum_{j'} c_{j'}^2 = 1$. Note that we have

$$\widehat{\lambda}_j \sum_{j'} c_{j'} U^{j'} = \widehat{\lambda}_j \widehat{U}^j = \widehat{M}\widehat{U}^j = \sum_{j'} \lambda_{j'} c_{j'} U^{j'} + (M - \widehat{M})\widehat{U}^j,$$

so $\{c_{j'}\}$ is the solution to the linear system

$$\sum_{j'} c_{j'} \cdot (\widehat{\lambda}_j - \lambda_{j'}) U^{j'} = (M - \widehat{M})\widehat{U}^j.$$

Recalling that $\|U^j\|_2 = \sqrt{m}$ and that $\sum c_{j'}^2 = 1$, we get that

$$\begin{aligned}\|\widehat{U}^j - U^j\|_2^2 &= \sum_{j' \neq j} c_{j'}^2 \|U^{j'}\|_2^2 + (c_j - 1)^2 \|U^j\|_2^2 \leq 2m \sum_{j' \neq j} c_{j'}^2 \\ &\leq \frac{8m\|U^{-1}(M - \widehat{M})\widehat{U}^j\|_2^2}{\Delta_D^2} \leq \frac{8m^2\|M - \widehat{M}\|_2^2}{\Delta_D^2 \sigma_{\min}(U)^2}.\end{aligned}$$

□

Finally, we must estimate $\|M - \widehat{M}\|_2^2$ in the bound in Lemma 9.10.2:

Lemma 9.10.3. *If $\eta' \leq \frac{\lambda_{\min}^2 \sigma_{\min}(V)^2}{6\sqrt{m}\kappa(V)^2}$, then $\|M - \widehat{M}\|_2 \leq \frac{9\eta'\sqrt{m}\kappa(V)^2}{\lambda_{\min}^2 \sigma_{\min}(V)^2}$.*

Proof. Define $Z_i \triangleq \widehat{E}_i - E_i$ for $i \in [2]$ so by taking Schur complements

$$M - \widehat{M} = E_1 E_2^{-1} - (E_1 + Z_1)(E_2 + Z_2)^{-1} = M Z_2 (\mathrm{Id} + E_2^{-1} Z_2)^{-1} E_2^{-1} + Z_1 E_2^{-1} \triangleq MH + G \tag{9.26}$$

Note that

$$\sigma_{\max}(H) \leq \frac{\|Z_2\|_2}{\sigma_{\min}(E_2) - \|Z_2\|_2} \leq \frac{\|Z_2\|_2}{\lambda_{\min}\sigma_{\min}(U)^2 - \|Z_2\|_2}, \quad \sigma_{\max}(G) \leq \frac{\sigma_{\max}(Z_1)}{\sigma_{\min}(E_2)} \leq \frac{\|Z_1\|_2}{\lambda_{\min}\sigma_{\min}(U)^2}$$

and furthermore for either $i \in [2]$, because $Z_i = \widehat{P}^\dagger \widetilde{\mathbf{T}}(\mathrm{Id},\mathrm{Id},e_i)\widehat{P} - P^\dagger \mathbf{T}(\mathrm{Id},\mathrm{Id},e_i)P$,

$$\|Z_i\|_2 \le \|P\|_2\|\mathbf{T}(\mathrm{Id},\mathrm{Id},e_i)\|_2\|P-\widehat{P}\|_2 + \|\widehat{P}\|_2\|\mathbf{T}(\mathrm{Id},\mathrm{Id},e_i)\|_2\|\widehat{P}-P\|_2 + \|\widehat{P}\|_2^2\|\widetilde{\mathbf{T}}(\mathrm{Id},\mathrm{Id},e_i)\|_2$$

$$\le 2\frac{\eta'\sqrt{m}}{\lambda_{\min}\sigma_{\min}(V)^2}\cdot \lambda_{\max}\sigma_{\max}(V)^2 + \lambda_{\max}\sigma_{\max}(V)^2 \le \frac{3\eta'\sqrt{m}\kappa(V)^2}{\lambda_{\min}} \tag{9.27}$$

Because $\sigma_{\min}(U)^2 = \sigma_{\min}(V)^2$, by the bound on η' in the hypothesis, $\sigma_{\max}(H) \le \frac{2\|Z_2\|_2}{\lambda_{\min}\sigma_{\min}(V)^2}$. Finally, noting that $\|M\|_2 \le \sigma_{\max}(D) = 1$, we conclude the proof from (9.26) and (9.27). □

It remains to bound Δ_D.

Lemma 9.10.4. *For any $\delta > 0$, with probability at least $1-\delta$, $\Delta_D \ge O\left(\frac{(c-\overline{\gamma})\delta'\Delta}{k^2}\right)$.*

Proof. Using the elementary inequality $|e^{-2\pi i x} - 1| \le 2\pi|x|$ for any $x \in \mathbb{R}$, we conclude that $|D_{j,j} - D_{j',j'}| \le |e^{-2\pi i\langle \mu_j - \mu_{j'}, v^{(1)} - v^{(2)}\rangle} - 1| \le 2\pi|\langle \mu_j - \mu_{j'}, v^{(1)} - v^{(2)}\rangle|$. By standard anti-concentration, for any $j \ne j'$ and $\delta' > 0$ we have that $|\langle \mu_j - \mu_{j'}, v^{(1)} - v^{(2)}\rangle| \le O(\delta'\|\mu_j - \mu_{j'}\|_2 \cdot \|v^{(1)} - v^{(2)}\|_2)$ with probability at most δ'. The proof follows by taking $\delta' = \delta/k^2$, union bounding, and recalling the definition of $v^{(1)}, v^{(2)}$ in TensorResolve. □

Combining (9.25) and Lemmas 9.10.1, 9.10.2, 9.10.3, 9.10.4, there exists a permutation τ for which

$$\|\widehat{V}^j - V^{\tau(j)}\|_2 \le \frac{\eta' m}{\lambda_{\min}\sigma_{\min}(V)^2} + \frac{27\eta' m^{3/2}\kappa(V)^2}{\Delta_D\lambda_{\min}^2\sigma_{\min}(V)^3} \le O\left(\frac{k^2\eta' m^{3/2}\kappa(V)^5}{(c-\overline{\gamma})\delta\Delta\lambda_{\min}^2}\right) \quad \forall j \in [k].$$

We conclude that for the permutation matrix Π corresponding to τ, $\|\widehat{V} - V\Pi\|_F \le \sqrt{k}\max_{j\in[k]}\|\widehat{V}^j - V^\tau(j)\|_2 \le O\left(\frac{k^{5/2}\eta' m^{3/2}\kappa(V)^5}{(c-\overline{\gamma})\delta\Delta\lambda_{\min}^2}\right)$ as claimed. □

9.11 Appendix: Generating Figure 9-3

Here we elaborate on how Figure 9-3 was generated. While Theorem 9.5.1 yields an explicit construction which rigorously demonstrates the phase transition at the diffraction limit, empirically we found that this phase transition was even more pronounced when we slightly modified the construction. Specifically, we empirically evaluated the following instance:

for even k, separation $\Delta > 0$, and $1 \le i \le k$, let $\mu_i = (a_i, 0)$ and let $\mu'_i = (b_i, 0)$ for $a_i \triangleq \frac{\Delta}{2} \cdot \left(2i - \frac{k+3}{2}\right)$ and $b_i \triangleq \frac{\Delta}{2} \cdot \left(2i - \frac{k+1}{2}\right)$, and take $\{\lambda_i\}$ and $\{\lambda'_i\}$ to be the unique solution to the affine system

$$\sum_{i=1}^{\lceil k/2 \rceil} \lambda_i = 1 \text{ and } \sum_{i=1}^{\lfloor k/2 \rfloor} \lambda'_i = 1 \qquad\qquad \sum_{i=1}^{\lceil k/2 \rceil} \lambda_i a_i^\ell = \sum_{i=1}^{\lfloor k/2 \rfloor} \lambda'_i b_i^\ell \;\; \forall \;\; 0 \le \ell < k-1.$$

These are the weights for which the superposition of point masses at $\{\mu_i\}$ with weights $\{\lambda_i\}$ matches the superposition of point masses at $\{\mu'_i\}$ with weights $\{\lambda'_i\}$ on all moments of degree at most $k-2$. While moment-matching does not directly translate to any kind of statistical lower bound, it is often the starting point for many such lower bounds in the distribution learning literature [MV10, DKS17, HP15, Kea98]. The "carefully chosen pair of superpositions" referenced in the caption of Figure 9-3 refers to this moment-matching construction. Henceforth refer to these two superpositions, both of which are Δ-separated superpositions of $k/2$ Airy disks, as $\mathcal{D}_0(\Delta, k)$ and $\mathcal{D}_1(\Delta, k)$ respectively. We will omit the parenthetical Δ, k when the context is clear.

Unfortunately, there is no closed form for the expression for $d_{\mathrm{TV}}(\mathcal{D}_0, \mathcal{D}_1)$. Instead, we estimated this via numerical integration. Direct evaluation of the integral $\int_{\mathbb{R}^2} |\mathcal{D}_0(\mathbf{x}) - \mathcal{D}_1(\mathbf{x})| \, dx$ poses issues because of the heavy tails of the Airy point spread function. To tame these tails, we used a carefully chosen proposal measure μ in order to rewrite $d_{\mathrm{TV}}(\mathcal{D}_0, \mathcal{D}_1)$ as $\int_{\mathbb{R}^2} \left| \frac{\mathcal{D}_0(\mathbf{x})}{\mu(\mathbf{x})} - \frac{\mathcal{D}_1(\mathbf{x})}{\mu(\mathbf{x})} \right| d\mu$. Because of the heavy tails, we needed to use a similarly heavy-tailed proposal distribution, so we took μ to be the convolution of the superposition of point masses at $\{\mu_i\} \cup \{\mu'_i\}$ having weights $\{\lambda_i\} \cup \{\lambda'_i\}$ with the following kernel $P(\cdot)$. To sample from the density over $\mathbb{R}^2$ correpsonding to P, with probability $1/2$ sample a radius r uniformly from $[0, 1]$ and output a random vector in $\mathbb{R}^2$ of norm r, and with the remaining probability $1/2$, sample from the Pareto distribution with parameter $2/3$ over $[1, \infty]$ and output a random vector of norm r. The motivation for P and in particular for the parameter $2/3$ is that it is a rough approximation to the tail behavior of the radial density $\frac{J_1(r)^2}{r}$ defining the Airy point spread function, which by Theorem 9.3.6 decays roughly as $r^{-5/3}$.

To generate the curves in Figure 9-3, for each $k \in [2, 4, 6, 12, 20, 30, 42, 56, 72, 90]$ and each $\Delta \in [-2, -1.92, -1.84, ..., 1.84, 1.92, 2]$, we simply estimated the corresponding $d_{\mathrm{TV}}(\mathcal{D}_0, \mathcal{D}_1)$

by sampling 10 million points $\mathbf{x}$ from μ and computing the empirical mean of the quantity $\left|\frac{\mathcal{D}_0(\mathbf{x})}{\mu(\mathbf{x})} - \frac{\mathcal{D}_1(\mathbf{x})}{\mu(\mathbf{x})}\right|$.

We have made the code for Figure 9-3 available at `https://github.com/secanth/airy/`.

Chapter 10

Quantum Memory-Sample Tradeoffs for Mixedness Testing

10.1 Introduction

In the last two chapters of this thesis, we study the problem of *quantum state certification*. We begin by recalling the setup and motivation. Recall from Definition 1.2.36 that in this problem, we are given N copies of an unknown mixed state $\rho \in \mathbb{C}^{d\times d}$ and a description of a known mixed state σ, and our goal is to make measurements on these copies and use the outcomes of these measurements to distinguish whether $\rho = \sigma$, or if it is ε-far from σ in trace norm. An important special case of this is when σ is the maximally mixed state, in which case the problem is known as *quantum mixedness testing*.

This problem is motivated by the need to verify the output of quantum computations. In many applications, a quantum algorithm is designed to prepare some known d-dimensional mixed state σ. However, due to the possibility of noise or device defects, it is unclear whether or not the output state is truly equal to σ. Quantum state certification allows us to verify the correctness of the quantum algorithm. In addition to this more practical motivation, quantum state certification can be seen as the natural non-commutative analogue of *identity testing* of (classical) probability distributions, a well-studied problem in statistics and theoretical computer science.

Recently, [OW15] demonstrated that $\Theta(d/\varepsilon^2)$ copies are necessary and sufficient to solve quantum mixedness testing with good confidence. Subsequently, [BOW19] demonstrated that the same copy complexity suffices for quantum state certification. Note that these copy complexities are sublinear in the number of parameters in ρ, and in particular, are less than the $\Theta(d^2/\varepsilon^2)$ copies necessary to learn ρ to ε error in trace norm [OW16, HHJ+17].

To achieve these copy complexities, the algorithms in [OW15, BOW19] heavily rely on entangled measurements. These powerful measurements allow them to leverage the representation theoretic structure of the underlying problem to dramatically decrease the copy complexity. However, this power comes with some tradeoffs. Entangled measurements require that all N copies of ρ are measured simultaneously. Thus, all N copies of ρ must be kept in quantum memory without any of them de-cohering. Additionally, the positive-operator valued measure (POVM) elements that formally define the quantum measurement must all be of size $d^N \times d^N$; in particular, the size of the POVM elements scales exponentially with N. Both of these issues are problematic for using any of these algorithms in practice [CW20]. Entangled measurements are also necessary for the only known sample-optimal algorithms for quantum tomography [OW16, HHJ+17, OW17].

This leads to the question: can these sample complexities be achieved using weaker forms of measurement? There are two natural classes of such restricted measurements to consider:

- an (unentangled) *nonadaptive measurement* fixes N POVMs ahead of time, measures each copy of ρ using one of these POVMs, then uses the results to make its decision.
- an (unentangled) *adaptive measurement* measures each copy of ρ sequentially, and can potentially choose its next POVM based on the results of the outcomes of the previous experiments.

It is clear that fully entangled measurements are strictly more general than adaptive measurements, which are in turn strictly more general than nonadaptive ones. However, both nonadaptive and adaptive measurements have the advantage that the quantum memory they require is substantially smaller than what is required for a generic entangled measurement. In particular, only one copy of ρ need be prepared at any given time, as opposed to the N copies that must simultaneously be created, if we use general entangled measurements.

Separating the power of entangled vs. nonentangled measurements for such quantum learning and testing tasks was posed as an open problem in [Wri16]. In this paper, we demonstrate the first such separations for quantum state certification, and to our knowledge, the first separation between adaptive measurements and entangled measurements without any additional assumptions on the measurements, for any quantum estimation task.

We first show a sharp characterization of the copy complexity of quantum mixedness testing with nonadaptive measurements:

Theorem 10.1.1. *If only unentangled, nonadaptive measurements are used, $\Theta(d^{3/2}/\varepsilon^2)$ copies are necessary and sufficient to distinguish whether $\rho \in \mathbb{C}^{d\times d}$ is the maximally mixed state, or if ρ has trace distance at least ε from the maximally mixed state, with probability at least 2/3.*

We defer a proof of the upper bound in this theorem to the next chapter (see Lemma 11.6.2).

Second, we show that $\omega(d)$ copies are necessary, even with adaptive measurements. We view this as our main technical contribution. Formally:

Theorem 10.1.2. *If only unentangled, possibly adaptive, measurements are used, $\Omega(d^{4/3}/\varepsilon^2)$ copies are necesssary to distinguish whether $\rho \in \mathbb{C}^{d\times d}$ is the maximally mixed state, or has trace distance at least ε from the maximally mixed state, with probability at least 2/3.*

As quantum state certification is a strict generalization of mixedness testing, Theorems 10.1.1 and 10.1.2 also immediately imply separations for that problem as well. Note that the constant 2/3 in the above theorem statements is arbitrary and can be replaced with any constant greater than 1/2. We also remark that our lower bounds make no assumptions on the number of outcomes of the POVMs used, which can be infinite (see Definition 1.3.47).

10.1.1 Overview of our techniques

In this section, we give a high-level description of our techniques. We start with the lower bounds.

"Lifting" classical lower bounds to quantum ones Our lower bound instance can be thought of as the natural quantum analogue of Paninski's for (classical) uniformity testing:

Theorem 10.1.3 (Theorem 4, [Pan08])**.** $\Omega(\sqrt{d}/\varepsilon^2)$ *samples are necessary to distinguish* *whether a distribution p over $\{1,\ldots,d\}$ is ε-far from the uniform distribution in total variation distance, with confidence at least* $2/3$*.*

At a high level, Paninski demonstrates that it is statistically impossible to distinguish between the distribution $p_0^{\leq N}$ of N independent draws from the uniform distribution, and the distribution $p_1^{\leq N}$ of N independent draws from a random perturbation of the uniform distribution, where the marginal probability of each element in $\{1,\ldots,d\}$ has been randomly perturbed by $\pm\varepsilon/d$ (see Example 10.2.7).

The hard instance we consider can be viewed as the natural quantum analogue of Paninski's construction. Roughly speaking, rather than simply perturbing the marginal probabilities of every element in $\{1,\ldots,d\}$, which corresponds to randomly perturbing the diagonal elements of the mixed state, we also randomly rotate it (see Construction 2). We note that this hard instance is not novel and has been considered before in similar settings [OW15, Wri16, HHJ$^+$17]. However, our analysis technique is quite different from previous bounds, especially in the adaptive setting.

The technical crux of Paninski's lower bound is to upper bound the total variation distance between $p_0^{\leq N}$ and $p_1^{\leq N}$ in terms of the χ^2-divergence between the two. This turns out to have a simple, explicit form, and can be calculated exactly. This works well because, conditioned on the choice of the random perturbation in $p_1^{\leq N}$, both of the distributions $p_0^{\leq N}$ and $p_1^{\leq N}$ have a product structure, as they consist of N independent samples.

This product structure still holds true in the quantum case when we restrict to nonadaptive measurements. This allows us to do a more involved version of Paninski's calculation in the quantum case and thus obtain the lower bound in Theorem 10.1.1.

However, this product structure breaks down completely in the adaptive setting, as now the POVMs, and hence, the measurement outcomes that we observe, for the t-th copy of ρ, can depend heavily on the previous outcomes. As a result, the χ^2-divergence between the analogous quantities to $p_0^{\leq N}$ and $p_1^{\leq N}$ no longer have a nice, closed form, and it is not clear how to proceed using Paninski's style of argument.

Instead, inspired by the literature on bandit lower bounds [ACBFS02, BCB12], we upper bound the total variation distance between $p_0^{\leq N}$ and $p_1^{\leq N}$ by the KL divergence between these two quantities. The primary advantage of doing so is that the KL divergence satisfies the chain rule. This allows us to partially disentangle how much information that the t-th copy of ρ gives the algorithm, conditioned on the outcomes of the previous experiments.

At present, this chain-rule formulation of Paninski's lower bound seems to be somewhat lossy. Even in the classical case, we need additional calculations tailored to Paninski's instance to recover the $\Omega(\sqrt{d}/\varepsilon^2)$ bound for uniformity testing (see Appendix 10.8), without which our approach can only obtain a lower bound of $\Omega(d^{1/3}/\varepsilon^2)$ (see Section 10.5). At a high level, this appears to be why we do not obtain a lower bound of $\Omega(d^{3/2}/\varepsilon^2)$ for adaptive measurements. We leave the question of closing this gap as an interesting future direction.

"Projecting" quantum upper bounds to classical ones While the lower bound techniques we employ are motivated by the lower bounds for classical testing, they do not directly use any of those results. In contrast, to obtain our upper bounds, we demonstrate a direct reduction from non-adaptive mixedness testing to classical uniformity testing. The reduction is as follows. First, we choose a random orthogonal measurement basis. Measuring ρ in this basis induces some distribution over $\{1, \ldots, d\}$. If ρ is maximally mixed, this distribution is the uniform distribution. Otherwise, if it is far from maximally mixed, then by similar concentration of measure phenomena as used in the proof of the lower bounds, with high probability this distribution will be quite far from the uniform distribution in L_2 distance. Thus, to distinguish these two cases, we can simply run a classical L_2 uniformity tester [CDVV14, DKN14, CDGR18]. See Section 11.6.1 for more details.

Concentration of measure over the unitary group In both our lower bounds and upper bounds, it will crucial to carefully control the deviations of various functions of Haar random unitary matrices. In fact, specializations of quantities we encounter have been extensively studied in the literature on quantum transport in mesoscopic systems, namely the conductance of a chaotic cavity [BB96, Bee97, BB00, KSS09, AOK09], though the tail bounds we need are not captured by these works (see Section 10.3.3 for more details). Instead, we will rely on more general tail bounds [MM13] that follow from log-Sobolev inequalities on the unitary group $U(d)$.

10.1.2 Related Work

The literature on quantum (and classical) testing and learning is vast and we cannot hope to do it justice here; for conciseness we only discuss some of the more relevant works below.

Quantum state certification fits into the general framework of quantum state property testing problems. Here the goal is to infer non-trivial properties of the unknown quantum state, using fewer copies than are necessary to fully learn the state. See [MdW16] for a more complete survey on property testing of quantum states. Broadly speaking, there are two regimes studied here: the asymptotic regime and the non-asymptotic regime.

In the asymptotic regime, the goal is to precisely characterize the exponential convergence of the error as $n \to \infty$ and d, ε are held fixed and relatively small. In this setting, quantum state certification is commonly referred to as *quantum state discrimination*. See e.g. [Che00, ANSV08, BC09] and references within. However, this allows for rates which could depend arbitrarily badly on the dimension.

In contrast, we work in the non-asymptotic regime, where the goal is to precisely characterize the rate of convergence as a function of d and ε. The closest work to ours is arguably [OW15] and [BOW19]. The former demonstrated that the copy complexity of quantum mixedness testing is $\Theta(d/\varepsilon^2)$, and the latter showed that quantum state certification has the same copy complexity. However, as described previously, the algorithms which achieve these copy complexities heavily rely on entangled measurements.

Another interesting line of work focuses on the case where the measurements are only allowed to be Pauli matrices [FL11, FGLE12, dSLCP11, AGKE15]. Unfortunately, even for pure states, these algorithms require $\Omega(d)$ copies of ρ. We note in particular the paper of [FGLE12], which gives a $\Omega(d)$ lower bound for the copy complexity of the problem, even when the Pauli measurements are allowed to be adaptively chosen. However, their techniques do not appear to generalize easily to arbitrary adaptive measurements.

We also mention [Yu19] which gives algorithms for various quantum property testing problems using *local measurements* which non-adaptively operate on each individual qubit. Because this is a more restrictive family of measurements, the sample complexity for these algorithms suffers some polynomial overhead as a function of d.

A related task is that of quantum tomography, where the goal is to recover ρ, typically to good fidelity or low trace norm error. The paper [HHJ+17] showed that $O(d^2 \log(d/\varepsilon)/\varepsilon^2)$ copies suffice to obtain ε trace error, and that $\Omega(d^2/\varepsilon^2)$ copies are necessary. Independently, [OW16] improved their upper bound to $O(d^2/\varepsilon^2)$. These papers, in addition to [OW17], also discuss the case when ρ is low rank, where $o(d^2)$ copy complexity can be achieved. Notably, all the upper bounds that achieve the tight bound heavily require entanglement. In [HHJ+17], they demonstrate that $\Omega(d^3/\varepsilon^2)$ copies are necessary, if the measurements are nonadaptive. It is a very interesting question to understand the power of adaptive measurements for this problem as well.

Quantum state certification and quantum mixedness testing are the natural quantum analogues of classical identity testing and uniformity testing, respectively, which both fit into the general setting of (classical) distribution testing. There is again a vast literature on this topic; see e.g. [Can20, Gol17] for a more extensive treatment of the topic. Besides the papers covered previously and in the surveys, we highlight a line of work on testing with *conditional sampling oracles* [CRS15, CFGM16, CRS14, ACK14, BC18, KT19], a classical model of sampling which also allows for adaptive queries. It would be interesting to see if the techniques we develop here can also be used to obtain stronger lower bounds in this setting. Adaptivity also plays a major role in property testing of functions [BB16, CWX17a, KS16, BCP+17, CWX17b, Bel18], although these problems appear to be technically unrelated to the ones we consider here.

Roadmap The rest of the paper is organized as follows:

- Section 10.2— We describe a generic setup that captures Paninski's and our settings as special cases and provide an overview of the techniques needed to show lower bounds in this setup.

- Section 10.3— We formalize the notion of quantum property testing via adaptive measurements, define our lower bound instance, and perform some preliminary calculations.

- Section 10.4— Proof of the lower bound in Theorem 10.1.1.

- Section 10.5— As a warmup to the proof of Theorem 10.1.2, we prove a weaker version of Paninski's lower bound using our chain rule approach.

- Section 10.6— Proof of our main result, Theorem 10.1.2.
- Section 10.7— Proof of certain tail bounds for Haar-random unitary matrices which are crucial to the proofs of Theorems 10.1.1 and 10.1.2.
- Appendix 10.8— A more ad hoc chain rule proof of Paninski's optimal $\Omega(\sqrt{d}/\varepsilon^2)$ lower bound.

10.2 Lower Bound Strategies

The lower bounds we show in this work are lower bounds on the number of observations needed to distinguish between a simple null hypothesis and a mixture of alternatives. For instance, in the context of classical uniformity testing, the null hypothesis is that the underlying distribution is the uniform distribution over $[d]$, and the mixture of alternatives considered in [Pan08] is that the underlying distribution was drawn from a particular *distribution over distributions* p which are ε-far in total variation distance from the uniform distribution (see Example 10.2.7). In our setting, the null hypothesis is that the underlying state is the maximally mixed state ρ_{mm}, and the mixture of alternatives will be a particular *distribution over quantum states* ρ which are ε-far in trace distance from ρ_{mm} (see Construction 2).

Note that in order to obtain dimension-dependent lower bounds, as in classical uniformity testing, it is essential that the alternative hypothesis be a mixture. If the task were instead to distinguish whether the underlying state was ρ_{mm} or some *specific* alternative state ρ, then if we make independent measurements in the eigenbasis of ρ, it takes only $O(1/\varepsilon^2)$ such measurements to tell apart the two scenarios.

For this reason we will be interested in the following abstraction which contains as special cases both Paninski's lower bound instance for uniformity testing [Pan08] and our lower bound instance for mixedness testing, and which itself is a special case of Le Cam's two-point method [LeC73]. We will do this in a few steps. First, we give a general formalism for what it means to perform possibly adaptive measurements:

Definition 10.2.1 (Adaptive measurements)**.** *Given an underlying space $\mathcal{S}$, a natural number $N \in \mathbb{N}$, and a (possibly infinite) universe $\mathcal{U}$ of measurement outcomes, a* measurement schedule *A using N measurements is any (potentially random) algorithm which outputs $M_1, \ldots, M_N : \mathcal{S} \to \mathcal{U}$, where each M_i is a potentially random function. We say that A is* nonadaptive *if the choice of M_i is independent of the choice of M_j for all $j \neq i$, and we say A is* adaptive *if the choice of M_t depends only on the outcomes of $M_1, \ldots, M_{t-1}$ for all $t \in [N]$.*

To instantiate this for the quantum setting, we let the underlying space $\mathcal{S}$ be the set of mixed states, and we restrict the measurement functions to be (possibly adaptively chosen) POVMs. See Definition 10.3.1 for a formal definition.

Recall the definition of a distinguishing task from Definition 1.2.35, reworded slightly here:

Definition 10.2.2. *A* distinguishing task *is specified by two disjoint sets $\mathcal{S}_0, \mathcal{S}_1$ in $\mathcal{S}$. For any $N \in \mathbb{N}$, and any measurement schedule A, we say that A solves the problem if there exists a (potentially random) post-processing algorithm $f : \mathcal{U}^N \to \{0, 1\}$ so that for any $\alpha \in \{0, 1\}$, if $D \in \mathcal{S}_\alpha$, then*

$$\Pr[f\left(M_1(D) \circ \cdots \circ M_N(D)\right) = \alpha] \geq 2/3 \ ,$$

where $M_1, \ldots, M_N$ are generated by A.

For instance, to instantiate the quantum mixedness testing setting, we let $\mathcal{S}$ be the set of mixed states, we let $\mathcal{S}_0 = \{\rho_{\mathsf{mm}}\}$ be the set containing only ρ_{mm}, the maximally mixed state, and we let $\mathcal{S}_1 = \{\rho : \|\rho - \rho_{\mathsf{mm}}\|_1 > \varepsilon\}$. Note that the choice of $2/3$ for the constant is arbitrary and can be replaced (up to constant factors in N) with any constant strictly larger than $1/2$. With this, we can now define our lower bound setup:

Definition 10.2.3 (Lower Bound Setup: Simple Null vs. Mixture of Alternatives)**.** *In the setting of Definition 10.2.2, a* distinguishing task *is specified by a null object $D_0 \in \mathcal{S}_0$, a set of alternate objects $\{D_\zeta\} \subseteq \mathcal{S}_1$ parametrized by ζ, and a distribution $\mathcal{D}$ over ζ.*

For any measurement schedule A which generates measurement functions $M_1, \ldots, M_N$, let $p_0^{\leq N} = p_0^{\leq N}(A)$ and $p_1^{\leq N} = p_1^{\leq N}(A)$ be distributions over strings $x_{\leq N} \in \mathcal{U}^N$, which we call transcripts of length N*. The distribution $p_0^{\leq N}$ corresponds to the distribution of $M_1(D_0) \circ \cdots \circ$*

$M_N(D_0)$. The distribution $p_1^{\leq N}$ corresponds to the distribution of of $M_1(D_\zeta) \circ \cdots \circ M_N(D_\zeta)$, where $\zeta \sim \mathcal{D}$.

The following is a standard result which allows us to relate this back to property testing:

Fact 10.2.4. *Let $\mathcal{S}_0, \mathcal{S}_1$ be a property, let $N \in \mathbb{N}$, and let $\mathcal{A}$ be a class of measurement schedules using N measurements. Suppose that there exists a distinguishing task so that for every $A \in \mathcal{A}$, we have that $d_{TV}(p_0^{\leq N}(A), p_1^{\leq N}(A)) \leq 1/3$. Then the distinguishing task cannot be solved with N samples by any algorithm in $\mathcal{A}$.*

For the remainder of the paper, we will usually implicitly fix a measurement schedule A, and just write $p_0^{\leq N}$ and $p_1^{\leq N}$. The properties that we assume (e.g. adaptive or nonadaptive) of this algorithm should be clear from context, if it is relevant.

We next define some important quantities which repeatedly arise in our calculations:

Definition 10.2.5. *In the setting of Definition 10.2.3, for any $t \in [N]$, define $p_0^t(\cdot | x_{<t}), p_1^t(\cdot | x_{<t})$ to be the respective conditional laws of the t-th entry, given preceding transcript $x_{<t}$. For any ζ, let $p_1^{\leq N} | \zeta$ be the distribution over transcripts from N independent observations from D_ζ.*

Assume additionally that $p_1^{\leq N} | \zeta$ are absolutely continuous with respect to $p_0^{\leq N}$, for every $\zeta \in \mathrm{supp}(\mathcal{D})$. Then, there will exist functions $\{g^\zeta_{x_{<t}}(\cdot)\}_{t \in [N], x_{<t} \in \mathcal{U}^{t-1}, \zeta \in \mathrm{supp}(\mathcal{D})}$, such that for any $\zeta, t, x_{\leq t}$, the Radon-Nikodym derivative satisfies

$$\frac{\mathrm{d}p_1^{\leq t} | \zeta}{\mathrm{d}p_0^{\leq t}}(x_{\leq t}) = \prod_{i=1}^{t} \left(1 + g^\zeta_{x_{<i}}(x_i)\right) . \tag{10.1}$$

We refer to the $g^\zeta_{x_{<t}}(\cdot)$ functions as likelihood ratio factors.

We emphasize that neither $p_0^{\leq N}$ nor any of the alternatives $p_1^{\leq N} | \zeta$ is necessarily a product measure. Indeed, this is one of the crucial difficulties of proving lower bounds in the adaptive setting. In the non-adaptive setting, the picture of Definition 10.2.3 simplifies substantially:

Definition 10.2.6 (Non-adaptive Testing Lower Bound Setup)**.** *In this case, in the notation of Definition 10.2.3, the measurement schedule A is nonadaptive, so $p_0^{\leq N}$ and all $p_1^{\leq N} | \zeta$ are product measures. Consequently, the functions $g^\zeta_{x_{<t}}$ will depend only on t and not on the particular transcript $x_{<t}$, so we will denote the functions by $\{g^\zeta_t(\cdot)\}_{t \in [N], \zeta \in \mathrm{supp}(\mathcal{D})}$.*

Paninski's lower bound for classical uniformity testing [Pan08] is an instance of the nonadaptive setup of Definition 10.2.6:

Example 10.2.7. *Let us first recall Paninski's construction. Here the set $\mathcal{S}$ is the set of distributions over $[d]$. Uniformity testing is the property $S_0 = \{U\}, S_1 = \{U' : d_{TV}(U, U') \geq \varepsilon\}$, where U is the uniform distribution over $[d]$. In the classical "sampling oracle" model of distribution testing, the measurements M_i simply take a distribution $D \in \mathcal{S}$ and output an independent sample from D. In particular, $\mathcal{U} = [d]$.*

To form Paninski's lower bound instance, take $\mathcal{D}$ to be the uniform distribution over $\{\pm 1\}^{d/2}$. Let the null hypothesis be D_0, and let the set of alternate hypotheses be given by $\{D_z\}_{z \in \{\pm 1\}^{d/2}}$, where D_z the distribution over $[d]$ whose x-th marginal is $D_z(x) = \frac{1}{d} + (-1)^x \cdot \frac{\varepsilon}{d} \cdot z_{\lceil x/2 \rceil}$ for any $x \in [d]$. Clearly $D_z \in \mathcal{S}_1$ for all z.

There is no obviously no adaptivity in what the tester does after seeing each new sample. So the family of likelihood ratio factors $\{g^z_t(\cdot)\}$ for which (10.1) holds is given by

$$g^z_t(x) = g^z(x) \triangleq \varepsilon(-1)^x \cdot z_{\lceil x/2 \rceil}. \tag{10.2}$$

The definition of $p_0^{\leq N}, p_1^{\leq N}$ in our proofs will be straightforward (see Construction 2), and by Fact 10.2.4, the key technical difficulty is to upper bound the total variation distance between $p_0^{\leq N}, p_1^{\leq N}$ in terms of N. In Section 10.2.1, we overview our approach for doing so in the non-adaptive setting of Definition 10.2.6, and in Section 10.2.2, we describe our techniques for extending these bounds to the generic, adaptive setting of Definition 10.2.3.

10.2.1 Non-Adaptive Lower Bounds

It is a standard trick to upper bound total variation distance between two distributions in terms of the χ^2-divergence, which is often more amenable to calculations. These calculations are especially straightforward in the non-adaptive setting of Definition 10.2.6 and is reminiscent of the so-called Ingster-Suslina method [IS12] for showing minimax bounds in classical settings (see Section 11.2).

Lemma 10.2.8. *Let $p_0^{\leq N}, p_1^{\leq N}, \mathcal{D}, \{g^\zeta_t(\cdot)\}_{t \in \mathbb{N}, \zeta \in \mathrm{supp}(\mathcal{D})}$ be defined as in Definition 10.2.6. As*

$p_0^{\leq N}$ is therefore a product measure, for every $t \in [N]$ denote its t-th marginal by p_0^t. Then

$$\frac{1}{2\ln 2} d_{TV}\left(p_1^{\leq N}, p_0^{\leq N}\right)^2 \leq \chi^2\left(p_1^{\leq N} \| p_0^{\leq N}\right) \leq \max_t \mathop{\mathbb{E}}_{\zeta,\zeta'}\left[\left(1 + \mathop{\mathbb{E}}_{x_t \sim p_0^t}\left[g_t^\zeta(x_t) g_t^{\zeta'}(x_t)\right]\right)^N\right] - 1.$$

Proof. The first inequality is just Pinsker's and the fact that chi-squared divergence upper bounds KL divergence. For the latter inequality, it will be convenient to define

$$g_S^\zeta(x_S) \triangleq \prod_{t \in S} g_t^\zeta(x_t).$$

Then for any ζ, ζ', S, the product structure implies

$$\mathop{\mathbb{E}}_{x_{\leq N} \sim p_0^{\leq N}}\left[g_S^\zeta(x_S) g_S^{\zeta'}(x_S)\right] = \prod_{t \in S} \mathop{\mathbb{E}}_{x_t \sim p_0^t}[g_t^\zeta(x_t) g_t^{\zeta'}(x_t)] \tag{10.3}$$

We then get that

$$\chi^2\left(p_1^{\leq N} \| p_0^{\leq N}\right) = \mathop{\mathbb{E}}_{x_{\leq N} \sim p_0^{\leq N}}\left[\left(\mathop{\mathbb{E}}_\zeta\left[\prod_{t=1}^N (1 + g_t^\zeta(x_t))\right] - 1\right)^2\right] = \mathop{\mathbb{E}}_{x_{\leq N}, \zeta, \zeta'}\left[\sum_{\emptyset \neq S, S' \subseteq [N]} g_S^\zeta(x_S) g_{S'}^{\zeta'}(x_{S'})\right]$$

$$\begin{aligned} &= \mathop{\mathbb{E}}_{x_{\leq N}, \zeta, \zeta'}\left[\sum_{S \neq \emptyset} g_S^\zeta(x_S) g_S^{\zeta'}(x_S)\right] = \mathop{\mathbb{E}}_{\zeta,\zeta'}\left[\prod_{t=1}^N \left(1 + \mathop{\mathbb{E}}_{x_t \sim p_0^t}\left[g_t^\zeta(x_t) g_t^{\zeta'}(x_t)\right]\right)\right] - 1 \\ &\leq \max_t \mathop{\mathbb{E}}_{\zeta,\zeta'}\left[\left(1 + \mathop{\mathbb{E}}_{x_t \sim p_0^t}\left[g_t^\zeta(x_t) g_t^{\zeta'}(x_t)\right]\right)^N\right] - 1, \end{aligned} \tag{10.4}$$

where the fourth step follows by (10.3), the last step follows by Holder's, and the third step follows by the fact that for $S \neq S'$ and any ζ, ζ',

$$\mathop{\mathbb{E}}_{x_{\leq N}}[g_S^\zeta(x_S) g_{S'}^{\zeta'}(x_{S'})] = \prod_{t \in S \cap S'} \mathop{\mathbb{E}}_{x_t}[g_t^\zeta(x_t) g_t^{\zeta'}(x_t)] \cdot \prod_{t \in S \backslash S'} \mathop{\mathbb{E}}_{x_t}[g_t^\zeta(x_t)] \cdot \prod_{t \in S' \backslash S} \mathop{\mathbb{E}}_{x_t}[g_t^{\zeta'}(x_t)] = 0,$$

□

The upshot of (10.4) is that the fluctuations of the quantities $\mathbb{E}_{x_t}[g_t^\zeta(x_t) g_t^{\zeta'}(x_t)]$ with respect to the randomness of ζ, ζ' dictate how large N must be for $p_0^{\le N}$ and $p_1^{\le N}$ to be distinguishable.

Example 10.2.9. *Recalling (10.2), the quantities $\mathbb{E}_{x_t}[g_t^{\zeta}(x_t)g_t^{\zeta'}(x_t)]$ take a particularly nice form in Paninski's setting. There we have*

$$\mathop{\mathbb{E}}_{x_t}[g_t^{\zeta}(x_t)g_t^{\zeta'}(x_t)] = \varepsilon^2 \cdot \mathop{\mathbb{E}}_{x\sim[d]}\left[z_{\lceil x/2\rceil} \cdot z'_{\lceil x/2\rceil}\right] = \frac{\varepsilon^2}{d}\sum_{x=1}^{d} \mathbb{1}[z_{\lceil x/2\rceil} = z'_{\lceil x/2\rceil}] = \frac{2\varepsilon^2}{d}\langle z, z'\rangle \tag{10.5}$$

Because $\langle z, z'\rangle$ is distributed as a shifted, rescaled binomial distribution, $\mathbb{E}_{x_t}[g_t^{\zeta}(x_t)g_t^{\zeta'}(x_t)]$ has sub-Gaussian tails and fluctuations of order $O(\varepsilon^2/\sqrt{d})$, implying that for N as large as $o(\sqrt{d}/\varepsilon^2)$, $\chi^2\left(p_1^{\le N} \| p_0^{\le N}\right) = o(1)$. While this is not exactly how Paninski's lower bound was originally proven, concentration of the binomial random variable $\langle z, z'\rangle$ lies at the heart of the lower bound and formalizes the usual intuition for the $\sqrt{d}$ scaling in the lower bound: to tell whether a distribution is far from uniform, it is necessary to draw $\Omega(\sqrt{d})$ samples just to see some element of $[d]$ appear twice.

In Section 10.4, we will show how to use Lemma 10.2.8 to prove Theorem 10.1.1. As it turns out, understanding the fluctuations of the random variable $\mathbb{E}_{x_t}[g_t^{\zeta}(x_t)g_t^{\zeta'}(x_t)]$ that arises in that setting will be one of the primary technical challenges of this work, both for our adaptive and non-adaptive lower bounds (see Section 10.7).

10.2.2 Adaptive Lower Bounds

As was discussed previously and is evident from the proof of Lemma 10.2.8, the lack of product structure for $p_0^{\le N}$ and $p_1^{\le N}|\zeta$ in the adaptive setting of Definition 10.2.3 makes it infeasible to directly estimate $\chi^2\left(p_1^{\le N} \| p_0^{\le N}\right)$. Inspired by the literature on bandit lower bounds [ACBFS02,BCB12], we instead upper bound $\mathrm{KL}\left(p_1^{\le N} \| p_0^{\le N}\right)$, for which we can appeal to the chain rule to tame the extra power afforded by adaptivity. To handle the mixture structure of $p_1^{\le N}$, we will upper bound each of the resulting *conditional* KL divergence terms by their corresponding conditional χ^2 divergence.

First, we introduce some notation essential to the calculations in this work.

Definition 10.2.10 (Key Quantities). *In the generic setup of Definition 10.2.3, for any*

$x_{\leq t} \in \mathcal{U}^t$, define

$$\Delta(x_{\leq t}) \triangleq \frac{\mathrm{d}p_1^{\leq t}}{\mathrm{d}p_0^{\leq t}}(x_{\leq t}),\ \phi^{\zeta,\zeta'}_{x_{\leq t}} \triangleq \mathop{\mathbb{E}}_{x\sim p_0^t(\cdot|x_{\leq t})}\Big[g^{\zeta}_{x_{\leq t}}(x)g^{\zeta'}_{x_{\leq t}}(x)\Big],\ \Psi^{\zeta,\zeta'}_{x_{\leq t}} \triangleq \prod_{i=1}^{t}(1+g^{\zeta}_{x_{<i}}(x_i))(1+g^{\zeta'}_{x_{<i}}(x_i)) \tag{10.6}$$

The following is a key technical ingredient of this work.

Lemma 10.2.11. *Let $p_0^{\leq N}, p_1^{\leq N}, \mathcal{D}, \{g^{\zeta}_{x_{<t}}(\cdot)\}$ be defined as in Definition 10.2.3. Then*

$$\frac{1}{2\ln 2}d_{TV}\left(p_0^{\leq N}, p_1^{\leq N}\right)^2 \leq KL\left(p_1^{\leq N}\|p_0^{\leq N}\right) \leq \sum_{t=1}^{N}\mathop{\mathbb{E}}_{x_{<t}\sim p_0^{\leq t-1}}\left[\frac{1}{\Delta(x_{<t})}\mathop{\mathbb{E}}_{\zeta,\zeta'\sim\mathcal{D}}\Big[\phi^{\zeta,\zeta'}_{x_{<t}}\cdot\Psi^{\zeta,\zeta'}_{x_{<t}}\Big]\right].$$

Proof. The first inequality is Pinsker's. For the second, by the chain rule for KL divergence and the fact that chi-squared divergence upper bounds KL, $\mathrm{KL}\left(p_1^{\leq (N)}\|p_0^{\leq N}\right)$ can be written as

$$\sum_{t=1}^{N}\mathop{\mathbb{E}}_{x_{<t}\sim p_1^{\leq t-1}}\left[\mathrm{KL}\left(p_1^t(\cdot|x_{<t})\|p_0^t(\cdot|x_{<t})\right)\right] \leq \sum_{t=1}^{N}\mathop{\mathbb{E}}_{x_{<t}\sim p_1^{\leq t-1}}\left[\chi^2\left(p_1^t(\cdot|x_{<t})\|p_0^t(\cdot|x_{<t})\right)\right].$$

By definition, the conditional densities $p_0^t(\cdot|x_{<t}), p_1^t(\cdot|x_{<t})$ satisfy

$$p_i^t(x_t|x_{<t}) = \frac{p_i^{\leq t}(x_{<t}\circ x_t)}{p_i^{\leq t-1}(x_{<t})} \quad \text{for } i = 0,1. \tag{10.7}$$

Therefore, we have:

$$\begin{aligned}\mathop{\mathbb{E}}_{x_{<t}\sim p_1^{\leq t-1}}\left[\chi^2\left(p_1^t(\cdot|x_{<t})\|p_0^t(\cdot|x_{<t})\right)\right] &= \mathop{\mathbb{E}}_{x_{<t}\sim p_1^{\leq t-1}}\left[\mathop{\mathbb{E}}_{x_t\sim p_0^t(\cdot|x_{<t})}\left[\left(\frac{\Delta(x_{<t}\circ x_t)}{\Delta(x_{<t})}-1\right)^2\right]\right]\\ &= \mathop{\mathbb{E}}_{x_{<t}\sim p_1^{\leq t-1}}\left[\frac{1}{\Delta(x_{<t})^2}\mathop{\mathbb{E}}_{x_t\sim p_0^t(\cdot|x_{<t})}\left[(\Delta(x_{<t}\circ x_t)-\Delta(x_{<t}))^2\right]\right]\\ &= \mathop{\mathbb{E}}_{x_{<t}\sim p_0^{\leq t-1}}\left[\frac{1}{\Delta(x_{<t})}\mathop{\mathbb{E}}_{x_t\sim p_0^t(\cdot|x_{<t})}\left[(\Delta(x_{<t}\circ x_t)-\Delta(x_{<t}))^2\right]\right]\end{aligned} \tag{10.8}$$

where the first step follows by (10.7) and the third step follows by a change of measure in the outer expectation.

By the assumption (10.1) and the definition of $\Delta(\cdot)$,

$$\Delta(x_{<t}) = \mathop{\mathbb{E}}_{\zeta}\left[\prod_{i=1}^{t-1}(1 + g^{\zeta}(x_i))\right]. \tag{10.9}$$

This yields

$$\begin{aligned}
\mathop{\mathbb{E}}_{x_t \sim p_0^t(\cdot|x_{<t})}\left[(\Delta(x_{<t} \circ x_t) - \Delta(x_{<t}))^2\right] &= \mathop{\mathbb{E}}_{x_t \sim p_0^t(\cdot|x_{<t})}\left[\mathop{\mathbb{E}}_{\mathcal{D}}\left[\prod_{i=1}^{t-1}(1 + g^{\zeta}_{x_{<i}}(x_i)) \cdot g^{\zeta}_{x_{<t}}(x_t)\right]^2\right] \\
&= \mathop{\mathbb{E}}_{\zeta,\zeta'}\left[\mathop{\mathbb{E}}_{x_t}\left[g^{\zeta}_{x_{<t}}(x_t) g^{\zeta'}_{x_{<t}}(x_t)\right]\prod_{i=1}^{t-1}(1 + g^{\zeta}_{x_{<i}}(x_i))(1 + g^{\zeta'}_{x_{<i}}(x_i))\right] \\
&= \mathop{\mathbb{E}}_{\zeta,\zeta'}\left[\phi^{\zeta,\zeta'}_{x_{<t}} \cdot \Psi^{\zeta,\zeta'}_{x_{<t}}\right],
\end{aligned}$$

from which the lemma follows by (10.8). □

10.3 Unentangled Measurements and Lower Bound Instance

In this section we provide some preliminary notions and calculations that are essential to understanding the proofs of Theorem 10.1.1 and 10.1.2. We first formalize the notion of quantum property testing with unentangled, possibly adaptive measurements in Section 10.3.1. Then in Section 10.3.2, we give our lower bound construction and instantiate it in the generic setup of Definition 10.2.3. Finally, in Section 10.3.3, we give some intuition for some of the key quantities that arise.

10.3.1 Testing with Unentangled Measurements

Definition 10.3.1. *Let* $N \in \mathbb{N}$. *An* unentangled, possibly adaptive POVM schedule $\mathcal{S}$ *is a type of measurement schedule specified by a (possibly infinite) collection of POVMs* $\{\mathcal{M}^{x_{<t}}\}_{t\in[N],x_{<t}\in\mathcal{T}_t}$ *where* $\mathcal{T}_1 \triangleq \{\emptyset\}$*, and for every* $t > 1$*,* $\mathcal{T}_t$ *denotes the set of all possible* *transcripts of measurement outcomes $x_{<t}$ for which $x_i \in \Omega(\mathcal{M}^{x_{<i}})$ for all $1 \le i \le t-1$ (recall that $x_{<i} \triangleq (x_1, ..., x_{i-1})$). The schedule works in the natural manner: at time t for $t = 1, \dots, N$, given a transcript $x_{<t} \in \mathcal{T}_t$, it measures the t-th copy of ρ using the POVM $\mathcal{M}^{x_{<t}}$.*

If in addition the resulting schedule is also a nonadaptive measurement schedule, we say it is an ℓ-entangled, nonadaptive POVM schedule and denote it simply by $\{\mathcal{M}^t\}_{t\in[N]}$.

10.3.2 Lower Bound Instance

Let $\mathcal{D}$ be the Haar measure over the unitary group $U(d)$. In place of ζ from Definition 10.2.3, we will denote elements from $\mathcal{D}$ by $\mathbf{U}$. $\Pr_{\mathbf{U}}[\cdot]$ and $\mathbb{E}_{\mathbf{U}}[\cdot]$ will be with respect to $\mathcal{D}$ unless otherwise specified.

Construction 2. *Let $\mathbf{X} \in \mathbb{R}^{d\times d}$ denote the diagonal matrix whose first $d/2$ diagonal entries are equal to ε, and whose last $d/2$ diagonal entries are equal to $-\varepsilon$. Let $\mathbf{X}' \triangleq \frac{1}{\varepsilon}\mathbf{X}$. Let $\mathbf{\Lambda} \triangleq \frac{1}{d}(Id + \mathbf{X})$.*

Our lower bound instance will be the distribution over densities $\mathbf{U}^\dagger\mathbf{\Lambda}\mathbf{U}$ for $\mathbf{U} \sim \mathcal{D}$. We remark that this instance, the quantum analogue of Paninski's lower bound instance [Pan08] for classial uniformity testing, has appeared in various forms throughout the quantum learning and testing literature [OW15, Wri16, HHJ+17].

Given $N \in \mathbb{N}$, define $\boldsymbol{\rho}_0^{\le N} \triangleq \rho_{\mathsf{mm}}^{\otimes N}$ and $\boldsymbol{\rho}_1^{\le N} \triangleq \mathbb{E}_{\mathbf{U}\sim\mathcal{D}}[(\mathbf{U}^\dagger\mathbf{\Lambda}\mathbf{U})^{\otimes N}]$. Take any POVM schedule $\mathcal{S} = \{\mathcal{M}^{x_{<t}}\}_{t\in[N], x_{<t}\in\mathcal{T}_t}$. Given $t \le N$, define $p_0^{\le t}$ and $p_1^{\le t}$ to be the distribution over the measurement outcomes when the first t steps of these POVM schedules are applied to the first t parts of $\boldsymbol{\rho}_0^{\le N}$ and $\boldsymbol{\rho}_1^{\le N}$ respectively. Equivalently, $p_1^{\le t}$ can be regarded as the distribution over sequences of t measurement outcomes arising from first sampling $\mathbf{U}$ according to the Haar measure $\mathcal{D}$ and then applying the first t steps of POVM schedule $\mathcal{S}$ to t copies of $\rho \triangleq \mathbf{U}^\dagger\mathbf{\Lambda}\mathbf{U}$.

Lemma 10.3.2. *For any POVM $\mathcal{M}$, define*

$$g^{\mathbf{U}}_{\mathcal{M}}(x) \triangleq \langle \widehat{M}^{x_{<t}}_x, \mathbf{U}^\dagger\mathbf{X}\mathbf{U}\rangle. \tag{10.10}$$

$p_1^{\leq N}$ is absolutely continuous with respect to $p_0^{\leq N}$, and the family of likelihood ratio factors $\{g^{\mathbf{U}}_{x_{<t}}(\cdot)\}$ for which (10.1) holds for $p_0^{\leq N}$ and $p_1^{\leq N}$ defined in Construction 2 is given by $g^{\mathbf{U}}_{x_{<t}}(\cdot) \triangleq g^{\mathbf{U}}_{\mathcal{M}^{x_{<t}}}$.

Proof. By taking a disjoint union over $\Omega(\mathcal{M}^{x_{<t}})$ for all $t \in \mathbb{N}$ and transcripts $x_{<t}$, we can assume without loss of generality that there is some space Ω^* for which $\Omega(\mathcal{M}^{x_{<t}})$ is a subspace of Ω^* for every $t, x_{<t}$. For the product space $(\Omega^*)^N$, equip the t-th factor with the σ-algebra given by the join of all σ-algebras associated to $\Omega(\mathcal{M}^{x_{\leq t}})$ for transcripts $x_{\leq t}$ of length t.

Then the measures μ in Definition 1.3.47 for all POVMs $\mathcal{M}^{x_{<t}}$ induce a measure μ^* over $(\Omega^*)^N$. Moreover, by definition, $p_0^{\leq N}$ and $p_1^{\leq N}$ correspond to probability measures over $(\Omega^*)^t$ which are absolutely continuous with respect to μ^*.

Because $\langle M_x, \rho_{\mathsf{mm}}\rangle > 0$ for any nonzero psd Hermitian matrix M_x, absolute continuity of $p_1^{\leq N}$ with respect to $p_0^{\leq N}$ follows immediately.

By the chain rule for Radon-Nikodym derivatives, we conclude that

$$\frac{\mathrm{d}p_1^{\leq t}|\mathbf{U}}{\mathrm{d}p_0^{\leq t}}(x_{\leq t}) = \frac{\prod_{i=1}^t \langle M^{x_{<i}}_{x_i}, \mathbf{U}^\dagger \mathbf{\Lambda} \mathbf{U}\rangle}{\prod_{i=1}^t \frac{1}{d}\operatorname{Tr}(M^{x_{<i}}_{x_i})} = \prod_{i=1}^t \langle \widehat{M}^{x_{<i}}_{x_i}, \mathbf{U}^\dagger(\mathrm{Id} + \mathbf{X})\mathbf{U}\rangle = \prod_{i=1}^t (1 + g^{\mathbf{U}}_{x_{<i}}(x_i))$$

as claimed. □

For any $\mathbf{U}, \mathbf{U}' \in U(d)$, the quantities $\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}}$ and $\phi^{\mathbf{U},\mathbf{U}'}_{x_{<t}}$ are given by (10.6). Given a POVM $\mathcal{M}$, also define $\phi^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}$ in the obvious way. Lastly, we record the following basic facts:

Fact 10.3.3. *For any POVM $\mathcal{M}$,*

(I) $\mathbb{E}_{x\sim p}[g^{\mathbf{U}}_{\mathcal{M}}(x)] = 0$ for any $\mathbf{U} \in U(d)$.

(II) For any measurement outcome x and $\mathbf{U}, \mathbf{U}' \in U(d)$, $|g^{\mathbf{U}}_{\mathcal{M}}(x)| \leq \varepsilon$ and thus $\phi^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}} \leq \varepsilon^2$.

10.3.3 Intuition for $\phi^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}$

Recall from Example 10.2.9 that for classical uniformity testing, $\phi^{z,z'} = \frac{2\varepsilon^2}{d}\langle z, z'\rangle$, and by Lemma 10.2.8, the $O(\varepsilon^2/\sqrt{d})$ fluctuations of $\phi^{z,z'}$ as a random variable in z, z' precisely dictate the sample complexity of uniformity testing.

One should therefore think of the distribution of the quantity $\phi_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}$ as a "quantum analogue" of the binomial distribution whose fluctuations are closely related to the scaling of the copy complexity of mixedness testing.

As we will show in Theorem 10.4.1, $\phi_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}$ has $O(\varepsilon^2/d^{3/2})$ fluctuations and concentrates well, from which it will follow by integration by parts that N can be taken as large as $o(d^{3/2}/\varepsilon^2)$, yielding the lower bound of Theorem 10.1.1.

To get some intuition for where these $O(\varepsilon^2/d^{3/2})$ fluctuations come from, suppose $\mathcal{M}$ were the orthogonal POVM given by the standard basis. Then

$$\phi_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'} = \frac{1}{d}\sum_{i=1}^{d}\left\langle \mathrm{diag}(\mathbf{U}^\dagger\mathbf{X}\mathbf{U}), \mathrm{diag}(\mathbf{U}'^\dagger\mathbf{X}\mathbf{U}')\right\rangle = \frac{1}{d}\sum_{i=1}^{d}\varepsilon^2\cdot\delta(\mathbf{U}_i)\cdot\delta(\mathbf{U}'_i),$$

where

$$\delta(v) \triangleq \sum_{i=1}^{d/2} v_i^2 - \sum_{i=d/2+1}^{d} v_i^2. \tag{10.11}$$

For any fixed i, $\mathbf{U}_i, \mathbf{U}'_i$ are independent random unit vectors, and the variance of $\delta(\mathbf{U}_i)\cdot\delta(\mathbf{U}'_i)$ is $O(1/d^2)$ (see Fact 10.7.2). If $\mathbf{U}_1, \mathbf{U}'_1 ..., \mathbf{U}_d, \mathbf{U}'_d$ were all independent, then $\phi_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}$ would thus have variance ε^4/d^3, suggesting $O(\varepsilon^2/d^{3/2})$ fluctuations as claimed. Of course we do not actually have this independence assumption; in addition, the other key technical challenges we must face to get Theorem 10.4.1 are 1) to go beyond just a second moment bound and show sufficiently strong concentration of $\phi_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}$, and 2) to show this is the case for *all POVMs*. We do this in Section 10.7.

10.4 Proof of Non-Adaptive Lower Bound

In this section we prove Theorem 10.1.1 by applying Lemma 10.2.8; the technical crux of the proof (and of our proof of Theorem 10.1.2 in the next section) is the following tail bound, whose proof we defer to Section 10.7:

Theorem 10.4.1. *Fix any POVM $\mathcal{M}$. There exists an absolute constant $c'' > 0$ such that* *for any $t > \Omega(\varepsilon^2/d^{1.99})$, we have*

$$\Pr_{\mathbf{U},\mathbf{U}'\sim\mathcal{D}}\left[\left|\phi_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}\right| > t\right] \le \exp\left(-c''\left\{\frac{d^3t^2}{\varepsilon^4} \wedge \frac{d^2t}{\varepsilon^2}\right\}\right)$$

Proof of Theorem 10.1.1. By Fact 10.2.4, it suffices to show that no nonadaptive POVM schedule can solve the distinguishing task given by Construction 2, unless $N = \Omega(d^{3/2}/\varepsilon^2)$. For a non-adaptive POVM schedule $\mathcal{S}$, let $\{\mathcal{M}^1, ..., \mathcal{M}^N\}$ denote the sequence of POVMs that are used. Recalling (10.10), the likelihood ratio factors $\{g_t^{\mathbf{U}}(\cdot)\}_{\mathbf{U}\in U(d),t\in[N]}$ for which (10.1) holds in the nonadaptive setting of Definition 10.2.6 are given by $g_{\mathcal{M}^t}^{\mathbf{U}}(\cdot)$. Similarly, denote $\phi_{x_{<t}}^{\mathbf{U},\mathbf{U}'}$ by $\phi_t^{\mathbf{U},\mathbf{U}'}$.

By Lemma 10.2.8, we have

$$\frac{1}{2\ln 2}d_{\mathrm{TV}}\left(p_1^{\le N}, p_0^{\le N}\right)^2 \le \max_t \mathop{\mathbb{E}}_{\zeta,\zeta'}\left[\left(1+\phi_t^{\mathbf{U},\mathbf{U}'}\right)^N\right] - 1.$$

To finish the proof, we will show that

$$\sup_{\mathcal{M}} \mathop{\mathbb{E}}_{\mathbf{U},\mathbf{U}'}\left[\left(1+\phi_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}\right)^N\right] = 1 + o(1)$$

for $N = o(d^{3/2}/\varepsilon^2)$, from which the proof is complete by (10.4).

We would like to apply integration by parts (Fact 1.3.30) to the random variable $Z \triangleq 1 + \phi_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}$ and the function $f(Z) \triangleq Z^N$. By Part (II) of Fact 10.3.3, this random variable is supported in $[1-\varepsilon^2, 1+\varepsilon^2]$. We can take the parameters in Fact 1.3.30 as follows: set $a \triangleq 1 + \varepsilon/(N^{1/2}d^{3/4})$, $b \triangleq 1+\varepsilon^2$, and tail bound function $\tau(x) = \exp\left(-c''\left\{\frac{d^3(x-1)^2}{\varepsilon^4} \wedge \frac{d^2(x-1)}{\varepsilon^2}\right\}\right)$. Note that for $N = o(d^{3/2}/\varepsilon^2)$, $(1+\tau(a))f(a) = 1+o(1)$. So by Fact 1.3.30 and Theorem 10.4.1,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{U}'}\left[\left(1+\phi_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}\right)^N\right] \\
&\quad\le 1 + o(1) + \int_{1+\varepsilon^2/d^{3/2}}^{1+\varepsilon^2} Nx^{N-1}\cdot\exp\left(-c''\left\{\frac{d^3(x-1)^2}{\varepsilon^4} \wedge \frac{d^2(x-1)}{\varepsilon^2}\right\}\right)\,\mathrm{d}x \\
&\quad\le 1 + o(1) + \int_{\varepsilon^2/d^{3/2}}^{\varepsilon^2} N(1+x)^{N-1}\left(\exp\left(-\frac{c''d^3x^2}{\varepsilon^4}\right) + \exp\left(-\frac{c''d^2x}{\varepsilon^2}\right)\right)\,\mathrm{d}x \\
&\quad\le 1 + o(1) + \int_{0}^{\infty} N(1+x)^{N-1}\left(\exp\left(-\frac{c''d^3x^2}{\varepsilon^4}\right) + \exp\left(-\frac{c''d^2x}{\varepsilon^2}\right)\right)\,\mathrm{d}x
\end{aligned}$$

$$= 1 + o(1) + (N/2)!(c''d^3/\varepsilon^4)^{-N/2} + N!(c''d^2/\varepsilon^2)^N = 1 + o(1),$$

where the final step uses that $N = o(d^{3/2}/\varepsilon^2)$. □

10.5 A Chain Rule Proof of Paninski's Theorem

As discussed previously, the proof of Theorem 10.1.1 completely breaks down when the POVM schedule $\mathcal{S}$ is adaptive, so we will instead use the chain rule, via Lemma 10.2.11, to prove Theorem 10.1.2.

As a warmup, in this section we will show how to use Lemma 10.2.11 to prove a lower bound for *classical* uniformity testing. As it turns out, it is possible to recover Paninski's optimal $\Omega(\sqrt{d}/\varepsilon^2)$ lower bound with this approach, the details of which we give in Appendix 10.8, but in this section we opt to present a proof which achieves a slightly weaker bound. The reason is that in our proof of Theorem 10.5.1, we will make minimal use of the kind of precise cancellations that would yield a tight bound but which, unfortunately, are specific to the product structure of the distribution of random signs z. As such, these steps will be general-purpose enough to extend to the quantum setting where the Haar measure over $U(d)$ enjoys no such product structure.

Specifically, we will use the chain rule to show the following:

Theorem 10.5.1 (Weaker Paninski Theorem). $\Omega(d^{1/3}/\varepsilon^2)$ *samples are necessary to test whether a distribution p is ε-far from the uniform distribution.*

In this section, let $p_0^{\leq N}, p_1^{\leq N}$ denote the distributions defined in Example 10.2.7. Recalling the notation from Example 10.2.7 and Definition 10.2.10, as well as the identities (10.2) and (10.5), we immediately get the following from Lemma 10.2.11:

Lemma 10.5.2.

$$KL\left(p_1^{\leq N} \| p_0^{\leq N}\right) \leq \sum_{t=1}^{N} Z_t \quad \textit{for} \quad Z_t \triangleq \mathop{\mathbb{E}}_{x_{<t} \sim U^{\otimes t-1}}\left[\frac{1}{\Delta(x_{<t})} \mathop{\mathbb{E}}_{z,z' \sim \{\pm 1\}^{d/2}}\left[\phi^{z,z'} \cdot \Psi^{z,z'}_{x_{<t}}\right]\right]. \quad (10.12)$$

We will also need the following two estimates (see below for their proofs).

Lemma 10.5.3. *For any transcript $x_{<t}$, $\Delta(x_{<t}) \geq (1-\varepsilon^2)^{(t-1)/2}$.*

Lemma 10.5.4. *For any $z, z' \in \{\pm 1\}^{d/2}$, $\mathbb{E}_{x_{<t} \sim U^{\otimes t-1}}[(\Psi^{z,z'}(x_{<t}))^2] \leq (1+O(\varepsilon^2))^{t-1}$.*

We now describe how to use these to bound the summands Z_t in (10.12). As discussed in Example 10.2.9, $\phi^{z,z'} = \frac{2\varepsilon^2}{d}\langle z, z'\rangle$ has $O(\varepsilon^2/\sqrt{d})$ fluctuations. If we pretended $\phi^{z,z'}$ was of this magnitude with probability one, then

$$Z_t \approx O(\varepsilon^2/\sqrt{d}) \cdot \mathop{\mathbb{E}}_{x_{<t} \sim U^{\otimes t-1}}\left[\frac{1}{\Delta(x_{<t})} \mathop{\mathbb{E}}_{z,z' \sim \{\pm 1\}^{d/2}}\left[\Psi^{z,z'}_{x_{<t}}\right]\right] = O(\varepsilon^2/\sqrt{d}),$$

where the last step follows because $\Delta(x_{<t})^2 = \mathbb{E}_{z,z'}[\Psi^{z,z'}_{x_{<t}}]$ and the likelihood ratio between two distributions always integrates to 1. Then by (10.12) we would in fact even recover Theorem 10.1.3.

Unfortunately, in reality $\phi^{z,z'}$ can be as large as order ε^2, albeit with exponentially small probability, so instead we will partition the space of $z, z' \in \{\pm 1\}^{d/2}$ into those for which $\phi^{z,z'}$ is either less than some threshold τ or greater. When $\phi^{z,z'} \leq \tau$, we can bound the total contribution to Z_t of such z, z' by τ. When $\phi^{z,z'} > \tau$, we will use the pointwise estimates from Lemmas 10.5.3 and 10.5.4 and argue that because $\Pr[\phi^{z,z'} > \tau]$ is so small, these z, z' contribute negligibly to Z_t. The reason we only get an $\Omega(d^{1/3}/\varepsilon^2)$ lower bound in the end is that we must take τ slightly larger than the fluctuations of $\phi^{z,z'}$ to balance the low probability of $\phi^{z,z'}$ exceeding τ with the pessimistic pointwise estimates of Lemmas 10.5.3 and 10.5.4.

Proof of Theorem 10.5.1. We fill in the details of the strategy outlined above. We will use Fact 10.2.4 with the construction in Example 10.2.7. Given a transcript $x_{<t}$ and $z, z' \in \{\pm 1\}^{d/2}$, let $\mathbb{1}[\mathcal{E}^{z,z'}(\tau)]$ denote the indicator of whether $\phi^{z,z'} > \tau$. We have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{z,z'}\left[\Psi^{z,z'}_{x_{<t}} \cdot \phi^{z,z'}\right] &= \mathop{\mathbb{E}}_{z,z'}\left[\Psi^{z,z'}_{x_{<t}} \cdot \phi^{z,z'} \cdot \left(\mathbb{1}[\mathcal{E}^{z,z'}(\tau)] + \mathbb{1}[\mathcal{E}^{z,z'}(\tau)^c]\right)\right] \\
&\leq \varepsilon^2 \cdot \mathop{\mathbb{E}}_{z,z'}\left[\Psi^{z,z'}_{x_{<t}} \cdot \mathbb{1}[\mathcal{E}^{z,z'}(\tau)]\right] + \tau \cdot \mathop{\mathbb{E}}_{z,z'}\left[\Psi^{z,z'}_{x_{<t}} \cdot \mathbb{1}[\mathcal{E}^{z,z'}(\tau)^c]\right] \\
&\leq \varepsilon^2 \cdot \underbrace{\mathop{\mathbb{E}}_{z,z'}\left[\Psi^{z,z'}_{x_{<t}} \cdot \mathbb{1}[\mathcal{E}^{z,z'}(\tau)]\right]}_{\text{Ⓑ}_{x_{<t}}} + \tau \cdot \underbrace{\mathop{\mathbb{E}}_{z,z'}\left[\Psi^{z,z'}_{x_{<t}}\right]}_{\text{Ⓖ}_{x_{<t}}},
\end{aligned}$$

where in the second step we used Part (II) of Fact 10.3.3. Note that for any transcript $x_{<t}$,

$\Delta(x_{<t})^2 = \mathbb{E}_{z,z'}[\Psi^{z,z'}_{x_{<t}}] = \textcircled{G}_{x_{<t}}$, so by this and the fact that the likelihood ratio between two distributions always integrates to 1,

$$\mathop{\mathbb{E}}_{x_{<t}\sim U^{\otimes t-1}}\left[\frac{1}{\Delta(x_{<t})}\cdot \textcircled{G}_{x_{<t}}\right] = \mathop{\mathbb{E}}_{x_{<t}}[\Delta(x_{<t})] = 1. \tag{10.13}$$

We conclude that

$$\begin{aligned} Z_t &\le \varepsilon^2 \cdot \mathop{\mathbb{E}}_{x_{<t}\sim U^{\otimes t-1}}\left[\frac{1}{\Delta(x_{<t})}\cdot \textcircled{B}_{x_{<t}}\right] + \tau \cdot \mathop{\mathbb{E}}_{x_{<t}\sim U^{\otimes t-1}}\left[\frac{1}{\Delta(x_{<t})}\cdot \textcircled{G}_{x_{<t}}\right] \\ &\le \varepsilon^2 \cdot (1+\varepsilon^2)^{(t-1)/2} \mathop{\mathbb{E}}_{x_{<t}}[\textcircled{B}_{x_{<t}}] + \tau, \end{aligned}$$

where the second step follows by Lemma 10.5.3 and (10.13). It remains to show that τ is the dominant quantity above, for appropriately chosen τ.

Pick $\tau = \Omega(\varepsilon^2/d^{1/3})$. To upper bound $\mathbb{E}_{x_{<t}}[\textcircled{B}_{x_{<t}}]$, first apply Cauchy-Schwarz to get

$$\begin{aligned} \mathop{\mathbb{E}}_{x_{<t}}[\textcircled{B}_{x_{<t}}] &\le \mathop{\mathbb{E}}_{x_{<t},z,z'}\left[\left(\Psi^{z,z'}_{x_{<t}}\right)^2\right]^{1/2} \cdot \mathop{\Pr}_{x_{<t},z,z'}\left[\phi^{z,z'} > \tau\right]^{1/2} \\ &\le (1+O(\varepsilon^2))^{(t-1)/2} \cdot \exp(-\Omega(d^{1/3})), \end{aligned}$$

where the second step follows by Lemma 10.5.4, (10.5), and standard binomial tail bounds. For $t = o(d^{1/3}/\varepsilon^2)$, this quantity is indeed negligible, concluding the proof that $\textcircled{*} \le O(\varepsilon^2/d^{1/3})$ and, by Lemma 10.5.2, that $\chi^2\left(p_1^{\le N} \| p_0^{\le N}\right) = o(1)$ for $N = o(d^{1/3}/\varepsilon^2)$. □

Deferred Proofs

Proof of Lemma 10.5.3. For any $x_{<t} \in [d]^{t-1}$, we have that

$$\begin{aligned} \mathop{\mathbb{E}}_{z}\left[\prod_{i=1}^{t-1}(1+g^z(x_i))\right] &\ge \left(\prod_{z\in\{\pm1\}^{d/2}}\prod_{i=1}^{t-1}(1+g^z(x_i))\right)^{2^{-d/2}} \\ &= \left(\prod_{z\in\{\pm1\}^{d/2}}\prod_{i=1}^{t-1}(1+g^z(x_i))^{1/2}(1+g^{-z}(x_i))^{1/2}\right)^{2^{-d/2}} \\ &= (1-\varepsilon^2)^{(t-1)/2}, \end{aligned}$$

where in the first step we used AM-GM, in the second step we used the fact that if z is chosen uniformly at random from $\{\pm 1\}^{d/2}$, then $-z$ is also distributed according to the uniform distribution over $\{\pm 1\}^{d/2}$, and in the third step we used that for any x, $(1+g^z(x))(1+g^{-z}(x)) = 1-\varepsilon^2$. □

Proof of Lemma 10.5.4. Note that by both parts of Fact 10.3.3,

$$\mathop{\mathbb{E}}_{x\sim U}[(1+g^z(x))^2(1+g^{z'}(x))^2] = 1 + \mathop{\mathbb{E}}_{x}[g^z(x)g^{z'}(x)] \leq 1 + O(\varepsilon^2).$$

Writing

$$\mathop{\mathbb{E}}_{x_{<t}}[\Psi^{z,z'}_{x_{<t}}] \leq \mathop{\mathbb{E}}_{x_{<t-1}}[\Psi^{z,z'}_{x_{<t-1}}] \cdot (1 + O(\varepsilon^2)),$$

we see that the claim follows by induction on t. □

Parallels to Proof of Theorem 10.1.2 Lastly, we comment on how these ingredients carry over to our proof of Theorem 10.1.2. Lemma 10.5.2 translates verbatim to the quantum setting (see Lemma 10.6.1), as does the final part of the proof where we partition based on the value of $\phi^{z,z'}$.

Lemma 10.6.2 will be the quantum analogue of Lemma 10.5.3, and its proof uses a similar trick of AM-GM plus averaging with an involution.

Lemma 10.6.4 will be the quantum analogue of Lemma 10.5.4. Unfortunately, as we will see later in Section 10.6, an analogously naive bound will not suffice in our proof of Theorem 10.1.2. The workaround is somewhat technical, and we defer the details to Lemma 10.6.4 and the discussion preceding it.

Finally, as in Section 10.2.1, the central technical ingredient in the proof of Theorem 10.5.1 is the concentration of $\phi^{z,z'}$. Analogously, in the proof of Theorem 10.1.2, we will need sufficiently strong tail bounds for $\phi^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}$, which we show in Theorem 10.4.1.

10.6 An Adaptive Lower Bound for Mixedness Testing

In this section we prove our main result, Theorem 10.1.2.

First, recalling the notation from Construction 2 and Definition 10.2.10, as well as the identity (10.10), we immediately get the following from Lemma 10.2.11:

Lemma 10.6.1.

$$KL\left(p_1^{\le N} \| p_0^{\le N}\right) \le \sum_{t=1}^{N} Z_t \quad \textit{for} \quad Z_t \triangleq \mathop{\mathbb{E}}_{x_{<t}\sim p_0^{\le t-1}}\left[\frac{1}{\Delta(x_{<t})} \mathop{\mathbb{E}}_{\mathbf{U},\mathbf{U}'}\left[\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}} \cdot \phi^{\mathbf{U},\mathbf{U}'}_{x_{<t}}\right]\right] \tag{10.14}$$

Take any $t \le N$. To bound Z_t in (10.14), we first estimate the likelihood ratio Δ for an arbitrary transcript, in analogy with Lemma 10.5.3 from Section 10.5 respectively:

Lemma 10.6.2. *For any transcript $x_{<t}$, $\Delta(x_{<t}) \ge (1 - O(\varepsilon^2/d))^{t-1}$.*

Proof. Recall (10.9). By convexity of the exponential function and the fact that $1+g^{\mathbf{U}}_{x_{<i}}(x_i) > 0$ for all $\mathbf{U}, i, x_i$,

$$\Delta(x_{<t}) \ge \exp\left(\mathop{\mathbb{E}}_{U\sim\mathcal{D}}\left[\sum_{i=1}^{t-1} \ln\left(1 + g^{\mathbf{U}}_{x_{<i}}(x_i)\right)\right]\right) = \prod_{i=1}^{t-1} \exp\left(\mathop{\mathbb{E}}_{\mathbf{U}\sim\mathcal{D}}\left[\ln(1 + g^{\mathbf{U}}_{x_{<i}}(x_i))\right]\right). \tag{10.15}$$

Define the unitary block matrix $\mathbf{T} = \begin{pmatrix} \mathbf{0} & \mathrm{Id}_{d/2} \\ \mathrm{Id}_{d/2} & \mathbf{0}. \end{pmatrix}$ As $\mathcal{D}$ is invariant with respect to left-multiplication by $\mathbf{T} \in U(d)$, for all $i < t$ we have that

$$\begin{aligned}
\exp\left(\mathop{\mathbb{E}}_{\mathbf{U}\sim\mathcal{D}}\left[\ln(1 + g^{\mathbf{U}}_{x_{<i}}(x_i))\right]\right) &= \exp\left(\frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{U}\sim\mathcal{D}}\left[\ln(1 + g^{\mathbf{U}}_{x_{<i}}(x_i)) + \ln(1 + g^{\mathbf{TU}}_{x_{<i}}(x_i))\right]\right) \\
&= \exp\left(\frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{U}\sim\mathcal{D}}\left[\ln(1 + g^{\mathbf{U}}_{x_{<i}}(x_i)) + \ln(1 - g^{\mathbf{U}}_{x_{<i}}(x_i))\right]\right) \\
&= \exp\left(\frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{U}\sim\mathcal{D}}\left[\ln(1 - g^{\mathbf{U}}_{x_{<i}}(x_i)^2)\right]\right) \\
&\ge 1 + \frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{U}\sim\mathcal{D}}\left[\ln(1 - g^{\mathbf{U}}_{x_{<i}}(x_i)^2)\right] \\
&\ge 1 - \mathop{\mathbb{E}}_{\mathbf{U}\sim\mathcal{D}}\left[g^{\mathbf{U}}_{x_{<i}}(x_i)^2)\right]
\end{aligned} \tag{10.16}$$

where the second step follows from the fact that $\mathbf{T}^\dagger\mathbf{X}\mathbf{T} = -\mathbf{X}$, the fourth step follows by the elementary inequality $\exp(x) \ge 1 + x$ for all x, and the fifth inequality follows by the elementary inequality $\log(1 - x) \ge -2x$ for all $0 \le x < 1/2$.

Finally, note that for any trace-one psd matrix M, we may write $M = \sum \lambda_i v_i v_i^\dagger$, and for any unit vector $v \in \mathbb{C}^n$, $\mathbb{E}_{\mathbf{U}}[\langle vv^\dagger, \mathbf{U}^\dagger \mathbf{X} \mathbf{U}\rangle^2] = O(\varepsilon^2/d)$. So

$$\mathop{\mathbb{E}}_{U}[\langle M, \mathbf{U}^\dagger \mathbf{X} \mathbf{U}\rangle^2] = \sum_{i,j} \lambda_i \lambda_j \, \mathbb{E}\Big[\langle v_i v_i^\dagger, \mathbf{U}^\dagger \mathbf{X} \mathbf{U}\rangle \langle v_j v_j^\dagger, \mathbf{U}^\dagger \mathbf{X} \mathbf{U}\rangle\Big] \le O(\varepsilon^2/d) \cdot \left(\sum_i \lambda_i\right)^2 = O(\varepsilon^2/d),$$

where the second step follows by Cauchy-Schwarz. From this we conclude that $\mathbb{E}_{\mathbf{U}\sim\mathcal{D}}[g^{\mathbf{U}}_{x_{<i}}(x_i)^2] \le O(\varepsilon^2/d)$ for all $i, x_{<i}, x_i$, and the lemma follows by (10.15) and (10.16). □

Next, in analogy with Lemma 10.5.4, we would like to control the expectation of $(\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}})^2$. We remark that like in the proof of Lemma 10.5.4, one can obtain a naive estimate of $(1 + O(\varepsilon^2))^{t-1}$ using just Fact 10.3.3, but unlike in the proof of Theorem 10.5.1, such a bound would not suffice here. Instead, we will need the following important moment bound, whose proof we defer to Section 10.7:

Theorem 10.6.3. *For any POVM $\mathcal{M}$, let p denote the distribution over outcomes from measuring ρ_{mm} with $\mathcal{M}$, and let $\gamma > 0$ be an absolute constant. Define the random variable*

$$K^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}} \triangleq \mathop{\mathbb{E}}_{x\sim p}\left[\left(g^{\mathbf{U}}_{\mathcal{M}}(x) + g^{\mathbf{U}'}_{\mathcal{M}}(x)\right)^2\right]$$

Then for any $n = o(d^2/\varepsilon^2)$, we have that

$$\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{U}'}\left[\left(1 + \gamma \cdot K^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}\right)^n\right] \le \exp(O(\gamma n \varepsilon^2/d)) \tag{10.17}$$

We will use this and a series of invocations of Holder's to prove the following sufficiently strong generalization of Lemma 10.5.4:

Lemma 10.6.4. *Suppose $t = o(d^2/\varepsilon^2)$. Then $\mathbb{E}_{x_{<t},\mathbf{U},\mathbf{U}'}[(\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}})^2] \le \exp(O(t \cdot \varepsilon^2/d))$.*

Proof. Consider any $a, b \in \mathbf{Z}$ for which $a \ge b$ and $a \ge 2$. For any $x_{<t-1}$, let p denote the distribution over measurement outcomes when the POVM $\mathcal{M}^{x_{<t-1}}$ is applied to ρ_{mm}. We have by Part (II) of Fact 10.3.3 that

$$\mathop{\mathbb{E}}_{x\sim p}[g^{\mathbf{U}}_{x_{<t-1}}(x)^a \cdot g^{\mathbf{U}'}_{x_{<t-1}}(x_t)^b] \le \varepsilon \mathop{\mathbb{E}}_{x\sim p}[g^{\mathbf{U}}_{x_{<t-1}}(x)^2].$$

Recalling Part (I) of Fact 10.3.3, we conclude that for any $x_{<t-1}$ and constant degree $c \geq 2$,

$$\begin{aligned}
&\underset{x\sim p}{\mathbb{E}}\Big[(1+g^{\mathbf{U}}_{x_{<t-1}}(x))^c(1+g^{\mathbf{U}'}_{x_{<t-1}}(x))^c\Big] \\
&\qquad \leq 1 + O_c(\underset{x\sim p}{\mathbb{E}}[g^{\mathbf{U}}_{x_{<t-1}}(x)^2]) + O_c(\underset{x\sim p}{\mathbb{E}}[g^{\mathbf{U}'}_{x_{<t-1}}(x)^2]) + O_c(\phi^{\mathbf{U},\mathbf{U}'}_{x_{<t-1}}) \triangleq 1 + Z^{\mathbf{U},\mathbf{U}'}_{x_{<t-1}}(c). \qquad (10.18)
\end{aligned}$$

By abuse of notation, for POVM $\mathcal{M}$, define $Z^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}(c)$ in the obvious way.

For $\alpha_i \triangleq 2 \cdot \left(\frac{t-1}{t-2}\right)^i$, we have that

$$\begin{aligned}
&\underset{x_{<t},\mathbf{U},\mathbf{U}'}{\mathbb{E}}\left[\left(\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}}\right)^{\alpha_i}\right] \\
&\quad \leq \underset{x_{<t-1},\mathbf{U},\mathbf{U}'}{\mathbb{E}}\left[\left(\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t-1}}\right)^{\alpha_i} \cdot \left(1 + Z^{\mathbf{U},\mathbf{U}'}_{x_{<t-1}}(\alpha_i)\right)\right] \qquad (10.19) \\
&\quad \leq \underset{x_{<t-1},\mathbf{U},\mathbf{U}'}{\mathbb{E}}\left[\left(\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t-1}}\right)^{\alpha_i(t-1)/(t-2)}\right]^{(t-2)/(t-1)} \cdot \underset{x_{<t-1},\mathbf{U},\mathbf{U}'}{\mathbb{E}}\left[\left(1 + Z^{\mathbf{U},\mathbf{U}'}_{x_{<t-1}}(\alpha_i)\right)^{t-1}\right]^{1/(t-1)} \qquad (10.20) \\
&\quad \leq \underset{x_{<t-1},\mathbf{U},\mathbf{U}'}{\mathbb{E}}\left[\left(\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t-1}}\right)^{\alpha_{i+1}(t-1)/(t-2)}\right] \cdot \underset{x_{<t-1},\mathbf{U},\mathbf{U}'}{\mathbb{E}}\left[\left(1 + Z^{\mathbf{U},\mathbf{U}'}_{x_{<t-1}}(\alpha_i)\right)^{t-1}\right]^{1/(t-1)}.
\end{aligned}$$

where (10.19) follows by (10.18), and (10.20) follows by Holder's. Unrolling this recurrence, we conclude that

$$\begin{aligned}
\underset{x_{<t},\mathbf{U},\mathbf{U}'}{\mathbb{E}}\left[\left(\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}}\right)^2\right] &\leq \prod_{i=1}^{t-1} \underset{x_{<i},\mathbf{U},\mathbf{U}'}{\mathbb{E}}\left[\left(1 + Z^{\mathbf{U},\mathbf{U}'}_{x_{<i}}(\alpha_{t-1-i})\right)^{t-1}\right]^{1/(t-1)} \\
&\leq \prod_{i=1}^{t-1} \underset{x_{<i},\mathbf{U},\mathbf{U}'}{\mathbb{E}}\left[\left(1 + Z^{\mathbf{U},\mathbf{U}'}_{x_{<i}}(2e)\right)^{t-1}\right]^{1/(t-1)}, \qquad (10.21) \\
&\leq \sup_{\mathcal{M}} \underset{\mathbf{U},\mathbf{U}'}{\mathbb{E}}\left[\left(1 + Z^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}(2e)\right)^{t-1}\right]
\end{aligned}$$

where (10.21) follows by the fact that for $1 \leq i \leq t-1$, $\alpha_{t-1-i} \leq 2\left(1+\frac{1}{t-2}\right)^{t-2} \leq 2e$, and the supremum in the last step is over all POVMs $\mathcal{M}$. The proof is complete upon noting that $Z^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}$ is at most a constant multiple of $K^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}$ defined in (10.17) and invoking Theorem 10.6.3. □

We can now complete the proof of Theorem 10.1.2. Note that the following argument is very similar to the argument we used to complete the proof of Theorem 10.5.1.

Proof of Theorem 10.1.2. Given a transcript $x_{<t}$ and $\mathbf{U}, \mathbf{U}' \in U(d)$, let $\mathbb{1}[\mathcal{E}^{\mathbf{U},\mathbf{U}'}_{x_{<t}}(\tau)]$ denote the indicator of whether $\phi^{\mathbf{U},\mathbf{U}'}_{x_{<t}} > \tau$. We have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{U}'}\left[\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}} \cdot \phi^{\mathbf{U},\mathbf{U}'}_{x_{<t}}\right] &= \mathop{\mathbb{E}}_{\mathbf{U},\mathbf{U}'}\left[\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}} \cdot \phi^{\mathbf{U},\mathbf{U}'}_{x_{<t}} \cdot \left(\mathbb{1}[\mathcal{E}^{\mathbf{U},\mathbf{U}'}_{x_{<t}}(\tau)] + \mathbb{1}[\mathcal{E}^{\mathbf{U},\mathbf{U}'}_{x_{<t}}(\tau)^c]\right)\right] \\
&\le \varepsilon^2 \cdot \mathop{\mathbb{E}}_{\mathbf{U},\mathbf{U}'}\left[\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}} \cdot \mathbb{1}[\mathcal{E}^{\mathbf{U},\mathbf{U}'}_{x_{<t}}(\tau)]\right] + \tau \cdot \mathop{\mathbb{E}}_{\mathbf{U},\mathbf{U}'}\left[\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}} \cdot \mathbb{1}[\mathcal{E}^{\mathbf{U},\mathbf{U}'}_{x_{<t}}(\tau)^c]\right] \\
&\le \varepsilon^2 \cdot \underbrace{\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{U}'}\left[\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}} \cdot \mathbb{1}[\mathcal{E}^{\mathbf{U},\mathbf{U}'}_{x_{<t}}(\tau)]\right]}_{\textcircled{\scriptsize B}_{x_{<t}}} + \tau \cdot \underbrace{\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{U}'}\left[\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}}\right]}_{\textcircled{\scriptsize G}_{x_{<t}}},
\end{aligned}$$

where in the second step we used Part (II) of Fact 10.3.3. Note that for any transcript $x_{<t}$, $\Delta(x_{<t})^2 = \mathbb{E}_{\mathbf{U},\mathbf{U}'}[\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}}] = \textcircled{\scriptsize G}_{x_{<t}}$, so by this and the fact that the likelihood ratio between two distributions always integrates to 1,

$$\mathop{\mathbb{E}}_{x_{<t}\sim p_0^{\le t-1}}\left[\frac{1}{\Delta^{(t-1)}(x_{<t})} \cdot \textcircled{\scriptsize G}_{x_{<t}}\right] = \mathop{\mathbb{E}}_{x_{<t}\sim p_0^{\le t-1}}[\Delta^{(t-1)}(x_{<t})] = 1. \tag{10.22}$$

We conclude that

$$\begin{aligned}
Z_t &\le \varepsilon^2 \cdot \mathop{\mathbb{E}}_{x_{<t}\sim p_0^{\le t-1}}\left[\frac{1}{\Delta^{(t-1)}(x_{<t})} \cdot \textcircled{\scriptsize B}_{x_{<t}}\right] + \tau \cdot \mathop{\mathbb{E}}_{x_{<t}\sim p_0^{\le t-1}}\left[\frac{1}{\Delta^{(t-1)}(x_{<t})} \cdot \textcircled{\scriptsize G}_{x_{<t}}\right] \\
&\le \varepsilon^2 \cdot (1 + O(\varepsilon^2/d))^{t-1} \mathop{\mathbb{E}}_{x_{<t}\sim p_0^{\le t-1}}\left[\textcircled{\scriptsize B}_{x_{<t}}\right] + \tau,
\end{aligned}$$

where the second step follows by Lemma 10.6.2 and (10.22). So the challenge is to show that τ is the dominant quantity above, for appropriately chosen τ.

Pick $\tau = \varepsilon^2/d^{4/3}$. To upper bound $\mathbb{E}_{x_{<t}\sim p_0^{\le t-1}}[\textcircled{\scriptsize B}_{x_{<t}}]$, apply Cauchy-Schwarz to get

$$\begin{aligned}
\mathop{\mathbb{E}}_{x_{<t}\sim p_0^{\le t-1}}\left[\textcircled{\scriptsize B}_{x_{<t}}\right] &\le \mathop{\mathbb{E}}_{x_{<t}\sim p_0^{\le t-1},\mathbf{U},\mathbf{U}'}\left[\left(\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}}\right)^2\right]^{1/2} \cdot \mathop{\Pr}_{x_{<t}\sim p_0^{\le t-1},\mathbf{U},\mathbf{U}'}\left[\mathcal{E}^{\mathbf{U},\mathbf{U}'}_{x_{<t}}(\tau)\right]^{1/2} \\
&\le \mathop{\mathbb{E}}_{x_{<t}\sim p_0^{\le t-1},\mathbf{U},\mathbf{U}'}\left[\left(\Psi^{\mathbf{U},\mathbf{U}'}_{x_{<t}}\right)^2\right]^{1/2} \cdot \exp\left(-\Omega(d^{1/3})\right),
\end{aligned}$$

where the second step follows by Theorem 10.4.1.

This, together with Lemma 10.6.4, says that $\mathbb{E}_{x_{<t}\sim p_0^{\le t-1}}[\textcircled{\scriptsize B}_{x_{<t}}]$ is indeed negligible for

$t = o(d^{4/3}/\varepsilon^2)$. For such t, $Z_t = O(\varepsilon^2/d^{4/3})$, so by Lemma 10.6.1, $\mathrm{KL}\left(p_1^{\le N} \| p_0^{\le N}\right) = o(1)$ as desired. The desired result follows from Fact 10.2.4. □

10.7 Haar Tail Bounds

In this section we complete the proof of the two key estimates, Theorems 10.6.3 and 10.4.1, which were crucial to our proof of Theorem 10.1.2. The following concentration inequality is key to our analysis:

Theorem 10.7.1 ([MM13], Corollary 17, see also [AGZ10], Corollary 4.4.28)**.** *Equip $M \triangleq U(d)^k$ with the L_2-sum of Frobenius metrics. If $F : M \to \mathbb{R}$ is L-Lipschitz, then for any $t > 0$:*

$$\Pr_{(\mathbf{U}_1,...,\mathbf{U}_k)\in M}[|F(\mathbf{U}_1,...,\mathbf{U}_k) - \mathbb{E}[F(\mathbf{U}_1,...,\mathbf{U}_k)]| \ge t] \le e^{-dt^2/12L^2},$$

where $\mathbf{U}_1,...,\mathbf{U}_k$ are independent unitary matrices drawn from the Haar measure.

10.7.1 Proof of Theorem 10.6.3

For convenience, Theorem 10.6.3 is restated below:

Theorem 10.6.3. *For any POVM $\mathcal{M}$, let p denote the distribution over outcomes from measuring ρ_{mm} with $\mathcal{M}$, and let $\gamma > 0$ be an absolute constant. Define the random variable*

$$K_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'} \triangleq \mathop{\mathbb{E}}_{x\sim p}\left[\left(g_{\mathcal{M}}^{\mathbf{U}}(x) + g_{\mathcal{M}}^{\mathbf{U}'}(x)\right)^2\right]$$

Then for any $n = o(d^2/\varepsilon^2)$, we have that

$$\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{U}'}\left[\left(1 + \gamma \cdot K_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}\right)^n\right] \le \exp(O(\gamma n\varepsilon^2/d)) \tag{10.17}$$

To get intuition for this, consider again the special case where $\mathcal{M}$ is an orthogonal POVM given by an orthonormal basis of $\mathbb{C}^d$. Then p is uniform over $[d]$ and

$$K_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'} = \frac{\varepsilon^2}{d}\sum_{i=1}^{d}(\delta(\mathbf{U}_i) + \delta(\mathbf{U}'_i))^2 \le \frac{2\varepsilon^2}{d}\sum_{i=1}^{d}(\delta(\mathbf{U}_i)^2 + \delta(\mathbf{U}'_i)^2), \tag{10.23}$$

where $\delta(\cdot)$ is defined in (10.11). The following is a standard fact:

Fact 10.7.2. *For random unit vector $v \in \mathbb{S}^{d-1}$, $\mathbb{E}[\delta(v)^2] = \frac{1}{d+1}$.*

While this follows immediately from moments of random unit vectors, for pedagogical purposes we will give a proof using Weingarten calculus, as it will be a crucial ingredient later on. Recall that for every $q \in \mathbb{N}$, there exists a corresponding Weingarten function $\mathrm{Wg}(\cdot, d) : \mathcal{S}_q \to \mathbb{R}$ [Wei78, Col03]. In the special case of $q = 2$, the symmetric group $\mathcal{S}_q$ consists of two elements e, τ^*, namely, the identity and non-identity permutation, respectively, and we have that $\mathrm{Wg}(e, d) = \frac{1}{d^2-1}$ and $\mathrm{Wg}(\tau^*, d) = -\frac{1}{d(d^2-1)}$. We then have:

Lemma 10.7.3 (Degree-2 case of [Col03], Lemma 4.3)**.** *Let e, τ^* denote the identity and non-identity permutation of $\mathcal{S}_2$ respectively. For $d \geq 2$ and any $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{d\times d}$, we have that*[1]

$$\mathop{\mathbb{E}}_{\mathbf{U}}[Tr((\mathbf{A}\mathbf{U}^\dagger\mathbf{B}\mathbf{U})^2)] = \sum_{\sigma,\tau\in\mathcal{S}_2} \langle \mathbf{A}\rangle_\sigma \langle \mathbf{B}\rangle_\tau \,\mathrm{Wg}(\sigma\tau^{-1}, d)\ .$$

Proof of Fact 10.7.2. Let $\mathbf{\Pi} \triangleq e_1 e_1^\dagger$ and note that $\delta(v)$ is identical in distribution to the quantity $\mathrm{Tr}(\mathbf{\Pi}\mathbf{U}^\dagger\mathbf{X}'\mathbf{U})$. By Lemma 10.7.3,

$$\mathop{\mathbb{E}}_{v}[\delta(v)^2] = \mathop{\mathbb{E}}_{\mathbf{U}}[\mathrm{Tr}(\mathbf{\Pi}\mathbf{U}^\dagger\mathbf{X}'\mathbf{U})^2] = \sum_{\sigma,\tau\in\mathcal{S}_2} \langle \mathbf{\Pi}\rangle_\sigma \langle \mathbf{X}'\rangle_\tau \,\mathrm{Wg}(\sigma\tau^{-1}, d).$$

Note that $\langle \mathbf{X}'\rangle_\tau = d \cdot \mathbb{1}[\tau = \tau^*]$ and $\langle \mathbf{\Pi}\rangle_\sigma = 1$ for all $\sigma \in \mathcal{S}_2$, so

$$\mathop{\mathbb{E}}_{v}[\delta(v)^2] = d\left(\frac{1}{d^2-1} - \frac{1}{d(d^2-1)}\right) = \frac{1}{d+1}$$

as claimed. □

Furthermore, it is known that $\delta(v)^2$ concentrates around its expectation. So if the columns of $\mathbf{U}$ were actually *independent* random unit vectors, we would conclude that $K_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'} = O(\varepsilon^2/d)$ with high probability and obtain (10.17) for the special case where $\mathcal{M}$ is orthogonal.

[1] Note that this looks different from the statement in [Col03] only because they work with normalized trace $\mathrm{tr}(\cdot) \triangleq \frac{1}{d}\mathrm{Tr}(\cdot)$.

To circumvent the issue of dependence among the columns of Haar-random $\mathbf{U}$, we will invoke Theorem 10.7.1. The following is a toy version of the more general result that we show later in our proof of Theorem 10.6.3 (see Lemma 10.7.6):

Lemma 10.7.4. *For any* $t > 0$, $\Pr_{\mathbf{U}\sim\mathcal{D}}\left[\left(\sum_{i=1}^d \delta(\mathbf{U}_i)^2\right)^{1/2} \geq 1+t\right] \leq \exp\left(-\Omega(dt^2)\right)$.

Proof. By Jensen's and Fact 10.7.2,

$$\mathbb{E}\left[\left(\sum_{i=1}^d \delta(\mathbf{U}_i)^2\right)^{1/2}\right] \leq \mathbb{E}\left[\sum_{i=1}^d \delta(\mathbf{U}_i)^2\right]^{1/2} = \left(\frac{d}{d+1}\right)^{1/2} \leq 1.$$

We wish to invoke Theorem 10.7.1, so it suffices to show that $G : \mathbf{U} \mapsto (\sum_{i=1}^d \delta(\mathbf{U}_i)^2)^{1/2}$ is $O(1)$-Lipschitz. Recalling the definition of $\mathbf{X}'$ from Construction 2, note that

$$\left(\sum_{i=1}^d \delta(\mathbf{U}_i)^2\right)^{1/2} = \|\mathrm{diag}(\mathbf{U}^\dagger\mathbf{X}'\mathbf{U})\|_F.$$

Take any $\mathbf{U}, \mathbf{V} \in U(d)$ and note

$$\begin{aligned} G(\mathbf{U}) - G(\mathbf{V}) &\leq \sqrt{\sum_{i=1}^d |(\mathbf{U}^\dagger\mathbf{X}'\mathbf{U})_{ii} - (\mathbf{V}^\dagger\mathbf{X}'\mathbf{V})_{ii}|^2} \\ &\leq \|\mathbf{U}^\dagger\mathbf{X}'\mathbf{U} - \mathbf{V}^\dagger\mathbf{X}'\mathbf{V}\|_F \\ &= \|\mathbf{U}^\dagger\mathbf{X}'(\mathbf{U}-\mathbf{V}) + (\mathbf{V}-\mathbf{U})^\dagger\mathbf{X}'\mathbf{V}\|_F \\ &\leq 2\|\mathbf{X}'\|_2\|\mathbf{U}-\mathbf{V}\|_F = 2\|\mathbf{U}-\mathbf{V}\|_F, \end{aligned}$$

where the first step follows by Cauchy-Schwarz. So $G(\mathbf{U})$ is 2-Lipschitz as desired. □

Eq. (10.23), Fact 10.7.2, and Lemma 10.7.4, together with integration by parts, allow us to conclude Theorem 10.6.3 in the special case where $\mathcal{M}$ is orthogonal. Guided by the arguments above, we now proceed to our actual proof of Theorem 10.6.3.

Proof of Theorem 10.6.3. Let $\mathcal{M}$ be an arbitrary POVM. We first show a bound on $\mathbb{E}_{\mathbf{U},\mathbf{U}'}[K_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}]$, generalizing Fact 10.7.2:

Lemma 10.7.5. $\mathbb{E}_{\mathbf{U},\mathbf{U}'}[K_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}] \leq \frac{\varepsilon^2}{d+1}$.

Proof. Note that $K_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'} = 2\,\mathbb{E}_{x\sim p}\big[g_{\mathcal{M}}^{\mathbf{U}}(x)^2\big] + 2\phi_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}$. We will suppress the subscripts for the rest of this proof. Clearly we have that $\mathbb{E}_{\mathbf{U},\mathbf{U}'}[\phi^{\mathbf{U},\mathbf{U}'}] = 0$, so it remains to bound $\mathbb{E}_{x\sim p,\mathbf{U},\mathbf{U}'}\big[g^{\mathbf{U}}(x)^2\big]$. Let $\tau^* \in \mathcal{S}_2$ denote the non-identity permutation. For any fixed x, by Lemma 10.7.3,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{U}}[g^{\mathbf{U}}(x)^2] &= \sum_{\sigma,\tau\in\mathcal{S}_2} \langle \mathbf{X}\rangle_\tau \langle \widehat{M}_x\rangle_\sigma \operatorname{Wg}(\sigma\tau^{-1},d) \\
&= \langle \mathbf{X}\rangle_{\tau^*}\left(\operatorname{Tr}(\widehat{M}_x^2)\cdot \operatorname{Wg}(e,d) + \operatorname{Tr}(\widehat{M}_x)^2\cdot\operatorname{Wg}(\tau^*,d)\right) \\
&= \varepsilon^2\cdot d\cdot\left(\frac{1}{d^2-1}\operatorname{Tr}(\widehat{M}_x^2) - \frac{1}{d(d^2-1)}\right) \le \frac{\varepsilon^2}{d+1},
\end{aligned} \tag{10.24}$$

where the second step follows by the fact that $\langle \mathbf{X}\rangle_\tau = \varepsilon^2\cdot d\cdot \mathbb{1}[\tau=\tau^*]$, and the last step follows by the fact that $\operatorname{Tr}(\widehat{M}_x^2)\le 1$. As (10.24) holds for any outcome x, $\mathbb{E}_{x\sim p,\mathbf{U},\mathbf{U}'}[g^{\mathbf{U}}(x)^2] \le \frac{\varepsilon^2}{d+1}$ as desired. □

We next show the following tail bound generalizing Lemma 10.7.4:

Lemma 10.7.6. *There are absolute constants $c, c' > 0$ such that*

$$\Pr_{\mathbf{U},\mathbf{U}'\sim\mathcal{D}}\Big[K_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'} > c\varepsilon^2/d + t\Big] \le \exp(-c'td^2/\varepsilon^2))$$

for any $t > c\varepsilon^2/d$.

Proof. We wish to apply Theorem 10.7.1. We will show that $F : (\mathbf{U},\mathbf{U}') \mapsto \Big(K_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}\Big)^{1/2}$ is L-Lipschitz for $L = O(\varepsilon/\sqrt{d})$. As $\mathbb{E}[F(\mathbf{U},\mathbf{U}')] \le \mathbb{E}_{\mathbf{U},\mathbf{U}'}\Big[K_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}\Big]^{1/2} \le \frac{\varepsilon}{\sqrt{d+1}}$ by Jensen's and Lemma 10.7.5, this would imply that for any $s > 0$,

$$\Pr_{\mathbf{U}\sim\mathcal{D}}\left[F(\mathbf{U},\mathbf{U}') > \frac{\varepsilon}{\sqrt{d+1}} + s\right] \le e^{-d^2s^2/12\varepsilon^2},$$

from which the lemma would follow by taking $s = \sqrt{t}$.

To show Lipschitz-ness, note that

$$F(\mathbf{U},\mathbf{U}') = \mathop{\mathbb{E}}_{x\sim p}\left[\left(g_{\mathcal{M}}^{\mathbf{U}}(x) + g_{\mathcal{M}}^{\mathbf{U}'}(x)\right)^2\right] \le \mathop{\mathbb{E}}_{x\sim p}\left[g_{\mathcal{M}}^{\mathbf{U}}(x)^2\right]^{1/2} + \mathop{\mathbb{E}}_{x\sim p}\left[g_{\mathcal{M}}^{\mathbf{U}'}(x)^2\right]^{1/2},$$

so the proof is complete given Lemma 10.7.7 below. □

Lemma 10.7.7. *The function* $G : \mathbf{U} \mapsto \mathbb{E}_{x\sim p}\big[g^{\mathbf{U}}_{\mathcal{M}}(x)^2\big]^{1/2}$ *is* $O(\varepsilon/\sqrt{d})$*-Lipschitz.*

Proof. Take any $\mathbf{U}, \mathbf{V} \in U(d)$ and note that by triangle inequality,

$$G(U) - G(V) \le \mathop{\mathbb{E}}_{x\sim p}\left[\left(g^{\mathbf{U}}_{\mathcal{M}}(x) - g^{\mathbf{V}}_{\mathcal{M}}(x)\right)^2\right]^{1/2},$$

so it suffices to show

$$\mathop{\mathbb{E}}_{x\sim p}\left[\left(g^{\mathbf{U}}_{\mathcal{M}}(x) - g^{\mathbf{V}}_{\mathcal{M}}(x)\right)^2\right] \le O(\varepsilon^2/d) \cdot \|\mathbf{U} - \mathbf{V}\|_F^2. \tag{10.25}$$

$\mathbf{A} \triangleq \mathbf{U}^\dagger\mathbf{X}\mathbf{U} - \mathbf{V}^\dagger\mathbf{X}\mathbf{V}$ is Hermitian, so write its eigendecomposition $\mathbf{A} = \mathbf{W}^\dagger\mathbf{\Sigma}\mathbf{W}$. Then

$$g^{\mathbf{U}}_{\mathcal{M}}(x) - g^{\mathbf{V}}_{\mathcal{M}}(x) = \left\langle \widehat{M}_x, \mathbf{A}\right\rangle = \left\langle \mathbf{W}\widehat{M}_x\mathbf{W}^\dagger, \mathbf{\Sigma}\right\rangle = \left\langle \mathrm{diag}(\mathbf{W}\widehat{M}_x\mathbf{W}^\dagger), \mathbf{\Sigma}\right\rangle,$$

so we may assume without loss of generality that $\mathbf{A}$ and $\widehat{M}_x$ are diagonal, in which case by Jensen's,

$$\left(g^{\mathbf{U}}_{\mathcal{M}}(x) - g^{\mathbf{V}}_{\mathcal{M}}(x)\right)^2 = \langle \widehat{M}_x, \mathbf{A}\rangle^2 = \left(\sum_{i=1}^d (\widehat{M}_x)_{ii}\mathbf{A}_{ii}\right)^2 \le \sum_{i=1}^d (\widehat{M}_x)_{ii}\mathbf{A}_{ii}^2.$$

Recalling the definition of p and letting μ denote the measure over $\Omega(\mathcal{M})$ associated to $\mathcal{M}$ (see Definition 10.2.3), we see that the left-hand side of (10.25) becomes

$$\begin{aligned}\frac{1}{d}\int_{\Omega(\mathcal{M})} \mathrm{Tr}(M_x)\cdot\langle\widehat{M}_x, \mathbf{A}\rangle^2\,\mathrm{d}\mu &= \frac{1}{d}\int_{\Omega(\mathcal{M})}\sum_{i\in[d]}\mathrm{Tr}(M_x)\cdot(\widehat{M}_x)_{ii}\mathbf{A}_{ii}^2\,\mathrm{d}\mu = \frac{1}{d}\int_{\Omega(\mathcal{M})}\sum_{i\in[d]}(M_x)_{ii}\mathbf{A}_{ii}^2 \\ &= \frac{1}{d}\|\mathbf{A}\|_F^2 = \frac{1}{d}\|\mathbf{U}^\dagger\mathbf{X}(\mathbf{U}-\mathbf{V}) + (\mathbf{U}-\mathbf{V})^\dagger\mathbf{X}\mathbf{V}\|_F^2 \le \frac{2\varepsilon^2}{d}\|\mathbf{U}-\mathbf{V}\|_F^2,\end{aligned}$$

where the third step follows from the fact that $\int_{\Omega(\mathcal{M})} M_x\,\mathrm{d}\mu = \mathrm{Id}$, completing the proof of (10.25). □

To complete the proof of Theorem 10.6.3, we would like to apply Fact 1.3.30 to the random variable $Z \triangleq 1 + \gamma\cdot K^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}$ and the function $f(Z) \triangleq Z^n$. Note that this random variable is nonnegative and upper bounded by $1+C\cdot\gamma\cdot\varepsilon^2$ for some absolute constant $C>0$. So

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{U}'}\left[\left(1+\gamma K_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}\right)^n\right] &\le 2(1+O(\gamma\varepsilon^2/d))^n+\int_{1+c\gamma\varepsilon^2/d}^{1+C\gamma\varepsilon^2} nt^{n-1}\cdot e^{-\Omega(t\cdot d^2/\varepsilon^2)}\,\mathrm{d}t\\ &\le 2(1+O(\gamma\varepsilon^2/d))^n+\int_0^\infty nt^{n-1}\cdot e^{-\Omega(t\cdot d^2/\varepsilon^2)}\,\mathrm{d}t\\ &= 2(1+O(\gamma\varepsilon^2/d))^n+n!\cdot O(\varepsilon^2/d^2)^n\\ &\le e^{O(\gamma\varepsilon^2 n/d)},\end{aligned}$$

where in the last step we used that $n!\cdot O(\varepsilon^2/d^2)^n$ is negligible when $n=o(d^2/\varepsilon^2)$ □

10.7.2 Proof of Theorem 10.4.1

For convenience, Theorem 10.4.1 is restated below. Recall from the discussion in Section 10.3.3 that this can be thought of as the "quantum analogue" of binomial tail bounds:

Theorem 10.4.1. *Fix any POVM $\mathcal{M}$. There exists an absolute constant $c''>0$ such that for any $t>\Omega(\varepsilon^2/d^{1.99})$, we have*

$$\Pr_{\mathbf{U},\mathbf{U}'\sim\mathcal{D}}\left[\left|\phi_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}\right|>t\right]\le\exp\left(-c''\left\{\frac{d^3t^2}{\varepsilon^4}\wedge\frac{d^2t}{\varepsilon^2}\right\}\right)$$

Proof of Theorem 10.4.1. Define G as in Lemma 10.7.7. Fix any $\mathbf{U}'$ and consider the function $F_{\mathbf{U}'}:\mathbf{U}\mapsto\phi_{\mathcal{M}}^{\mathbf{U},\mathbf{U}'}$. First note that

$$\mathop{\mathbb{E}}_{\mathbf{U}}[F_{\mathbf{U}'}(\mathbf{U})]=\mathop{\mathbb{E}}_{x\sim p}[g_{\mathcal{M}}^{\mathbf{U}'}(x)\cdot\mathop{\mathbb{E}}_{\mathbf{U}}[g_{\mathcal{M}}^{\mathbf{U}}(x)]]=0$$

by Part (I) of Fact 10.3.3. Next, note that by Cauchy-Schwarz,

$$F_{\mathbf{U}'}(\mathbf{U})-F_{\mathbf{U}'}(\mathbf{V})\le\mathop{\mathbb{E}}_{x\sim p}\left[\left(g_{\mathcal{M}}^{\mathbf{U}}(x)-g_{\mathcal{M}}^{\mathbf{V}}(x)\right)^2\right]^{1/2}\cdot G(\mathbf{U}'),$$

which by (10.25) is $O(\varepsilon/\sqrt{d})\cdot G(\mathbf{U}')$-Lipschitz. So for any fixed $\mathbf{U}'$, Theorem 10.7.1 implies

$$\Pr_{\mathbf{U}}[|F_{\mathbf{U}'}(\mathbf{U})| > t] \le \exp\left(-C\cdot\frac{d^2t^2}{\varepsilon^2 G(\mathbf{U}')^2}\right) \tag{10.26}$$

for some absolute constant $C > 0$. We would like to integrate over $\mathbf{U}'$ to get a tail bound for $\phi^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}$ as a function of both $\mathbf{U}$ and $\mathbf{U}'$.

To this end, we can apply Fact 1.3.30 to the random variable $Y \triangleq G(\mathbf{U}') \in [0,\varepsilon]$. Recall from (10.24) and Jensen's that $\mathbb{E}[Y] \le \varepsilon/\sqrt{d+1}$. Furthermore, by Lemma 10.7.7 and Theorem 10.7.1, there is an absolute constant $C' > 0$ such that

$$\Pr[Y > \varepsilon/\sqrt{d+1} + s] \le \exp\left(-C'd^2s^2/\varepsilon^2\right).$$

So we can take the parameters in Fact 1.3.30 as follows: set $a \triangleq 2\varepsilon/\sqrt{d+1}$, tail bound function $\tau(x) \triangleq \exp\left(-C'\cdot\frac{d^2}{\varepsilon^2}\left(x - \frac{\varepsilon}{\sqrt{d+1}}\right)^2\right)$ for absolute constant $C' > 0$, and $f(Y) \triangleq \exp\left(-C\cdot\frac{d^2t^2}{\varepsilon^2Y^2}\right)$. By (10.26), $\Pr_{\mathbf{U},\mathbf{U}'}\left[\left|\phi^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}\right| > t\right] \le \mathbb{E}[f(Y)]$, and by Fact 1.3.30,

$$\mathbb{E}[f(Y)] \le 2\exp\left(-\Omega(d^3t^2/\varepsilon^4)\right) + \int_{2\varepsilon/\sqrt{d+1}}^{\varepsilon} \frac{2Cd^2t^2}{\varepsilon^2x^3}\exp\left(-\frac{d^2}{\varepsilon^2}\left(\frac{Ct^2}{x^2} + C'\left(x - \frac{\varepsilon}{\sqrt{d+1}}\right)^2\right)\right)\mathrm{d}x.$$

Note that by AM-GM, for $x \ge 2\varepsilon/\sqrt{d+1}$ we have that

$$\frac{Ct^2}{x^2} + C'\left(x - \frac{\varepsilon}{\sqrt{d+1}}\right)^2 \ge 2t\cdot(C\cdot C')^{1/2}\cdot\left(1 - \frac{\varepsilon/\sqrt{d+1}}{x}\right) \ge t\cdot(C\cdot C')^{1/2}.$$

We conclude that

$$\Pr_{\mathbf{U},\mathbf{U}'}\left[\left|\phi^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}\right| > t\right] \le 2\exp\left(-\Omega(d^3t^2/\varepsilon^4)\right) + \frac{2Cd^3t^2}{\varepsilon^4}\cdot\exp\left(-\Omega\left(d^2t/\varepsilon^2\right)\right).$$

In particular, for $t \ge \Omega(\varepsilon^2/d^{1.99})$, we have that

$$\Pr_{\mathbf{U},\mathbf{U}'}\left[\left|\phi^{\mathbf{U},\mathbf{U}'}_{\mathcal{M}}\right| > t\right] \ge \exp\left(-\Omega\left(\frac{d^3t^2}{\varepsilon^4} \wedge \frac{d^2t}{\varepsilon^2}\right)\right)$$

as claimed. □

10.8 Appendix: Chain Rule Proof of Theorem 10.1.3

Here we give a proof of the tight $\Omega(\sqrt{d}/\varepsilon^2)$ bound from Theorem 10.1.3 using the chain rule. The only part of the proof of the weaker Theorem 10.5.1 that we need here is Lemma 10.5.2, which we restate here for convenience.

Lemma 10.5.2.

$$KL\left(p_1^{\leq N} \| p_0^{\leq N}\right) \leq \sum_{t=1}^{N} Z_t \quad \textit{for} \quad Z_t \triangleq \mathop{\mathbb{E}}_{x_{<t} \sim U^{\otimes t-1}}\left[\frac{1}{\Delta(x_{<t})} \mathop{\mathbb{E}}_{z,z' \sim \{\pm 1\}^{d/2}}\left[\phi^{z,z'} \cdot \Psi^{z,z'}_{x_{<t}}\right]\right]. \tag{10.12}$$

Define $\varepsilon' \triangleq \log \frac{1+\varepsilon}{1-\varepsilon}$ and note that for $\varepsilon \leq 1/2$, $\varepsilon' \leq 3\varepsilon$. Given $x_{<t} \in [d]^{t-1}$, let $h(x_{<t}) \in [d]^d$ denote the vector whose j-th entry is the number of occurrences of element $j \in [d]$ in $x_{<t}$. For any $h_1, h_2 \in \mathbb{N}$, define

$$A^{h_1,h_2} \triangleq \frac{1}{2}\left((1-\varepsilon)^{h_1}(1+\varepsilon)^{h_2} + (1-\varepsilon)^{h_2}(1+\varepsilon)^{h_1}\right)$$

$$B^{h_1,h_2} \triangleq \frac{1}{2}\left((1-\varepsilon)^{h_1}(1+\varepsilon)^{h_2} - (1-\varepsilon)^{h_2}(1+\varepsilon)^{h_1}\right).$$

Fact 10.8.1. *For any $x_{<t} \in [d]^{t-1}$,*

$$\Delta(x_{<t}) = \prod_{a=1}^{d/2} A^{h(x_{<t})_{2a-1}, h(x_{<t})_{2a}}$$

$$\mathop{\mathbb{E}}_{z,z' \sim \{\pm 1\}^{d/2}}\left[\langle z, z' \rangle \cdot \Psi^{z,z'}_{x_{<t}}\right] = \sum_{a=1}^{d/2} \left(B^{h(x_{<t})_{2a-1}, h(x_{<t})_{2a}}\right)^2 \cdot \prod_{a' \neq a} \left(A^{h(x_{<t})_{2a-1}, h(x_{<t})_{2a}}\right)^2.$$

We can now complete the proof of Theorem 10.1.3.

Proof of Theorem 10.1.3. We will show that as long as $t \leq O(\sqrt{d}/\varepsilon^2)$, Z_t defined in (10.12) is no greater than $O(\varepsilon^2/\sqrt{d})$, from which the theorem follows. Fix any $t \leq N$. Let $\mathrm{Mul}_t(U)$ denote the multinomial distribution over d-tuples $\mathbf{h}$ for which $\sum_{i=1}^d h_i = t$. By Fact 10.8.1 and (10.5),

$$Z_t = \frac{2\varepsilon^2}{d} \sum_{a=1}^{d/2} \mathop{\mathbb{E}}_{\mathbf{h} \sim \mathrm{Mul}_t(U)}\left[\frac{(B^{h_{2a-1},h_{2a}})^2}{A^{h_{2a-1},h_{2a}}} \cdot \prod_{a' \neq a} A^{h_{2a'-1},h_{2a'}}\right] \triangleq \frac{2\varepsilon^2}{d} \sum_{a=1}^{d/2} C_t^{(a)}. \tag{10.27}$$

Fix any $a \in [d]$; without loss of generality suppose $a = 1$. Then

$$\begin{aligned}
C_t^{(1)} &= \frac{1}{d^t} \sum_t \binom{\mathbf{h}}{h_1 \cdots h_d} \frac{(B^{h_1,\ell-h_1})^2}{A^{h_1,\ell-h_1}} \cdot \prod_{a' \neq 1} A^{h_{2a'-1},h_{2a'}} \\
&= \frac{1}{d^t} \sum_{\ell=0}^{t} \sum_{h_1=0}^{\ell} \frac{t!}{h_1!(\ell-h_1)!(t-\ell)!} \frac{(B^{h_1,\ell-h_1})^2}{A^{h_1,\ell-h_1}} \sum_{h_3+\cdots+h_d=t-\ell} \binom{t-\ell}{h_3 \cdots h_d} \prod_{a' \neq 1} A^{h_{2a'-1},h_{2a'}} \\
&= \mathop{\mathbb{E}}_{\ell \sim \mathrm{Bin}(t,2/d)} \left[\mathop{\mathbb{E}}_{h_1 \sim \mathrm{Bin}(\ell,1/2)} \left[\frac{(B^{h_1,\ell-h_1})^2}{A^{h_1,\ell-h_1}} \right] \right]. \qquad (10.28)
\end{aligned}$$

Next, note that for any h_1, h_2,

$$\frac{(B^{h_1,h_2})^2}{A^{h_1,h_2}} = A^{h_1,h_2} - \frac{2}{\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^{h_1-h_2} + \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^{h_2-h_1}} \le A^{h_1,h_2} - \exp\left(-(h_1-h_2)^2 \varepsilon'^2/2\right).$$

Clearly $\mathbb{E}_{h_1 \sim \mathrm{Bin}(\ell,1/2)}[A^{h_1,\ell-h_1}] = 1$, so for any $0 \le \ell \le t$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{h_1 \sim \mathrm{Bin}(\ell,1/2)} \left[\frac{(B^{h_1,\ell-h_1})^2}{A^{h_1,\ell-h_1}} \right] &\le \mathop{\mathbb{E}}_{h_1 \sim \mathrm{Bin}(\ell,1/2)} \left[1 - \exp\left(-(2h_1-\ell)^2 \varepsilon'^2/2\right) \right] \\
&\le \mathop{\mathbb{E}}_{h_1 \sim \mathrm{Bin}(\ell,1/2)} \left[2(h_1 - \ell/2)^2 \varepsilon'^2 \right] = \varepsilon'^2 \cdot \ell,
\end{aligned}$$

where in the last step we used the expression for the variance of a binomial distribution. Substituting this into (10.28), we conclude that $C_t^{(a)} \le 2\varepsilon'^2 t/d \le 18\varepsilon^2 t/d$ for all $a \in [d]$, so for $t = O(\sqrt{d}/\varepsilon^2)$, (10.27) is at most $O(\varepsilon^2/\sqrt{d})$ as desired. □

Chapter 11

Instance-Optimal Quantum State Certification

11.1 Introduction

In this chapter, we turn to the more general problem of quantum state certification. As we saw in the previous chapter, even for the special case of mixedness testing, $\Omega(d^{4/3}/\varepsilon^2)$ copies are necessary if the learner can only make unentangled measurements, even if the measurements can be chosen adaptively as a function of the previous measurement outcomes. And when the measurements are additionally chosen non-adaptively, we demonstrated that $\Omega(d^{3/2}/\varepsilon^2)$ copies are necessary.

This naturally leads to another important question, namely, for which states σ can we perform state certification with lower copy complexity? As discussed in Section 1.2.4, this is inspired by a line of work in (classical) distribution testing which achieves so-called "instance optimal" bounds for identity testing. This culminated in a result due to [VV17], which states that, for any distribution p over d elements, there is a tester which requires

$$n = C_1 \cdot \max\left(\frac{1}{\varepsilon}, \frac{\left\|p^{-\max}_{-\varepsilon/16}\right\|_{2/3}}{\varepsilon^2}\right)$$

samples for some universal constant $C_1 > 0$, and distinguishes with high probability between the case where $q = p$, or when $\|q - p\|_1 \geq \varepsilon$. Here, $\| \cdot \|_{2/3}$ is the $\ell_{2/3}$-quasinorm, and $p_{-\delta}^{-\max} : [d] \to \mathbb{R}$ is the function that results after zeroing out the largest entry in the probability mass function of p, as well as zeroing out the bottom δ mass of it. To complement this bound, they also demonstrate there is a universal constant $C_2 > 0$ so that no tester can succeed with good confidence at this task, if the number of samples N satisfies

$$n \leq C_2 \cdot \max\left(\frac{1}{\varepsilon}, \frac{\left\|p_{-\varepsilon}^{-\max}\right\|_{2/3}}{\varepsilon^2}\right) .$$

Together, these two bounds give a striking and more or less tight characterization of the sample complexity landscape for this problem. Note that when p is uniform over d elements, this recovers the well-known sample complexity bound of $\Theta(\sqrt{d}/\varepsilon^2)$ for uniformity testing [Pan08].

Given this, it is natural to ask whether we can get similar σ-dependent bounds for state certification, for every σ. The bounds from the previous chapter are certainly not tight for all σ: for instance, if σ is a known pure state, it is not hard to see that $O(1/\varepsilon^2)$ copies suffice. One could hope *a priori* that a simple functional similar to the $\ell_{2/3}$ norm could also tightly characterize the instance optimal copy complexity of state certification.

11.1.1 Our Results

In this work, we present a similar instance-optimal characterization of the copy complexity of quantum state certification with non-adaptive measurements. Surprisingly, our results demonstrate that the behavior of quantum state certification is qualitatively quite different from that of classical identity testing. More formally, our main result is the following:

Theorem 11.1.1 (Informal, see Theorems 11.5.1 and 11.6.1)**.** *Given any mixed state $\sigma \in \mathbb{C}^{d\times d}$, there are mixed states $\overline{\sigma}$ and $\underline{\sigma}$ respectively given by zeroing out $\Theta(\varepsilon^2)$ and $\Theta(\varepsilon)$ total mass from σ and normalizing, such that the following holds.*

Let $\overline{\mathsf{d}}_{\mathsf{eff}}$ and $\underline{\mathsf{d}}_{\mathsf{eff}}$ be the number of nonzero entries of $\overline{\sigma}$ and $\underline{\sigma}$ respectively. The optimal copy complexity N of state certification with respect to σ to trace distance error ε using *nonadaptive, unentangled measurements satisfies*[1]

$$\widetilde{\Omega}\left(\frac{d \cdot \underline{d}_{\mathsf{eff}}^{1/2}}{\varepsilon^2} \cdot F(\underline{\sigma}, \rho_{\mathsf{mm}})\right) \le N \le \widetilde{O}\left(\frac{d \cdot \overline{d}_{\mathsf{eff}}^{1/2}}{\varepsilon^2} \cdot F(\overline{\sigma}, \rho_{\mathsf{mm}})\right).$$

Here F denotes the fidelity between two quantum states. Qualitatively, our result says that unless σ puts $1 - \mathrm{poly}(\varepsilon)$ mass on $o(d)$ dimensions, the copy complexity of state certification is equal to the worst-case copy complexity of state certification, times the fidelity between σ and the maximally mixed state. Surprisingly, unlike in the classical case, our bound demonstrates that there is no clean dimension-independent functional which controls the complexity of quantum state certification. Rather, there is some inherent "curse of dimensionality" for this problem. Also note that in the quantum case, unlike in the classical case, we do not remove the largest element from the spectrum of σ.

Example 11.1.2. *To elaborate on this qualitative difference, consider the following example. Let $\sigma \in \mathbb{C}^{(d+1)\times(d+1)}$ be the mixed state given by $\sigma = diag(1 - 1/d, 1/d^2, \ldots, 1/d^2)$. The classical analog of certifying this state is identity testing to the distribution p over $d + 1$ elements which has one element with probability $1 - 1/d$, and d elements with probability $1/d^2$.*

For the classical case, the bound from [VV17] demonstrates that the sample complexity of identity testing to p is $\Theta\left(\frac{1}{d^{4/3}\varepsilon^2}\right)$ for sufficiently small ε. In particular, in this regime the sample complexity actually decreases as we increase d. This phenomena is not too surprising—this distribution is very close to being a point distribution, and the only "interesting" part of it, namely, the tail, only has total mass $1/d$, which vanishes as we increase d.

In contrast, Theorem 11.1.1 shows that the copy complexity of the quantum version of this problem using unentangled measurements is $\widetilde{\Theta}(d^{1/2}/\varepsilon^2)$. Here, the copy complexity for this family of instances always increases as we scale d. This may be surprising, in light of the classical sample complexity. At a high level, it is because the unknown state may share the same diagonal entries with σ but may not commute with it, and therefore the "interesting" behavior need not be constrained to the subspace spanned by the small eigenvalues of σ. Concretely, the issue is that one must ensure that the off-diagonal entries of the first row

[1] Throughout, we use $\widetilde{\Omega}(\cdot)$ and $\widetilde{O}(\cdot)$ solely to suppress factors of $\log(d/\varepsilon)$.

and column of the unknown state are small. From a lower bounds perspective, this allows the lower bound instance many more degrees of freedom, which results in a much stronger resulting bound.

We can also show a lower bound against algorithms that use *adaptive*, unentangled measurements, generalizing the adaptive lower bound of the previous chapter.

Theorem 11.1.3 (Informal, see Theorem 11.7.1)**.** *In the notation of Theorem 11.1.1, the optimal copy complexity N of state certification with respect to σ to error ε using adaptive, unentangled measurements satisfies*

$$N \geq \widetilde{\Omega}\left(\frac{d \cdot d_{\mathsf{eff}}^{1/3}}{\varepsilon^2} \cdot F(\underline{\sigma}, \rho_{\mathsf{mm}})\right)$$

Since non-adaptive measurements are a subset of adaptive ones, Theorem 11.6.1 also provides a per-instance upper bound for this problem, although it does not match the lower bound we prove here. Obtaining tight bounds in this setting is an interesting open question, though we reiterate that this is not known even for mixedness testing.

11.1.2 Related Work

Apart from the works on quantum state certification and tomography discussed in Section 10.1.2, here we mention some works on instance-optimal *distribution testing*, the direct classical analog of the problem we consider in this paper. The works of [ADJ$^+$11, ADJ$^+$12] consider sample complexity bounds which improve upon the worst case sample complexity for different choices of probability distributions. The setting that we consider is most directly inspired by the work of [VV17]. Subsequent work has re-proven and/or derived new instance-optimal bounds for identity testing and other problems as well, see e.g. [DK16, BCG19, JHW18].

11.2 Overview of Techniques

As in the previous chapter, our general approach is based on showing hardness for distinguishing between a simple "null hypothesis" and a "mixture of alternatives," i.e. whether the unknown state ρ that we get copies of is equal to σ or was randomly sampled at the outset from some distribution $\mathcal{D}$ over states ε-far from σ. Recall that when $\sigma = \frac{1}{d}\,\mathrm{Id}$, a standard choice for $\mathcal{D}$ is the distribution over mixed states of the form $\frac{1}{d}\left(\mathrm{Id} + \varepsilon \cdot \mathbf{U}^\dagger \operatorname{diag}(\varepsilon, \ldots, -\varepsilon, \ldots)\mathbf{U}\right)$. Indeed, our proof builds upon the general framework introduced in the previous chapter (see Section 11.4 for a distillation of the key insights from the previous chapter) but differs in crucial ways.

To get a sense for what the right distinguishing task(s) to consider for general σ are, it is instructive to see first how to prove instance-optimal bounds for classical distribution testing.

11.2.1 Instance-Optimal Lower Bounds for Identity Testing

Here we sketch how to prove (a slightly worse version of) the lower bound of [VV17] for identity testing. Recall this is the setting where one gets access to independent samples from unknown distribution p over d elements and would like to test whether $p = q$ or $\|p - q\|_1 > \varepsilon$ for a known distribution q.

When q is the uniform distribution over d elements, a classical result of [Pan08] demonstrates that the fundamental bottleneck is distinguishing whether the samples come from p, or if the samples come from a version of q where each marginal has been perturbed by $\pm\varepsilon/d$. The main conceptual challenge to designing the lower bound for more general q is that there may be marginals of q at many different scales, and whatever lower bound instance one designs must be sensitive to these scales. One could consider the following mixture of alternatives: for a randomly sampled $\zeta \in \{\pm 1\}^d$, define q^ζ to be the distribution such that for every $i \in [d]$, q^ζ places mass $q_i + \zeta_i \cdot \varepsilon_i$ on element i, where the perturbations $\{\varepsilon_i\}$ are carefully tuned.

Pretending for now that q^ζ is a *bona fide* probability distribution, we need to upper bound the total variation distance between the distribution over N i.i.d. draws from q and the distribution over N i.i.d. draws from q^ζ where ζ was sampled uniformly at random from $\{\pm 1\}^d$ in advance. A common analytical trick for carrying out this bound— and the approach that [VV17] take— is to first Poissonize, that is, take N to be a Poisson random variable. Unfortunately, Poissonization does not seem to have any straightforward analogue in the quantum setting, where the choice of measurement can vary across copies, so we eschew this technique in favor of alternative approaches.

A somewhat simpler approach is to "bucket" the marginals, where each given bucket contains all marginals within a fixed multiplicative factor of one another. Within each bucket, one can use the simple Paninski-style lower bound on the distribution conditioned on the event that the sample lies within the bucket. Since within each bucket, the conditional distribution is close to uniform, Paninski's lower bound construction will give a hard instance for that conditional distribution. Combining these constructions across buckets after appropriately renormalizing them thus gives a natural hard instance for testing against the overall distribution q. Indeed, one can show that if one combines this with the same preprocessing steps done in the [VV17] paper (i.e. removing the largest element and the smallest ε mass), one can recover the same bound as [VV17] up to poly-logarithmic factors in $1/\varepsilon$. It is this approach that we will generalize to the quantum setting.

Ingster-Suslina Method and Moment Bounds Apart from Poissonization, another way to bound the total variation distance between the above two distributions is to pass to chi-squared divergence and invoke the Ingster-Suslina method [IS12]. At a high level, this approach amounts to bounding higher-order moments of the *relative chi-squared divergence*

$$\phi^{\zeta,\zeta'} \triangleq \mathop{\mathbb{E}}_{i}[(\Delta_\zeta(i) - 1)(\Delta_{\zeta'}(i) - 1)]$$

as a random variable in ζ, ζ'. Here, the expectation is over sample $i \in [d]$ drawn from q, and

$$\Delta_\zeta(i) = 1 + \zeta_i \cdot \varepsilon_i / q_i$$

is the *likelihood ratio* between the probability of drawing i when $p = q^\zeta$ versus the probability of drawing i when $p = q$. Concretely, if one can show that

$$\mathop{\mathbb{E}}_{\zeta,\zeta'}\left[\left(1 + \phi^{\zeta,\zeta'}\right)^t\right] = 1 + o(1) \tag{11.1}$$

for some t, this implies a sample complexity lower bound of t for testing identity to q.

To control the moments of $\phi^{\zeta,\zeta'}$, note that

$$\phi^{\zeta,\zeta'} = \sum_{i=1}^{d} \zeta_i \cdot \zeta'_i \cdot \varepsilon_i^2/q_i.$$

$\zeta_i \cdot \zeta'_i$ is uniformly random over $\{\pm 1\}$, so $\phi^{\zeta,\zeta'}$ has fluctuations of size roughly $(\sum_{i=1}^{d} \varepsilon_i^4/q_i^2)^{1/2}$, and one can show that (11.1) holds for $t = o\left((\sum_{i=1}^{d} \varepsilon_i^4/q_i^2)^{-1/2}\right)$.

At this point, it remains to optimize $\{\varepsilon_i\}$ among all perturbations for which $d_{\mathrm{TV}}(q, q^{\zeta}) \geq \varepsilon$ for all ζ. For this, one can take $\varepsilon_i \triangleq \min(q_i, \alpha q_i^{2/3})$ for α satisfying $\sum \varepsilon_i = \varepsilon$, thus recovering the instance-optimal lower bound of [VV17]. For instance, for sufficiently small ε, this lower bound specializes to $\|q\|_{2/3}/\varepsilon^2$.

11.2.2 Passing to the Quantum Setting

We now describe how to extend some of these ideas to quantum state certification.

A Generalized Quantum Paninski Instance for General σ Recall that in the case where $\sigma = \rho_{\mathsf{mm}}$, the "mixture of alternatives" we consider is given by perturbing every eigenvalue of ρ_{mm} and then randomly rotating by a Haar-unitary over $\mathbb{C}^d$.

For general σ, one cannot afford to use a global rotation. Just as we needed to be sensitive to the different scales when picking the perturbations ε_i in the classical setting, in the quantum setting we need to be sensitive to these scales not only in picking the perturbations to the eigenvalues of σ but also in picking the ensemble of rotations.

Motivated by the classical construction described above, one way to handle this is to group the eigenvalues into *buckets*, where a given bucket contains all eigenvalues within a fixed multiplicative factor of each other, and consider $\mathcal{D}$ defined as follows. $\mathcal{D}$ is a distribution over mixed states of the form $\sigma + \mathbf{U}^{\dagger}\mathbf{X}\mathbf{U}$, where now $\mathbf{U}$ is a block-diagonal unitary matrix whose blocks are Haar-random and whose block structure corresponds to the buckets, and $\mathbf{X}$ is a direct sum of multiples of Id with dimensions and magnitudes corresponding to the sizes and relative magnitudes of the buckets.

For instance, if $\sigma = \left(\frac{1}{2\sqrt{d}}\,\mathrm{Id}_{\sqrt{d}}\right) \oplus \left(\frac{1}{2(d-\sqrt{d})}\,\mathrm{Id}_{d-\sqrt{d}}\right)$, we can take $\mathbf{U}$ to be distributed as $\mathbf{U}_1 \oplus \mathbf{U}_2$, where $\mathbf{U}_1 \in U(\sqrt{d})$ and $\mathbf{U}_2 \in U(d-\sqrt{d})$ are Haar-random, and $\mathbf{X} = \left(\frac{\varepsilon_1}{2\sqrt{d}}\,\mathrm{Id}_{\sqrt{d}}\right) \oplus$

$\left(\frac{\varepsilon_2}{2(d-\sqrt{d})}\,\mathrm{Id}_{d-\sqrt{d}}\right)$ for appropriately chosen $\varepsilon_1, \varepsilon_2$ summing to 2.

Our analysis for this instance follows the Ingster-Suslina method in the nonadaptive case and the general framework of the previous chapter in the adaptive case, both of which reduce to proving that under any single-copy POVM, the relative chi-squared divergence

$$\phi^{\mathbf{U},\mathbf{V}} \triangleq \mathop{\mathbb{E}}_{z}[(\Delta_{\mathbf{U}}(z)-1)(\Delta_{\mathbf{V}}(z)-1)].$$

concentrates as a random variable in $\mathbf{U},\mathbf{V} \sim \mathcal{D}$. Analogous to the classical setup described above, here the expectation is over POVM outcomes z if one measures a single copy of the state $\rho = \sigma$, and $\Delta_{\mathbf{U}}(z)$ is the likelihood ratio between the probability of observing POVM outcome z when $\rho = \sigma + \mathbf{U}^\dagger\mathbf{X}\mathbf{U}$ versus the probability of observing the same outcome when $\rho = \sigma$ (see Section 11.4 for formal definitions).

If $\mathbf{U},\mathbf{V}$ were Haar-random over $\mathbf{U}(d)$, one could invoke standard concentration of measure for Haar-random unitary matrices [AGZ10,MM13]; this is the approach of our previous chapter, but for general σ we need to control the tails of $\phi^{\mathbf{U},\mathbf{V}}$ when $\mathbf{U},\mathbf{V}$ have the above-mentioned block structure, for which off-the-shelf tail bounds will not suffice. Instead, we argue that because we can assume without loss of generality that the optimal POVMs to use to distinguish $\rho = \sigma$ from $\rho = \sigma + \mathbf{U}^\dagger\mathbf{X}\mathbf{U}$ respect the block structure, $\phi^{\mathbf{U},\mathbf{V}}$ is a weighted sum of relative chi-squared divergences $\phi^{\mathbf{U},\mathbf{V}}_j$ for many independent sub-problems, one for each "bucket" j (see (11.8)). These are independent random variables, each parametrized by an independent Haar-random unitary matrix in a lower-dimensional space, so we can show a tail bound for $\phi^{\mathbf{U},\mathbf{V}}$ by combining the tail bounds for $\{\phi^{\mathbf{U},\mathbf{V}}_j\}$ (see Section 11.5.2).

Finally, in Section 11.5.2, we show how to tune the entries of $\mathbf{X}$ appropriately to get a copy complexity lower bound of essentially $\|\sigma\|_{2/5}/\varepsilon^2$ (see Lemma 11.5.5).

A Twist: Perturbing the Off-Diagonals The surprising thing is that this is not the best one can do! It can be shown that $\|\sigma\|_{2/5}/\varepsilon^2$ is actually dominated by the lower bound in Theorem 11.1.1 for sufficiently small ε. For instance, consider the instance in Example 11.1.2. For sufficiently small ε, the lower bound in Theorem 11.1.1 scales as $\Omega(\sqrt{d}/\varepsilon^2)$, whereas $\|\sigma\|_{2/5}/\varepsilon^2$ scales as $\Omega(1/\varepsilon^2)$. Where does the extra dimension factor come from?

This is the juncture at which instance-optimal state certification deviates significantly from its classical analogue: for sufficiently small ε, there is a distinguishing task in the former setting which is even harder than the generalized Paninski instance described above!

For simplicity, consider a mixed state σ with exactly two buckets, e.g. $\sigma = (\lambda_1 \operatorname{Id}_{d_1}) \oplus (\lambda_2 \operatorname{Id}_{d_2})$ where $d_1 \geq d_2$. In this case, one can regard the generalized Paninski instance as a family of perturbations of the two principal submatrices indexed by the coordinates $\{1, \ldots d_1\}$ in bucket 1 and the coordinates $\{d_1 + 1, \ldots, d\}$ in bucket 2 respectively. But one could also perturb σ by considering matrices of the form

$$\sigma + \begin{pmatrix} \mathbf{0}_{d_1} & (\varepsilon/2d_2) \cdot \mathbf{W} \\ (\varepsilon/2d_2) \cdot \mathbf{W}^\dagger & \mathbf{0}_{d_2} \end{pmatrix} \tag{11.2}$$

parametrized by Haar-random $\mathbf{W} \in \mathbb{C}^{d_1 \times d_2}$ consisting of orthonormal columns. One can show that as long as $\varepsilon \leq d_{j_1} \cdot \sqrt{\lambda_1 \lambda_2}$, then (11.2) is a valid density matrix (Lemma 11.5.18) and is ε-far in trace distance from σ. In this regime, we show a lower bound of $\Omega(d_1\sqrt{d_2}/\varepsilon^2)$ for distinguishing whether $\rho = \sigma$ or whether ρ is given by a matrix (11.2) where $\mathbf{W}$ is sampled Haar-randomly at the outset.

For general σ, by carefully choosing which pair of buckets to apply this construction to, we obtain the lower bound of Theorem 11.1.1 for very small ε. For larger ε we show that if the lower bound from the generalized Paninski instance were inferior to the lower bound of Theorem 11.1.1, then we would arrive at a contradiction of the assumption that ε is large (see Section 11.5.5).

Finally we remark that the ideas above can also be implemented in the setting where one can choose unentangled POVMs adaptively (Theorem 11.1.3). The reason the lower bound we obtain is not instance-optimal is the same technical reason that we were not able to obtain an optimal lower bound for mixedness testing, namely that there is some lossy balancing step (see the proof of Theorem 11.4.10 in Appendix 11.8.1).

Handling the Largest Eigenvalue We highlight one more feature of Theorems 11.1.1 and 11.1.3 which is unique to the quantum setting. In the classical setting, the instance-optimal sample complexity of testing identity to a given distribution p is essentially given by

$\frac{1}{\varepsilon} \vee \frac{\|p'\|_{2/3}}{\varepsilon^2}$, where p' is derived from p by zeroing out not just the bottom $O(\varepsilon)$ mass from p but also the *largest* entry of p. To see why the latter and the additional $\frac{1}{\varepsilon}$ term are necessary, consider a discrete distribution p which places $1-\varepsilon/100$ mass on some distinguished element of the domain, call it x^*. The $\frac{1}{\varepsilon} \vee \frac{\|p'\|_{2/3}}{\varepsilon^2}$ lower bound would yield $\Omega(1/\varepsilon)$ sample complexity, and an algorithm matching this bound would simply be to estimate the mass the unknown distribution places on x^*. The reason is that because p places total mass $\varepsilon/100$ on elements distinct from x^*, any distribution which is ε-far from p in ℓ_1-distance must place at most $1 - 101\varepsilon/100$ mass on x^*, which can be detected in $O(1/\varepsilon)$ samples.

In stark contrast, in the quantum setting if σ had an eigenvalue of $1 - \varepsilon/100$, then the copy complexity of state certification with respect to σ scales with $1/\varepsilon^2$. The reason is that there is "room in the off-diagonal entries" for a state ρ to be ε-far from σ. Indeed, we can formalize this by considering a lower bound instance similar to (11.2). In fact it is even simpler, because for mixed states whose largest eigenvalue is particularly large, it suffices to randomly perturb a single pair of off-diagonal entries! To analyze the resulting distinguishing task, we eschew the framework of the previous chapter and directly bound the likelihood ratio between observing any given sequence of POVM outcomes under the alternative hypothesis versus under the null hypothesis (see Section 11.5.4 and Lemma 11.5.24 in particular).

Upper Bound The algorithm we give for state certification is reminiscent of the instance near-optimal algorithm for identity testing from [DK16]. As in our lower bound proof, we will partition the spectrum of σ into buckets. We will also place all especially small eigenvalues of σ in a single bucket of their own.

Our starting point is a basic algorithm for state certification when the eigenvalues of σ all fall within the same bucket: this algorithm is based on measuring our copies of unknown state ρ in a Haar-random basis and running a classical identity tester [DK16]. This basic algorithm yields the upper bound in Theorem 10.1.1 from the previous chapter.

Now for a general mixed state σ, suppose there are m buckets in total. At a high level, if the state ρ to which we get copy access is ε-far in trace distance from σ, then by triangle inequality at least one of four things can happen. There may be a bucket consisting of moderately large eigenvalues for which the corresponding principal submatrix of σ is

$\Omega(\varepsilon/m^2)$-far from that of ρ, in which case we can run the abovementioned basic algorithm. Otherwise, there may be two buckets consisting of moderately large eigenvalues for which the corresponding pair of off-diagonal blocks in σ are $\Omega(\varepsilon/m^2)$-far from the corresponding submatrix in ρ.

If neither of these two cases happen, then for the single bucket consisting of all especially small eigenvalues of σ, the corresponding principal submatrix of σ is $\Omega(\varepsilon^2)$-far from that of ρ, in which case we can use a two-element POVM, one element given by the projector to that submatrix, to estimate the total mass of ρ within that submatrix. In this case, $O(1/\varepsilon^2)$ copies suffice. Finally, it could be that the none of the above three cases hold, which would imply that ρ and σ differ primarily in the off-diagonal block with rows indexed by the bucket of small eigenvalues and columns indexed by the complement. But by basic linear algebra (Fact 11.3.5), this would yield a contradiction!

11.3 Technical Preliminaries

As demonstrated in Chapter 10, our techniques generalize easily to POVMs for which $\Omega(\mathcal{M})$ is infinite, so for simplicity we will simply consider the finite case in this chapter. We retain the terminology and notation from Section 10.2.

11.3.1 Miscellaneous Technical Facts

The following elementary facts will be useful:

Fact 11.3.1. *Let S be any set of distinct positive integers. Given a collection of numbers $\{d_j\}_{j\in S}$ satisfying $\sum_j d_j 2^{-j} \le 2$, let p be the vector with d_j entries equal to 2^{-j} for every $j \in S$. Then $\max_j d_j^b 2^{-aj} \ge |S|^{-b}\|p\|_{a/b}^{-a}$ for any $a, b > 0$.*

Proof. Let j^* be the index attaining the minimum. By minimality we know $d_{j^*}2^{-aj/b} \ge \frac{1}{|S|}\sum_j d_j \cdot 2^{-aj/b}$. Raising both sides to the b-th power and taking reciprocals, we conclude that $2^{aj}/d_j^b \le |S|^b \|p\|_{a/b}^a$. □

Fact 11.3.2. *Let $c > 1$ and $p, q > 0$. Given a vector v with entries $v_1 > \cdots > v_m > 0$ for which $v_i \ge c \cdot v_{i+1}$ for every i, we have that $\|v\|_p \ge (1 - c^{-q})^{1/q} \cdot \|v\|_q$.*

Proof. We have that $\|v\|_q^q \le \sum_{i=1}^{\infty}(c \cdot v_i)^q = \frac{v_1^q}{1-c^{-q}}$, so $\|v\|_p \ge v_1 \ge \|v\|_q \cdot (1-c^{-q})^{1/q}$. □

Block Matrices and Tensors Here we record some basic facts about block matrices and tensors.

Fact 11.3.3. *Given matrices $A, B \in \mathbb{C}^{d\times d}$,*

$$A^{\otimes \ell} - B^{\otimes \ell} = \mathrm{sym}\left((A-B) \otimes \sum_{i=0}^{\ell-1} A^{\otimes i} \otimes B^{\otimes \ell-1-i}\right).$$

Fact 11.3.4 (Schur complements)**.** *For a block matrix $\rho = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\dagger & \mathbf{C}, \end{pmatrix}$ for which $\mathbf{A}$ and $\mathbf{C}$ are positive definite, ρ is positive definite if and only if Schur complement $\mathbf{C} - \mathbf{B}^\dagger \mathbf{A}^{-1}\mathbf{B}$ is positive definite.*

Fact 11.3.5. *For psd block matrix $\rho = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\dagger & \mathbf{C}, \end{pmatrix}$, where $\mathbf{A}$ and $\mathbf{C}$ are square, we have that $Tr(\mathbf{A})\ Tr(\mathbf{C}) \ge \|\mathbf{B}\|_1^2$. In particular, $\|\mathbf{B}\|_1 \le Tr(\rho)/2$.*

Proof. Without loss of generality suppose that $\mathbf{A}$ has at least as many rows/columns as $\mathbf{C}$. First note that we may assume $\mathbf{B}$ is actually square. Indeed, consider the matrix ρ' given by padding ρ with zeros,

$$\rho' = \begin{pmatrix} \mathbf{A} & \mathbf{B} & \mathbf{0} \\ \mathbf{B}^\dagger & \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}, \end{pmatrix}$$

so that $\mathbf{A}$ and $\mathbf{C}' \triangleq \begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ have the same dimensions. Clearly, $\|\begin{pmatrix} \mathbf{B} & \mathbf{0} \end{pmatrix}\|_1 = \|\mathbf{B}\|_1$, and $\|\mathbf{C}'\|_1 = \|\mathbf{C}\|_1$, so to show Fact 11.3.5 for ρ it suffices to prove it for ρ'. So henceforth, assume $\mathbf{B}$ is square.

We will further assume that $\mathbf{B}$ is diagonal. To see why this is without loss of generality, write the singular value decomposition $\mathbf{B} = \mathbf{U}^\dagger \mathbf{\Sigma} \mathbf{V}$ and note that

$$\begin{pmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{pmatrix} \rho \begin{pmatrix} \mathbf{U}^\dagger & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^\dagger \end{pmatrix} = \begin{pmatrix} \mathbf{U}^\dagger \mathbf{A} \mathbf{U} & \mathbf{\Sigma} \\ \mathbf{\Sigma} & \mathbf{V}^\dagger \mathbf{C} \mathbf{V}. \end{pmatrix}$$

If $\mathbf{B}$ is diagonal, then for every diagonal entry $\mathbf{B}_{i,i}$, we have that $\mathbf{B}_{i,i}^2 \le \mathbf{A}_{i,i}\mathbf{C}_{i,i}$, so

$$\|\mathbf{B}\|_1^2 = \left(\sum_i \mathbf{B}_{i,i}\right)^2 \le \left(\sum_i \mathbf{A}_{i,i}^{1/2}\mathbf{B}_{i,i}^{1/2}\right)^2 \le \mathrm{Tr}(\mathbf{A})\,\mathrm{Tr}(\mathbf{B}),$$

where the last step is by Cauchy-Schwarz.

The second part of the claim follows by AM-GM. □

11.3.2 Instance-Optimal Distribution Testing

We recall the precise statement of the instance-optimal lower bound from [VV17].

Theorem 11.3.6 ([VV17], Theorem 1)**.** *Given a known distribution p and samples from an unknown distribution q, any tester that can distinguish between $q = p$ and $\|p - q\|_1 \ge \varepsilon$ with probability 2/3 must draw at least $\Omega(1/\varepsilon \vee \|p_{-\varepsilon}^{-\max}\|_{2/3}/\varepsilon^2)$ samples.*

Note that this immediately implies a lower bound for state certification:

Corollary 11.3.7. *Given a known mixed state ρ and copies of an unknown mixed state σ, any tester that can distinguish between $\sigma = \rho$ and $\|\rho - \sigma\|_1 \ge \varepsilon$ with probability 2/3 using measurements on the copies of ρ must use at least $\Omega(\|\rho_{-\varepsilon}^{-\max}\|_{2/3}/\varepsilon^2)$ samples.*

We will use this corollary in our proof to handle mixed states whose eigenvalues are all pairwise separated by at least a constant factor. Intuitively, these mixed states are close to being low-rank, and one would expect that the copy complexity for testing identity to such a state is $\widetilde{\Theta}(1/\varepsilon^2)$. We show that this is indeed the case (see Lemma 11.5.12).

11.4 Generic Lower Bound Framework

In this section we abstract out some of the analysis from the previous chapter. As before, all of our lower bounds will be based on analyzing a suitable null vs. mixture distinguishing problem. Concretely, we will lower bound N for which it is possible to distinguish, using an unentangled POVM schedule $\mathcal{S}$ (recall Definition 10.2.1), between $\rho^{\otimes N}$ and $\mathbb{E}_{\mathbf{U}\sim\mathcal{D}}[\rho_{\mathbf{U}}^{\otimes N}]$ for some prior distribution $\mathcal{D}$.

As in the previous chapter, given unentangled POVM schedule $\mathcal{S}$, let $p_0^{\le N}$ (resp. $p_1^{\le N}$)

denote the distribution over transcripts given by measuring $\rho^{\otimes N}$ (resp. $\mathbb{E}_{\mathbf{U}}[\rho_{\mathbf{U}}^{\otimes N}]$ with $\mathcal{S}$.

A key component of our analysis is to bound how well a single step of $\mathcal{S}$ can distinguish between a single copy of ρ and a single copy of $\rho_{\mathbf{U}}$ for $\mathbf{U} \sim \mathcal{D}$:

Definition 11.4.1. *A* single-copy sub-problem $\mathcal{P} = (\mathcal{M}, \rho, \{\rho_{\mathbf{U}}\}_{\mathbf{U}\sim\mathcal{D}})$ *consists of the following data: a POVM $\mathcal{M}$ over $\mathbb{C}^d$, a mixed state $\rho \in \mathbb{C}^{d\times d}$, and a distribution over mixed states $\rho_{\mathbf{U}} \in \mathbb{C}^{d\times d}$ where $\mathbf{U}$ is drawn from some distribution $\mathcal{D}$.*

Definition 11.4.2. *Given a single-copy sub-problem $\mathcal{P} = (\mathcal{M}, \rho, \{\rho_{\mathbf{U}}\}_{\mathbf{U}\sim\mathcal{D}})$, let $p_0(\mathcal{M})$ denote the distribution over outcomes upon measuring ρ using $\mathcal{M} = \{M_z\}$. Given POVM outcome z, and $\mathbf{U}, \mathbf{V} \in supp(\mathcal{D})$, define the quantities*

$$g_{\mathcal{P}}^{\mathbf{U}}(z) \triangleq \frac{\langle M_z, \rho_{\mathbf{U}}\rangle}{\langle M_z, \rho\rangle} - 1 \qquad \phi_{\mathcal{P}}^{\mathbf{U},\mathbf{V}} \triangleq \mathop{\mathbb{E}}_{z\sim p_0(\mathcal{M})}\big[g_{\mathcal{P}}^{\mathbf{U}}(z) \cdot g_{\mathcal{P}}^{\mathbf{V}}(z)\big].$$

We will omit the subscript $\mathcal{P}$ when the context is clear.

11.4.1 Helpful Conditions on $g_{\mathcal{P}}^{\mathbf{U}}(z)$

For all of our lower bounds, we will design $\{\rho_{\mathbf{U}}\}$ in such a way that the following holds.

Condition 1. *For any $z \in \Omega(\mathcal{M})$, $\mathbb{E}_{\mathbf{U}}[g_{\mathcal{P}}^{\mathbf{U}}(z)] = 0$.*

One natural choice for $\mathcal{D}$ is the Haar measure over $U(d)$. In this case, many of our lower bounds are derived from showing the following two conditions hold, in addition to Condition 1.

Condition 2. $\mathbb{E}_{\mathbf{U}\sim\mathcal{D}, z\sim p_0(\mathcal{M})}[g_{\mathcal{P}}^{\mathbf{U}}(z)^2] \le \varsigma^2$.

Condition 3. $\mathbb{E}_{z\sim p_0(\mathcal{M})}[(g_{\mathcal{P}}^{\mathbf{U}}(z) - g_{\mathcal{P}}^{\mathbf{V}}(z))^2]^{1/2} \le L \cdot \|\mathbf{U} - \mathbf{V}\|_F$ *for any $\mathbf{U}, \mathbf{V} \in supp(\mathcal{D})$.*

Example 11.4.3. *In the previous chapter, we showed that if $\rho = \rho_{\mathsf{mm}}$ and $\rho_{\mathbf{U}} = \rho_{\mathsf{mm}} + \mathbf{U}^\dagger$ diag$(\frac{\varepsilon}{d}, \ldots, -\frac{\varepsilon}{d}, \ldots)\mathbf{U}$, Conditions 1, 2 and 3 hold for $\varsigma, L = O(\varepsilon/\sqrt{d})$.*

Under these conditions, we now collect some useful generic bounds. In the rest of this subsection, fix arbitrary single-copy subproblem $\mathcal{P}$ given by POVM $\mathcal{M}$; we will omit subscripts accordingly. For any $\mathbf{V} \in U(d)$, define the functions $F_{\mathbf{V}} : U(d) \to \mathbb{R}$ and $G(\mathbf{U})$

by

$$F_{\mathbf{V}}(\mathbf{U}) \triangleq \phi^{\mathbf{U},\mathbf{V}}_{\mathcal{M}} \qquad G(\mathbf{U}) \triangleq \mathop{\mathbb{E}}_{z\sim p_0(\mathcal{M})}[g^{\mathbf{U}}_{\mathcal{P}}(z)^2]^{1/2}.$$

Lemma 11.4.4. *If Conditions 1 and 3 hold, then for any $\mathbf{V} \in U(d)$, $F_{\mathbf{V}}$ is $G(\mathbf{V})\cdot L$-Lipschitz and satisfies $\mathbb{E}_{\mathbf{U}}[F_{\mathbf{V}}] = 0$.*

Proof. For any $\mathbf{U}, \mathbf{U}' \in U(d)$, we have that

$$\begin{aligned} F_{\mathbf{V}}(\mathbf{U}) - F_{\mathbf{V}}(\mathbf{U}') &= \mathop{\mathbb{E}}_{z\sim p_0(\mathcal{M})}[g^{\mathbf{V}}(z)\cdot(g^{\mathbf{U}}(z) - g^{\mathbf{U}'}(z))] \\ &\le \mathop{\mathbb{E}}_{z}[g^{\mathbf{V}}(z)^2]^{1/2}\cdot \mathop{\mathbb{E}}_{z}[(g^{\mathbf{U}}(z) - g^{\mathbf{U}'}(z))^2]^{1/2} \le G(\mathbf{V})\cdot L\cdot \|\mathbf{U}-\mathbf{U}'\|_F, \end{aligned}$$

where the first inequality is by Cauchy-Schwarz, and the second is by Condition 3.

The second part of the lemma immediately follows from Condition 1. □

Lemma 11.4.5. *If Conditions 2 and 3 hold, then for any $s > 0$,*

$$\Pr_{\mathbf{U}}[G(\mathbf{U}) > \varsigma + s] \le \exp(-\Omega(ds^2/L^2)).$$

Proof. The function G is L-Lipschitz. To see this, note that for any $\mathbf{U}, \mathbf{V} \in U(d)$,

$$G(\mathbf{U}) - G(\mathbf{V}) \le \mathop{\mathbb{E}}_{z\sim p_0(\mathcal{M})}[(g^{\mathbf{U}}_{\mathcal{P}}(z) - g^{\mathbf{V}}_{\mathcal{P}}(z))^2]^{1/2} \le L\cdot\|\mathbf{U}-\mathbf{V}\|_F,$$

where the first step is triangle inequality and the second is by Condition 3.

By Condition 2 and Jensen's, $\mathbb{E}[G(\mathbf{U})] \le \mathbb{E}[g^{\mathbf{U}}(z)^2]^{1/2} \le \varsigma$. The claim then follows by Theorem 1.3.52 and Condition 2. □

Lemma 11.4.6. *If Conditions 1, 2, and 3 hold, then for any $s > 0$,*

$$\Pr_{\mathbf{U},\mathbf{V}}\left[|\phi^{\mathbf{U},\mathbf{V}}| > s\right] \le \exp\left(-\Omega\left(\frac{ds^2}{L^2\varsigma^2} \wedge \frac{ds}{L^2}\right)\right).$$

Proof. By Lemma 11.4.4 and Theorem 1.3.52,

$$\Pr_{\mathbf{U}}[|\phi^{\mathbf{U},\mathbf{V}}| > s] \le \exp\left(-\Omega\left(\frac{ds^2}{L^2 G(\mathbf{V})^2}\right)\right). \tag{11.3}$$

We can apply Fact 1.3.30 to the random variable $Y \triangleq G(\mathbf{V})$ by taking the parameters as follows. Set $a \triangleq 2\varsigma$, $\tau(x) \triangleq \exp(-cd(x-\varsigma)^2/L^2)$, and $f(x) \triangleq \exp(-c'ds^2/L^2x^2)$ for appropriate constants $c, c' > 0$. By (11.3), $\Pr_{\mathbf{U},\mathbf{V}}[|\phi^{\mathbf{U},\mathbf{V}}| > s] \le \mathbb{E}[f(Y)]$, and by Fact 1.3.30 and Lemma 11.4.5,

$$\mathbb{E}[f(Y)] \le 2\exp\left(-\frac{c'ds^2}{L^2\varsigma^2}\right) + \int_{2\varsigma}^{\infty} \frac{2c'ds^2}{L^2x^3} \cdot \exp\left(-\frac{d}{L^2}\left(c(x-\varsigma)^2 + c's^2/x^2\right)\right) \mathrm{d}x$$

Note that for $x \ge 2\varsigma$, by AM-GM,

$$c(x-\varsigma)^2 + c's^2/x^2 \ge \Omega(s(1-\varsigma/x)) \ge \Omega(s),$$

so we can bound

$$\mathbb{E}[f(Y)] \le 2\exp\left(-\frac{c'ds^2}{L^2\varsigma^2}\right) + \Omega\left(\frac{ds^2}{L^2\varsigma^2}\right) \cdot \exp(-\Omega(ds/L^2)) \le \exp\left(-\Omega\left(\frac{ds^2}{L^2\varsigma^2} \wedge \frac{ds}{L^2}\right)\right)$$

as claimed. □

Corollary 11.4.7. *If Conditions 1, 2, and 3 hold, then for any $t \ge 1$,*

$$\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}}\left[\left(\phi^{\mathbf{U},\mathbf{V}}\right)^t\right]^{1/t} \le O\left(\varsigma L\sqrt{t/d} \vee L^2t/d\right) \le O(t \cdot L \cdot \{\varsigma \vee L\}/\sqrt{d}) \tag{11.4}$$

Proof. Define $\overline{\phi}^{\mathbf{U},\mathbf{V}} \triangleq \phi^{\mathbf{U},\mathbf{V}}$. We have

$$\begin{aligned}\mathbb{E}\left[\left(\overline{\phi}^{\mathbf{U},\mathbf{V}}\right)^t\right] &= \int_0^\infty \Pr\left[\left|\overline{\phi}^{\mathbf{U},\mathbf{V}}\right| > s^{1/t}\right] \mathrm{d}s \\ &\le \int_0^\infty \exp\left(-\Omega\left(\frac{ds^{2/t}}{L^2\varsigma^2}\right)\right) \mathrm{d}s + \int_0^\infty \exp\left(-\Omega\left(\frac{ds^{1/t}}{L^2}\right)\right) \mathrm{d}s \\ &= \Gamma(1+t/2) \cdot O(L\varsigma/\sqrt{d})^t + \Gamma(1+t) \cdot O(L^2/d)^t,\end{aligned}$$

from which the claim follows by triangle inequality. □

For our nonadaptive bounds, the weaker bound in (11.4) will suffice.

11.4.2 Nonadaptive Lower Bounds

Our nonadaptive lower bounds are based on the Ingster-Suslina method [IS12]. In the previous chapter, the main ingredients of this method are stated in the preceding notation as follows:

Lemma 11.4.8 (Restatement of Lemma 10.2.8)**.** *If the unentangled POVM schedule $\mathcal{S}$ is nonadaptive and consists of POVMs $\mathcal{M}_1, ..., \mathcal{M}_N$, then if $\mathcal{P}_t = (\mathcal{M}_t, \rho, \{\rho_{\mathbf{U}}\}_{\mathbf{U}\sim\mathcal{D}})$ denotes the t-th single-copy sub-problem for an arbitrary $\mathcal{D}$, then*

$$\chi^2\left(p_1^{\leq N} \| p_0^{\leq N}\right) \leq \max_{t\in[N]} \mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}\sim\mathcal{D}}\left[\left(1+\phi_{\mathcal{P}_t}^{\mathbf{U},\mathbf{V}}\right)^N\right] - 1 \tag{11.5}$$

Suppose Conditions 2 and 3 both hold. Then Corollary 11.4.7 immediately gives a way to upper bound the right-hand side of (11.5) and conclude a copy complexity lower bound.

Lemma 11.4.9. *Suppose Conditions 1, 2, and 3 hold for single-copy sub-problem $\mathcal{P} = (\mathcal{M}, \rho, \{\rho_{\mathbf{U}}\}_{\mathbf{U}\sim\mathcal{D}})$ for any unentangled POVM $\mathcal{M}$ and $\mathcal{D}$ the Haar measure over $U(d)$.*

Then distinguishing $\rho^{\otimes N}$ from $\mathbb{E}_{\mathbf{U}}[\rho_{\mathbf{U}}^{\otimes N}]$ with probability at least 2/3 using an unentangled, nonadaptive POVM schedule $\mathcal{S}$ requires $N = \Omega\left(\sqrt{d}L^{-1} \cdot \{\varsigma^{-1} \wedge L^{-1}\}\right)$.

Proof. For any unentangled POVM $\mathcal{M}$, we can expand

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}}\left[\left(1+\phi_{\mathcal{P}}^{\mathbf{U},\mathbf{V}}\right)^N\right] &= \sum_{2\leq t\leq N \text{ even}} \binom{N}{t} \mathbb{E}\left[\left(\phi_{\mathcal{P}}^{\mathbf{U},\mathbf{V}}\right)^t\right] \\ &\leq \sum_{2\leq t\leq N \text{ even}} \left(\frac{e\cdot N}{t}\right)^t \cdot O\left(t\cdot L\cdot\{\varsigma \vee L\}/\sqrt{d}\right)^t,\end{aligned}$$

where the first step follows by binomial theorem and the second part of Lemma 11.4.4, and the second step follows by Corollary 11.4.7. So when $N = o(\sqrt{d}L^{-1}\cdot\{\varsigma^{-1}\wedge L^{-1}\})$, by Lemma 11.4.8, $\chi^2\left(p_1^{\leq N}\|p_0^{\leq N}\right) = 1 + o(1)$. The lemma then follows from Pinsker's. □

11.4.3 Adaptive Lower Bounds

For our adaptive lower bounds, we follow the chain rule-based framework introduced in the last chapter.

Fix an unentangled, adaptive POVM schedule $\mathcal{S}$. Given a transcript of measurement outcomes $z_{<t}$ up to time t, if $\mathcal{M}^{z_{<t}}$ is the POVM used in time step t, then for convenience we will denote $g^{\mathbf{U}}_{\mathcal{P}}$, $\phi^{\mathbf{U},\mathbf{V}}_{\mathcal{P}}$, and $K^{\mathbf{U},\mathbf{V}}_{\mathcal{P}}$ by $g^{\mathbf{U}}_{z_{<t}}$, $\phi^{\mathbf{U},\mathbf{V}}_{z_{<t}}$, $K^{\mathbf{U},\mathbf{V}}_{z_{<t}}$.

Let $p_0^{\le t}$ (resp. $p_1^{\le t}$) denote the distribution over transcripts $z_{\le t}$ of outcomes up to and including time t under measuring ρ (resp. $\rho_{\mathbf{U}}$ for $\mathbf{U} \sim \mathcal{D}$) with the first t steps of $\mathcal{S}$, and define the quantities

$$\Delta(z_{\le t}) \triangleq \frac{\mathrm{d}p_1^{\le t}}{\mathrm{d}p_0^{\le t}}(z_{\le t}) \qquad \Psi^{\mathbf{U},\mathbf{V}}_{z_{<t}} \triangleq \prod_{i=1}^{t-1}(1+g^{\mathbf{U}}_{z_{<i}})(1+g^{\mathbf{V}}_{z_{<i}}).$$

The proofs for all of our adaptive lower bounds will be derived from verifying that, in addition to Conditions 1, 2, 3 hold, the following holds:

Condition 4. *For any $z \in \Omega(\mathcal{M})$, $|g^{\mathbf{U}}_{\mathcal{P}}(z)| \le 1/2$ almost surely.*

The following is implicit in the chain rule technology built in the previous chapter.

Theorem 11.4.10. *Suppose Conditions 1, 2, 3, and 4 hold for single-copy sub-problem $\mathcal{P} = (\mathcal{M}, \rho, \{\rho_{\mathbf{U}}\}_{\mathbf{U}\sim\mathcal{D}})$ for any entangled POVM $\mathcal{M}$. Then for any $\tau > 0$ and $N = o(d/L^2)$,*

$$KL\left(p_1^{\le N} \| p_0^{\le N}\right) \le N\tau + O(N) \cdot \exp\left(-\Omega\left(\left\{\frac{d\tau^2}{L^2\varsigma^2} \wedge \frac{d\tau}{L^2}\right\} - N \cdot \varsigma^2\right)\right). \tag{11.6}$$

We give a self-contained proof of this statement in Appendix 11.8.1 for the reader's convenience.

Example 11.4.11. *In the previous chapter, we showed that if $\rho = \rho_{\mathsf{mm}}$ and $\rho_{\mathbf{U}} = \rho_{\mathsf{mm}} + \mathbf{U}^\dagger\, diag(\frac{\varepsilon}{d}, \ldots, -\frac{\varepsilon}{d}, \ldots)\mathbf{U}$, then Condition 4 holds. So by taking $\tau = \varepsilon^2/d^{4/3}$, one gets that for $N = o(d^{4/3}/\varepsilon^2)$, the KL divergence in (11.6) is $o(1)$.*

11.5 Nonadaptive Lower Bound for State Certification

In this section we will show our instance-near-optimal lower bounds for state certification using nonadaptive, unentangled measurements.

Theorem 11.5.1. *There is an absolute constant $c > 0$ for which the following holds for any*

$0 < \varepsilon < c$.[2] *Let $\sigma \in \mathbb{C}^{d\times d}$ be a diagonal density matrix. There is a matrix σ^{**} given by zeroing out at most $O(\varepsilon)$ mass from σ (see Definition 11.5.2 and Fact 11.5.3 below), such that the following holds:*

Let $\widehat{\sigma}^{**} \triangleq \sigma^{**}/\, Tr(\sigma^{**})$, and let d_{eff} denote the number of nonzero entries of σ^{**}. Then any algorithm for state certification to error ε with respect to σ using nonadaptive, unentangled measurements has copy complexity at least

$$\Omega\left(d\sqrt{d_{\mathsf{eff}}}\cdot F(\widehat{\sigma}^{**},\rho_{\mathsf{mm}})/(\varepsilon^2\operatorname{polylog}(d/\varepsilon))\right).$$

In Section 11.5.1, we describe a bucketing scheme that will be essential to the core of our analysis. In Section 11.5.2 we describe and analyze the first of our two lower bound instances, a distinguishing problem based on an extension of the standard quantum Paninski construction. Specifically, in Section 11.5.2, we give a generic copy complexity lower bound for this problem, and in Section 11.5.2 we show how to tune the relevant parameters to obtain a copy complexity lower bound based on the Schatten $2/5$-quasinorm of σ. In Section 11.5.3, we describe and analyze the second of our two lower bound instances, a distinguishing problem based on perturbing the off-diagonal entries of an appropriately chosen principal submatrix of σ, obtaining for restricted choices of ε a copy complexity lower bound based on the effective dimension and Schatten $1/2$-quasinorm of σ. In Section 11.5.5, we put together the analyses of our two lower bound instances to conclude the proof of Theorem 11.5.1.

11.5.1 Bucketing and Mass Removal

We may without loss of generality assume that σ is some diagonal matrix $\operatorname{diag}(\lambda_1,\dots,\lambda_d)$.

For $j \in \mathbf{Z}_{\geq 0}$, let S_j denote the set of indices $i \in [d]$ for which $\lambda_i \in [2^{-j-1}, 2^{-j}]$; denote $|S_j|$ by d_j. Let $\mathcal{J}$ denote the set of j for which $S_j \neq \emptyset$. We will refer to $j \in \mathcal{J}$ as *buckets*. It will be convenient to refer to the index of the bucket containing a particular index $i \in [d]$ as $j(i)$. Also let S_{sing} denote the set of $i \in [d]$ belonging to a size-1 bucket S_j for some $j \in \mathcal{J}$, and let S_{many} denote the set of $i \in [d]$ which lie in a bucket S_j of size greater than 1 for some

[2] As presented, our analysis yields c within the vicinity of $1/3$, but we made no attempt to optimize for this constant.

$j \in \mathcal{J}$.

Our bounds are based on the following modification of σ obtained by zeroing out a small fraction of its entries:

Definition 11.5.2 (Removing low-probability elements- nonadaptive lower bound)**.** *Without loss of generality, suppose that $\lambda_1, \ldots, \lambda_d$ are sorted in ascending order according to $\lambda_i/d^2_{j(i)}$.*[3] *Let $d' \le d$ denote the largest index for which $\sum_{i=1}^{d'} \lambda_i \le 3\varepsilon$. Let $S_{\mathsf{tail}} \triangleq [d']$, and let S_{light} be the set of $i \in \{d'+1, \ldots, d\}$ for which $\sum_{i' \in S_{j(i)} \backslash S_{\mathsf{tail}}} \lambda_{i'} \le 2\varepsilon/\log(d/\varepsilon)$.*

Let $i_{\max}$ denote the index of the largest entry of σ. Let σ' denote the matrix given by zeroing out the largest entry of σ and the entries indexed by S_{tail}, and let σ^* denote the matrix given by zeroing out the entries indexed by $S_{\mathsf{tail}} \cup S_{\mathsf{light}}$. Finally, let σ^{**} denote the matrix given by further zeroing out from σ^* as many of the smallest entries as possible without removing more than 2ε mass.

Lastly, it will be convenient to define $\mathcal{J}'$ (resp. $\mathcal{J}^*$) to be the set of $j \in \mathcal{J}$ for which S_j has nonempty intersection with $(([d]\backslash\{i_{\max}\}) \cap S_{\mathsf{many}})\backslash S_{\mathsf{tail}}$ (resp. $[d]\backslash(S_{\mathsf{tail}} \cup S_{\mathsf{light}})$). Note that by design, $\mathcal{J}'$ and $\mathcal{J}^*$ denote the indices of the nonzero diagonal entries of σ' and σ^* respectively.

We will use the following basic consequence of bucketing:

Fact 11.5.3. *There are at most $O(\log(d/\varepsilon))$ indices $j \in \mathcal{J}$ for which S_j and S_{tail} are disjoint. As a consequence, $Tr(\sigma^{**}) \ge 1 - O(\varepsilon)$.*

Proof. For any $i_1 \notin S_{\mathsf{tail}}$ and $i_2 \in S_{\mathsf{tail}}$, we have that $p_{i_1}/d^2_{j(i_1)} \ge p_{i_2}/d^2_{j(i_2)}$, so $p_{i_1} \ge p_{i_2}/d^2$. In particular, summing over $i_2 \in S_{\mathsf{tail}}$, we conclude that $p_{i_1} \cdot |S_{\mathsf{tail}}| \ge \varepsilon/d^2$, so $p_{i_1} \ge \varepsilon/d^3$. By construction of the buckets S_j, the first part of the claim follows. For the second part, by definition we have that $\sum_{i \in [d']} \lambda_i \le O(\varepsilon)$. Furthermore, $\sum_{i \in S_{\mathsf{light}}} \lambda_i = O(\varepsilon)$ because of the first part of the claim. The second part of the claim follows by triangle inequality. □

Lastly, we will use the following shorthand: for any $j \in \mathcal{J}$ and any matrix $\mathbf{A}$, we will let $\mathbf{A}_j \in \mathbb{R}^{d \times d}$ denote the matrix which is zero outside of the principal submatrix indexed by S_j and which agrees with $\mathbf{A}$ within this submatrix.

[3] The only place where we need this particular choice of sorting is in the proof of Corollary 11.5.17 below.

11.5.2 Lower Bound Instance I: General Quantum Paninski

We will analyze the following distinguishing problem. We will pick a diagonal matrix $\boldsymbol{\mathcal{E}}$ as follows:

Definition 11.5.4 (Perturbation matrix $\boldsymbol{\mathcal{E}}$)**.** *For any $i \notin S_{\mathsf{many}}$, we will take the i-th diagonal entry of $\boldsymbol{\mathcal{E}}$ to be zero. For any bucket j of size at least 2, we will take the nonzero diagonal entries of $\boldsymbol{\mathcal{E}}_j$ to be $(\varepsilon_j, \cdots, -\varepsilon_j, \cdots)$ where there are $\lfloor d_j/2 \rfloor$ copies of ε_j and $\lfloor d_j/2 \rfloor$ copies of $-\varepsilon_j$, for ε_j to be optimized later.*

Given $\mathbf{U} \in U(d)$, define $\sigma_{\mathbf{U}} \triangleq \sigma + \mathbf{U}^\dagger \boldsymbol{\mathcal{E}} \mathbf{U}$.

Throughout this subsection, let $\mathcal{D}$ denote the distribution over block-diagonal unitary matrices $\mathbf{U}$ which are zero outside of the principal submatrices indexed by S_j for some $j \in \mathcal{J}$ with $d_j > 1$, and which within each each submatrix indexed by such an S_j is an independent Haar-random unitary if d_j is even, and otherwise is an independent Haar-random unitary in the submatrix consisting of the first $2\lfloor d_j/2 \rfloor$ rows/columns. This distinction will not be particularly important in the sequel, so the reader is encouraged to imagine that d_j is always even when $d_j > 1$.

The objective of this subsection is to show the following lower bound:

Lemma 11.5.5. *Fix $0 < \varepsilon < c$ for sufficiently small absolute constant $c > 0$. Let $\sigma \in \mathbb{C}^{d \times d}$ be a diagonal density matrix. There is a choice of $\boldsymbol{\mathcal{E}}$ in Definition 11.5.4 for which distinguishing between whether $\rho = \sigma$ or whether $\rho = \sigma + \mathbf{U}^\dagger \boldsymbol{\mathcal{E}} \mathbf{U}$ for $\mathbf{U} \sim \mathcal{D}$ using nonadaptive, unentangled measurements has copy complexity at least $\Omega(\|\sigma'\|_{2/5}/(\varepsilon^2 \log(d/\varepsilon)))$.*

By definition of $\mathcal{D}$, ρ is block-diagonal in either scenario, and the block-diagonal structure depends only on $\{S_j\}$. In particular, this implies that we can without loss of generality assume that the POVMs the tester uses respect this block structure. More precisely:

Lemma 11.5.6. *Let $\rho \in \mathbb{C}^{d \times d}$ be any density matrix which is zero outside of the principal submatrices indexed by the subsets $\{S_j\}_{j \in \mathcal{J}}$. Given an arbitrary POVM $\mathcal{M} = \{M_z\}$, there is a corresponding POVM $\mathcal{M}'$ satisfying the following. Let p, p' be the distributions over measurement outcomes from measuring ρ with $\mathcal{M}, \mathcal{M}'$ respectively. Then:*

- *For every $z \in \Omega(\mathcal{M}')$, there exists $j \in \mathcal{J}$ for which M'_z is zero outside of the principal submatrix indexed by S_j*

- *There is a function $f : \Omega(\mathcal{M}') \to \Omega(\mathcal{M})$ for which the pushforward of p' under f is p.*

Proof. For every $z \in \Omega(\mathcal{M})$ and every $j \in \mathcal{J}$, define a POVM element $M_{j,z} \triangleq \Pi_j M_z \Pi_j$, where $\Pi_j \in \mathbb{C}^{d\times d}$ is the matrix which is equal to the identity in the principal submatrix indexed by S_j and is zero elsewhere. Clearly $\{M_{j,z}\}_{j\in\mathcal{J},z\in\Omega(\mathcal{M})}$ is still a POVM because $\sum \Pi_j = \mathrm{Id}$; let $\mathcal{M}'$ be this POVM. Let f be given by $f((j,z)) = z$. The pushforward of p' under f places mass

$$\sum_{j\in\mathcal{J}} \langle \rho, \Pi_j M_z \Pi_j \rangle = \left\langle \sum_{j\in\mathcal{J}} \Pi_j \rho \Pi_j, M_z \right\rangle = \langle \rho, M_z \rangle$$

on $z \in \Omega(\mathcal{M})$ as claimed, where the penultimate step follows by the assumption that ρ is zero outside of the principal submatrices indexed by the subsets $\{S_j\}$. □

By Lemma 11.5.6, we will henceforth only work with POVMs like $\mathcal{M}'$. If $\mathcal{M}^t$ is the t-th POVM used by the tester, we may assume without loss of generality that its outcomes $\Omega(\mathcal{M}^t)$ consist of pairs (j,z), where the POVM element corresponding to such a pair has nonzero entries in the principal submatrix indexed by S_j. Henceforth, fix an arbitrary such POVM $\mathcal{M}$ (we will drop subscripts accordingly) and denote its elements by $\{M_{j,z}\}$ for $j \in \mathcal{J}$. We will denote by Ω_j the set of z for which there is an element $M_{j,z}$.

Let p denote the distribution over $\mathcal{J}$ induced by measuring σ with $\mathcal{M}$ and recording which bucket the outcome belongs to. Concretely, p places mass $p_j \triangleq \sum_{z\in\Omega_j} \langle M_{j,z}, \sigma_j \rangle = \mathrm{Tr}(\sigma_j)$ on bucket $j \in \mathcal{J}$. Similarly, define q^j to be the distribution over Ω_j conditioned on the outcome falling in bucket j, that is, q^j places mass $q^j_z \triangleq \frac{1}{p_j}\langle M_{j,z}, \sigma_j \rangle$ on $z \in \Omega_j$.

For every $j \in \mathcal{J}$, let $\mathcal{P}_j$ denote the d_j-dimensional sub-problem given by restricting to the coordinates indexed by S_j and using the POVM $\mathcal{M}_j \triangleq \{(M_{j,z})_j\}_{z\in\Omega_j}$. Formally, $\mathcal{P}_j$ is specified by the data $(\mathcal{M}_j, \widehat{\sigma}_j, \{(\widehat{\sigma}_{\mathbf{U}})_j\}_{\mathbf{U}\sim\mathcal{D}_j})$, where $\mathcal{D}_j$ is the Haar measure over $U(d_j)$ if d_j is even and is otherwise the distribution over $d_j \times d_j$ matrices which are Haar-random unitary in the first $2\lfloor d_j/2 \rfloor$ rows/columns and zero elsewhere. Note that the density matrix $(\widehat{\sigma}_{\mathbf{U}})_j$ can be written as $\widehat{\sigma}_j + \mathbf{U}^\dagger \boldsymbol{\mathcal{E}}'_j \mathbf{U}$ for $\boldsymbol{\mathcal{E}}'_j \triangleq \boldsymbol{\mathcal{E}}_j / p_j$.

For any $j \in \mathcal{J}$, $z \in \Omega_j$, it will be convenient to define $\widetilde{M}_{j,z} \triangleq \frac{1}{\langle M_{j,z}, \sigma_j \rangle} M_{j,z}$. We can write

$$g^{\mathbf{U}_j}_{\mathcal{P}_j}(z) = \frac{\langle M_{j,z}, \mathbf{U}_j^\dagger \boldsymbol{\mathcal{E}}'_j \mathbf{U}_j \rangle}{\langle M_{j,z}, \widehat{\sigma}_j \rangle} = \frac{\langle M_{j,z}, \mathbf{U}_j^\dagger \boldsymbol{\mathcal{E}}_j \mathbf{U}_j \rangle}{\langle M_{j,z}, \sigma_j \rangle} = \langle \widetilde{M}_{j,z}, \mathbf{U}_j^\dagger \boldsymbol{\mathcal{E}}_j \mathbf{U}_j \rangle. \tag{11.7}$$

Because $M_{j,z}$ is zero outside of the principal submatrix indexed by S_j, we thus have

$$g^{\mathbf{U}}(z) = \frac{\langle M_{j,z}, \mathbf{U}^\dagger \boldsymbol{\mathcal{E}} \mathbf{U} \rangle}{\langle M_{j,z}, \sigma \rangle} = \frac{\langle M_{j,z}, \mathbf{U}_j^\dagger \boldsymbol{\mathcal{E}}_j \mathbf{U}_j \rangle}{\langle M_{j,z}, \sigma_j \rangle} = g^{\mathbf{U}_j}_{\mathcal{P}_j}(z)$$

and

$$\phi^{\mathbf{U},\mathbf{V}} = \mathop{\mathbb{E}}_{j,z}\left[\frac{\langle M_{j,z}, \mathbf{U}_j^\dagger \boldsymbol{\mathcal{E}}_j \mathbf{U}_j \rangle \langle M_{j,z}, \mathbf{V}_j^\dagger \boldsymbol{\mathcal{E}}_j \mathbf{V}_j \rangle}{\langle M_{j,z}, \sigma_j \rangle^2}\right] = \sum_{j \in \mathcal{J}} p_j \cdot \phi^{\mathbf{U}_j,\mathbf{V}_j}_{\mathcal{P}_j}. \tag{11.8}$$

We now give a generic lower bound for the distinguishing problem in Lemma 11.5.5 that depends on the entries of $\boldsymbol{\mathcal{E}}$. After that, we show how to tune the entries of $\boldsymbol{\mathcal{E}}$ to complete the proof of Lemma 11.5.5.

Bound Under General Perturbations

Our goal is first to show the following generic bound:

Lemma 11.5.7. *Distinguishing $\sigma^{\otimes N}$ from $\mathbb{E}_{\mathbf{U}}[\sigma_{\mathbf{U}}^{\otimes N}]$ with probability at least 2/3 using an unentangled, adaptive POVM schedule $\mathcal{S}$ requires*

$$N = \Omega\left(\left(\sum_{j \in \mathcal{J}} \frac{2^{2j} \varepsilon_j^4}{d_j}\right)^{-1/2}\right) \tag{11.9}$$

By Lemma 11.4.8, it suffices to show that for any POVM $\mathcal{M}$, $\mathbb{E}_{\mathbf{U},\mathbf{V}}\left[\left(1 + \phi^{\mathbf{U},\mathbf{V}}_{\mathcal{M}}\right)^N\right] = 1 + o(1)$ for N smaller than the claimed bound. To do this, we will bound the moments of each $\phi^{\mathbf{U},\mathbf{V}}_{\mathcal{P}_j}$ individually.

As the relevant matrices $(M_{j,z})_j$ are zero outside of the principal submatrix indexed by S_j, we will abuse notation and refer to them as $M_{j,z}$ in the sequel whenever the context is clear. Likewise, we will refer to $\mathbf{U}_j \sim \mathcal{D}_j$ as $\mathbf{U}$.

Lemma 11.5.8. *For any $z \in \Omega_j$, $\mathbb{E}_{\mathbf{U}_j}[g^{\mathbf{U}_j}_{\mathcal{P}_j}(z)] = 0$, so Condition 1 holds.*

Proof. By Fact 1.3.48, $\mathbb{E}_{\mathbf{U}}[g^{\mathbf{U}}_{\mathcal{P}_j}(z)] = \mathrm{Tr}(\widetilde{M}_{j,z}) \cdot \mathrm{Tr}(\boldsymbol{\mathcal{E}}_j) = 0$. □

We will now show that Conditions 2 and 3 hold for appropriate choices of ς and L. It will be convenient to define $\widetilde{M}_{j,z} \triangleq \frac{1}{\langle M_{j,z}, \sigma_j\rangle} M_{j,z}$

Lemma 11.5.9. *$\mathbb{E}_{\mathbf{U}}[g^{\mathbf{U}}_{\mathcal{P}_j}(z)^2]^{1/2} \le O(2^j \varepsilon_j/\sqrt{d_j})$ for any $z \in \Omega_j$.*

Proof. Let $\tau^* \in \mathcal{S}_2$ denote transposition. For any $z \in \Omega_j$, by (11.7) and Lemma 1.3.49,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{U}}[g^{\mathbf{U}}_{\mathcal{P}_j}(z)^2] &= \mathbb{E}\left[\left\langle \widetilde{M}_{j,z}, \mathbf{U}^\dagger \boldsymbol{\mathcal{E}}_j \mathbf{U}\right\rangle^2\right] \\
&= \sum_{\pi,\tau \in \mathcal{S}_2} \langle \boldsymbol{\mathcal{E}}_j\rangle_\tau \langle \widetilde{M}_{j,z}\rangle_\pi \, \mathrm{Wg}(\pi\tau^{-1}, d_j) \\
&= \langle \boldsymbol{\mathcal{E}}_j\rangle_{\tau^*} \left(\mathrm{Tr}(\widetilde{M}^2_{j,z}) \cdot \mathrm{Wg}(e, d_j) + \mathrm{Tr}(\widetilde{M}_{j,z})^2 \cdot \mathrm{Wg}(\tau^*, d_j)\right) \\
&\le d_j \cdot \varepsilon_j^2 \cdot \frac{\mathrm{Tr}(M_{j,z})^2}{\langle M_{j,z}, \sigma_j\rangle^2} \left(\frac{1}{d_j^2 - 1}\mathrm{Tr}(\widehat{M}^2_{j,z}) - \frac{1}{d_j(d_j^2-1)} \cdot \mathrm{Tr}(\widehat{M}_{j,z})^2\right) \\
&\le \frac{\varepsilon_j^2}{d_j + 1} \cdot \frac{\mathrm{Tr}(M_{j,z})^2}{\langle M_{j,z}, \sigma_j\rangle^2} \le 2 \cdot 2^{2j} \varepsilon_j^2/d_j,
\end{aligned}$$

where in the last step we used the fact that $\mathrm{Tr}(\widehat{M}^2) \le 1$ for any matrix $\widehat{M}$ of trace 1. □

Lemma 11.5.10. *$\mathbb{E}_{z\sim q^j}[(g^{\mathbf{U}}_{\mathcal{P}_j}(z) - g^{\mathbf{V}}_{\mathcal{P}_j}(z))^2]^{1/2} \le O((2^j/p_j)^{1/2}\varepsilon_j) \cdot \|\mathbf{U} - \mathbf{V}\|_F$ for any $\mathbf{U}, \mathbf{V} \in U(d)$.*

Proof. The matrix $\mathbf{A} \triangleq \mathbf{U}^\dagger \boldsymbol{\mathcal{E}}_j \mathbf{U} - \mathbf{U}'^\dagger \boldsymbol{\mathcal{E}}_j \mathbf{U}'$ is Hermitian, so write its eigendecomposition $\mathbf{A} = \mathbf{W}^\dagger \boldsymbol{\Sigma} \mathbf{W}$. Define $M'_{j,z} \triangleq \mathbf{W} M_{j,z} \mathbf{W}^\dagger$ so that $\sum_{z\in\Omega_j} M'_{j,z} = \mathrm{Id}_{d_j}$ and

$$\begin{aligned}
\mathop{\mathbb{E}}_{z\sim q^j}[(g^{\mathbf{U}}_{\mathcal{P}_j}(z) - g^{\mathbf{V}}_{\mathcal{P}_j}(z))^2] &= \mathop{\mathbb{E}}_{z\sim q_j}\left[\left(\frac{1}{\langle M_{j,z}, \sigma_j\rangle} \sum_{i=1}^{d_j} (M'_{j,z})_{ii} \boldsymbol{\Sigma}_{ii}\right)^2\right] \\
&\le \mathop{\mathbb{E}}_{z\sim q_j}\left[\left(\frac{1}{\langle M_{j,z}, \sigma_j\rangle} \sum_{i=1}^{d_j} (M'_{j,z})_{ii} \Sigma^2_{ii}\right)\left(\frac{1}{\langle M_{j,z}, \sigma_j\rangle} \sum_{i=1}^{d_j} (M'_{j,z})_{ii}\right)\right] \\
&\le \frac{1}{p_j} \sum_{z\in\Omega_j} \frac{\mathrm{Tr}(M_{j,z})}{\langle M_{j,z}, \sigma_j\rangle} \cdot \sum_{i=1}^{d_j} (M'_{j,z})_{ii} \boldsymbol{\Sigma}^2_{ii} \\
&\le \frac{1}{p_j} 2^{j+1} \cdot \sum_{i=1}^{d_j} \boldsymbol{\Sigma}^2_{ii} \sum_{z\in\Omega_j} (M'_{j,z})_{ii} = \frac{1}{p_j} 2^{j+1} \|\boldsymbol{\Sigma}\|_F^2
\end{aligned}$$

where in the second step we used Cauchy-Schwarz, in the third step we used that $\mathrm{Tr}(M'_{j,z}) = \mathrm{Tr}(M_{j,z})$, in the fourth step we used the fact that the entries of diagonal matrix σ_j are lower bounded by 2^{-j-1}, and in the fifth step we used that $\sum_z \Omega'_{j,z} = \mathrm{Id}_{d_j}$. To upper bound $\|\mathbf{\Sigma}\|_F$, note

$$\|\mathbf{\Sigma}\|_F = \|\mathbf{U}^\dagger \boldsymbol{\mathcal{E}}_j \mathbf{U} - \mathbf{U}'^\dagger \boldsymbol{\mathcal{E}}_j \mathbf{U}'\|_F = \|\mathbf{U}^\dagger \boldsymbol{\mathcal{E}}_j (\mathbf{U} - \mathbf{U}') + (\mathbf{U}' - \mathbf{U})^\dagger \boldsymbol{\mathcal{E}}_j \mathbf{U}'\|_F \le \varepsilon_j \|\mathbf{U} - \mathbf{U}'\|_F,$$

from which we conclude that $\mathbb{E}_{z\sim q^j}[(g^{\mathbf{U}}_{\mathcal{P}_j}(z) - g^{\mathbf{V}}_{\mathcal{P}_j}(z))^2]^{1/2} \le (2^{j+1}/p_j)^{1/2}\varepsilon_j \|\mathbf{U} - \mathbf{U}'\|_F$. □

By applying Corollary 11.4.7, we get the following bound:

Lemma 11.5.11. *For any odd t, $\mathbb{E}_{\mathbf{U},\mathbf{V}\sim\mathcal{D}_j}\left[\left(\phi^{\mathbf{U},\mathbf{V}}_{\mathcal{P}_j}\right)^t\right] = 0$, and for any even t,*

$$\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}\sim\mathcal{D}_j}\left[\left(\phi^{\mathbf{U},\mathbf{V}}_{\mathcal{P}_j}\right)^t\right]^{1/t} \le O\left(2^{2j}\varepsilon_j^2/d_j \cdot \left\{\sqrt{t/d_j} \vee t/d_j\right\}\right) \le O\left(t \cdot 2^{2j} \cdot \varepsilon_j^2/d_j^{3/2}\right).$$

Proof. By Lemma 11.5.8 and the definition of $\phi^{\mathbf{U},\mathbf{V}}_{\mathcal{P}_j}$, $\mathbb{E}[\phi^{\mathbf{U},\mathbf{V}}_{\mathcal{P}_j}] = 0$. By Lemmas 11.5.9 and 11.5.10, we can take $\varsigma = O(2^j\varepsilon_j/\sqrt{d_j})$ and $L = O((2^j/p_j)^{1/2}\varepsilon_j)$ when invoking Corollary 11.4.7. Note that $p_j \ge d_j 2^{-j-1}$, so $L \le O(\varsigma)$. The claim follows. □

Lemma 11.5.11, Lemma 1.3.25, and (11.8) immediately imply Lemma 11.5.7.

Proof of Lemma 11.5.7. From Lemma 1.3.25, Lemma 11.5.11, and (11.8), we have that

$$\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}\sim\mathcal{D}}\left[\left(\phi^{\mathbf{U},\mathbf{V}}\right)^t\right]^{1/t} \le t\left(\sum_{j\in\mathcal{J}} p_j^2 \cdot O\left(\frac{2^{4j}\varepsilon_j^4}{d_j^3}\right)\right)^{1/2} \le t\left(\sum_{j\in\mathcal{J}} O\left(\frac{2^{2j}\varepsilon_j^4}{d_j}\right)\right)^{1/2}$$

where in the second step we used that $p_j \le d_j 2^{-j}$. We can thus expand

$$\mathbb{E}\left[\left(1+\phi^{\mathbf{U},\mathbf{V}}\right)^N\right] = \sum_{2\le t\le N \text{ even}} \binom{N}{t} \mathbb{E}[(\phi^{\mathbf{U},\mathbf{V}})^t] \le \left(\frac{e\cdot N}{t}\right)^t \cdot O\left(t^2 \sum_{j\in\mathcal{J}} \frac{2^{2j}\varepsilon_j^4}{d_j}\right)^{t/2},$$

from which the claim follows by Lemma 11.4.8. □

Tuning the Perturbations

Before we explain how to tune $\boldsymbol{\mathcal{E}}_j$, we address a minor corner case. Recall from Definition 11.5.4 that $\boldsymbol{\mathcal{E}}_j$ is zero for buckets j for which $|S_j| = 1$. In the extreme case where all buckets after removal of S_{tail} are of this type, then $\boldsymbol{\mathcal{E}} = 0$ and the problem of distinguishing between σ and $\sigma + \mathbf{U}^\dagger \boldsymbol{\mathcal{E}} \mathbf{U}$ would be vacuous. Fortunately, we can show that if the Schatten 2/5-quasinorm of σ' is dominated by such buckets, then the resulting state certification problem requires many copies because of existing *classical* lower bounds.

Lemma 11.5.12. *If $\sum_{i\in S_{\mathsf{sing}}\backslash S_{\mathsf{tail}}} \lambda_i^{2/5} \geq \frac{1}{2}\|\sigma'\|_{2/5}^{2/5}$, then state certification with respect to σ using nonadaptive, unentangled measurements has copy complexity at least $\Omega(\|\sigma'\|_{2/5}/\varepsilon^2)$.*

Proof. Intuitively in this case, the spectrum of σ is dominated by eigenvalues in geometric progression, and in fact the instance-optimal lower bound for *classical* identity testing [VV17] already implies a good enough copy complexity lower bound (even against entangled measurements).

Formally, Corollary 11.3.7 implies a copy complexity lower bound of $\Omega(1/\varepsilon \vee \|\sigma^{-\max}_{-\varepsilon}\|_{2/3}/\varepsilon^2)$. We would like to relate this to

$$\left(\sum_{i\in S_{\mathsf{sing}}\backslash S_{\mathsf{tail}}} \lambda_i^{2/3}\right)^{3/2} \geq (1-2^{-2/5})^{5/2} \cdot \left(\sum_{i\in S_{\mathsf{sing}}\backslash S_{\mathsf{tail}}} \lambda_i^{2/5}\right)^{5/2} \geq \Omega(\|\sigma'\|_{2/5}), \tag{11.10}$$

where the first step follows by Fact 11.3.2, and the last step follows by the hypothesis of the lemma.

Suppose that there is some i for which $d_{j(i)} = 1$ and i is not among the indices removed in the definition of $\sigma^{-\max}_{-\varepsilon}$. Then we can lower bound $\|\sigma^{-\max}_{-\varepsilon}\|_{2/3}$ by λ_i, which is at least $(1-2^{-2/3})^{3/2} = \Omega(1)$ times the left-hand side of (11.10).

On the other hand, suppose that all i for which $d_{j(i)} = 1$ are removed in the definition of $\sigma^{-\max}_{-\varepsilon}$. As long as $\sigma^{-\max}_{-\varepsilon}$ has some nonzero entry, call it λ_{i^*}, then $\lambda_{i^*} \geq \max_{i\in S_{\mathsf{sing}}\backslash S_{\mathsf{tail}}} \lambda_i$, so we can similarly guarantee that $\|\sigma^{-\max}_{-\varepsilon}\|_{2/3} \geq \lambda_{i^*}$ is at least $(1-2^{-2/3})^{3/2} = \Omega(1)$ times the left-hand side of (11.10). Otherwise, we note that σ' is zero as well, in which case we are also done. □

It remains to consider the primary case where the hypothesis of Lemma 11.5.12 does not hold, and this is where we will use Lemma 11.5.7. The following together with Lemma 11.5.12 will complete the proof of Lemma 11.5.5:

Lemma 11.5.13. *If $\sum_{i\in S_{\mathsf{sing}}\backslash S_{\mathsf{tail}}} \lambda_i^{2/5} < \frac{1}{2}\|\sigma'\|_{2/5}^{2/5}$, then state certification with respect to σ using nonadaptive, unentangled measurements has copy complexity at least $\Omega(\|\sigma'\|_{2/5}/(\varepsilon^2\log(d/\varepsilon)))$.*

The proof of Lemma 11.5.13 requires some setup. First, obviously the hypothesis of the lemma can equivalently be stated as

$$\sum_{i\in S_{\mathsf{many}}\backslash S_{\mathsf{tail}}} \lambda_i^{2/5} > \frac{1}{2}\|\sigma'\|_{2/5}^{2/5}. \tag{11.11}$$

Definition 11.5.14 (Choice of ε_j)**.** *For every $i \in S_{\mathsf{many}}$, for $j \in \mathcal{J}$ the index of the bucket containing i, define $\varepsilon_j \triangleq 2^{-j-1} \wedge \zeta 2^{-2/3(j+1)} d_j^{2/3}$ for normalizing quantity ζ satisfying*

$$\sum_{j\in\mathcal{J}:d_j>1} 2\lfloor d_j/2\rfloor \cdot \left\{2^{-j-1} \wedge \zeta 2^{-2/3(j+1)} d_j^{2/3}\right\} = \varepsilon. \tag{11.12}$$

Note that by ensuring that $\varepsilon_j \le 2^{-j-1}$, we ensure that $\sigma + \mathbf{U}^\dagger \boldsymbol{\mathcal{E}} \mathbf{U}$ has nonnegative spectrum, while (11.12) ζ ensures that for any $\mathbf{U}$ in the support of $\mathcal{D}$, $\|\boldsymbol{\mathcal{E}}\|_1 = \varepsilon$.

The rest of the proof is devoted to showing that for this choice of $\{\varepsilon_j\}$, the lower bound in (11.9) is at least the one in Lemma 11.5.13. The main step is to upper bound the normalizing quantity ζ.

Lemma 11.5.15. *For ζ defined in Definition 11.5.14,*

$$\zeta \le O(\varepsilon) \cdot \left(\sum_{j\in\mathcal{J}',i\in S_j} \lambda_i^{2/3} d_j^{5/3}\right)^{-1}. \tag{11.13}$$

We will need the following elementary fact.

Fact 11.5.16. *Let $u_1 < \cdots < u_m$ and $v_1 \le \cdots \le v_n$ be numbers for which $u_{i+1} \ge 2u_i$ for all i. Let $d_1, \ldots, d_n > 1$ be arbitrary integers. Let $w_1 \le \cdots \le w_{m+n}$ be these numbers in sorted order. For $i \in [m+n]$, define d_i^* to be 1 if w_i corresponds to some u_j, and d_j if w_i corresponds to some v_j.*

Let s be the largest index for which $\sum_{i=1}^{s} w_i d_i^* \leq 3\varepsilon$. Let a, b be the largest indices for which u_a, v_b are present among $w_1, \dots, w_s$ (if none exists, take it to be 0). Then either $b = n$ or $\sum_{i=1}^{b+1} v_i d_i > \varepsilon$.

This allows us to deduce the following bound for buckets not removed in Definition 11.5.2.

Corollary 11.5.17. *Under the hypothesis of Lemma 11.5.13, $S_{\mathsf{many}} \backslash S_{\mathsf{tail}}$ is nonempty, and there exists an absolute constant $c > 0$ such that for any $i \in S_{\mathsf{many}} \backslash S_{\mathsf{tail}}$ in some bucket j, $\zeta \cdot 2^{-2/3(j+1)} d_j^{2/3} \leq c \cdot 2^{-j-1}$.*

Proof. The first part immediately follows from (11.11). For the second part, take some constant c to be optimized later and suppose to the contrary that for some $i^* \in S_{\mathsf{many}} \backslash S_{\mathsf{tail}}$, lying in some bucket j^*, we have that $c \cdot 2^{-j^*-1} < \zeta \cdot 2^{-2/3(j^*+1)} d_j^{2/3}$, or equivalently $2^{-j^*-1}/d_j^2 < \zeta^3/c^3$. Because in the definition of S_{tail}, we sorted by $\lambda_i / d_{j(i)^2}$, for any $i \in S_{\mathsf{tail}}$, and because $\lambda_i \in [2^{-j(i)-1}, 2^{-j(i)}]$, we also have that $2^{-j(i)-1}/d_{j(i)}^2 < \zeta^3/c^3$, or equivalently, $c \cdot 2^{-j(i)-1} < \zeta \cdot 2^{-2/3(j+1)} d_{j(i)}^{2/3}$.

So the sum on the left-hand side of (11.12) is at least

$$\sum_{j \in \mathcal{J} : j \geq j^*, d_j > 1} 2\lfloor d_j/2 \rfloor \cdot (c \cdot 2^{-j-1}) \geq \sum_{j \in \mathcal{J} : j \geq j^*, d_j > 1} (2d_j/3) \cdot (c \cdot 2^{-j-1}) \geq \sum_{i \in S_{\mathsf{many}}, i \leq i^*} \lambda_i > \varepsilon,$$

where in the first step we used that for $d_j > 1$, $2\lfloor d_j/2 \rfloor \geq 2d_j/3$, in the second step we took $c = 3$ and used that $\lambda_i \leq 2^{-j}$ for $i \in S_j$, and in the third step we used Fact 11.5.16 applied to the numbers $\{u_i\} \triangleq \{\lambda_i\}_{i \in S_{\mathsf{sing}}}$, $\{v_i\} \triangleq \{\lambda_i / d_{j(i)}^2\}_{i \in S_{\mathsf{many}}}$ and $\{d_i\} \triangleq \{d_{j(i)}^2\}_{i \in S_{\mathsf{many}}}$. This contradicts (11.12). □

We are finally ready to upper bound ζ.

Proof of Lemma 11.5.15. We can now upper bound ζ as follows. We have

$$\begin{aligned} \varepsilon &\geq \Omega(\zeta) \cdot \sum_{j \in \mathcal{J}'} 2\lfloor d_j/2 \rfloor \cdot 2^{-2/3(j+1)} d_j^{2/3} \\ &\geq \Omega(\zeta) \sum_{j \in \mathcal{J}'} 2^{-2j/3} d_j^{5/3} \end{aligned}$$

where in the first step we used (11.12) and Corollary 11.5.17, and in the second step we again used the fact that for $d_j > 1$, $2\lfloor d_j/2 \rfloor \geq 2d_j/3$. The claimed bound follows. □

We are now ready to complete the proof of Lemma 11.5.13:

Proof. Substituting our choice of $\{\varepsilon_j\}$ in Definition 11.5.14 into the lower bound of Lemma 11.5.7 gives

$$\begin{aligned}
\left(\sum_{j\in\mathcal{J}} 2^{2j}\|\boldsymbol{\mathcal{E}}_j\|_2^4/d_j\right)^{-1/2} &\geq \left(\sum_{j\in\mathcal{J}:d_j>1}\left\{\frac{2^{-2j-4}}{d_j} \wedge \zeta^4 2^{-2/3j-8/3} d_j^{5/3}\right\}\right)^{-1/2} \\
&\geq \left(\sum_{j\in\mathcal{J}:d_j>1}\left\{\zeta^3 2^{-j-3} d_j \wedge \zeta^4 2^{-2/3j} d_j^{5/3}\right\}\right)^{-1/2} \\
&\geq \Omega(\zeta^{-3/2})\left(\sum_{j\in\mathcal{J}:d_j>1} 2\lfloor d_j/2\rfloor\left\{2^{-j-1} \wedge \zeta 2^{-2/3(j+1)} d_j^{2/3}\right\}\right)^{-1/2} \\
&= \Omega(\zeta^{-3/2})\cdot\varepsilon^{-1/2} \\
&\geq \varepsilon^{-2}\cdot\left(\sum_{j\in\mathcal{J}',i\in S_j}\lambda_i^{2/3} d_j^{5/3}\right)^{3/2} \\
&\geq \max_{j\in\mathcal{J}',i\in S_j}\lambda_i d_j^{5/2}/\varepsilon^2 \\
&\geq \left(\sum_{j\in\mathcal{J}',i\in S_j}\lambda_i^{2/5} d_j\right)^{5/2}\cdot\log(d/\varepsilon)^{-1} \\
&\geq \|\sigma'\|_{2/5}\cdot\log(d/\varepsilon)^{-1},
\end{aligned}$$

where in the second step we used that the minimum of two nonnegative numbers increases if we replace one of them by a weighted geometric mean of the two numbers, in the third step we use the fact that $\lfloor d_j/2\rfloor$ and d_j are equivalent up to constant factors if $d_j > 1$, in the fourth step we use (11.12), in the fifth step we use (11.13), in the penultimate step we used Fact 11.5.3, and in the last step we used (11.11) and the fact that for any j, there are at most d_j indices $i \in S_{\mathsf{many}}\backslash S_{\mathsf{tail}}$ within bucket S_j.

□

Proof of Lemma 11.5.5. This follows immediately from Lemmas 11.5.12 and 11.5.13. □

11.5.3 Lower Bound Instance II: Perturbing Off-Diagonals

In many cases, the following lower bound instance will yield a stronger lower bound than the preceding argument, at the cost of applying to a limited range of ε. Take any $j, j' \in \mathcal{J}^*$ for which $d_j \geq d_{j'}$. As we will explain below, if $d_j > 1$, then j and j' need not be distinct.

If j and j' are distinct, then given a matrix $\mathbf{W}^{d_j \times d_{j'}}$ with orthonormal columns, let $\sigma_{\mathbf{W}}$ be the matrix $\sigma + D_{\mathbf{W}}$ where $D_{\mathbf{W}} \in \mathbb{C}^{d \times d}$ is the matrix which is zero outside of the principal submatrix indexed by $S_j \cup S_{j'}$ and which is equal to the matrix

$$\left(\begin{array}{c|c} \mathbf{0}_{d_j} & (\varepsilon/2d_{j'}) \cdot \mathbf{W} \\ \hline (\varepsilon/2d_{j'}) \cdot \mathbf{W}^{\dagger} & \mathbf{0}_{d_{j'}}. \end{array}\right) \tag{11.14}$$

On the other hand, if $j = j'$ and $d_j > 1$, then partition S_j into contiguous sets S^1_j, S^2_j of size $\lceil d_j/2 \rceil$ and $\lfloor d_j/2 \rfloor$, and given a matrix $\mathbf{W}^{\lceil d_j/2 \rceil \times \lfloor d_j/2 \rfloor}$ with orthonormal columns, define $D_{\mathbf{W}} \in \mathbb{C}^{d \times d}$ to be the matrix which is zero outside the principal submatrix indexed by $S^1_j \times S^2_j$ and which is equal to the matrix

$$\left(\begin{array}{c|c} \mathbf{0}_{\lceil d_j/2 \rceil} & (\varepsilon/2\lfloor d_j/2 \rfloor) \cdot \mathbf{W} \\ \hline (\varepsilon/2\lfloor d_j/2 \rfloor) \cdot \mathbf{W}^{\dagger} & \mathbf{0}_{d_{\lfloor d_j/2 \rfloor}}. \end{array}\right) \tag{11.15}$$

In the rest of this subsection, we will consider the case where $j \neq j'$, but as will become evident, all of the following arguments easily extend to the construction for $j = j'$ when $d_j > 1$ by replacing S_j and $S_{j'}$ with S^1_j and S^2_j respectively.

Lemma 11.5.18. *If $\varepsilon \leq d_{j'} \cdot 2^{-j/2 - j'/2}$, then $\|\sigma - \sigma_{\mathbf{W}}\|_1 \geq \varepsilon$ and $\sigma_{\mathbf{W}}$ is a density matrix.*

Proof. For the first part, note that

$$\|\sigma - \sigma_{\mathbf{W}}\|_1 = \|D_{\mathbf{W}}\| = 2 \cdot (\varepsilon/2d_{j'}) \|\mathbf{W}\|_1 = \varepsilon,$$

where in the second equality we used that $D_{\mathbf{W}}$ is the Hermitian dilation of $(\varepsilon/d_{j'}) \cdot \mathbf{W}$, and in the last equality we used the fact that $\mathbf{W}$ consists of $d_{j'}$ orthogonal columns.

For the second part, first note that regardless of the choice of ε, we have that $\mathrm{Tr}(D_{\mathbf{W}}) = 0$, so $\mathrm{Tr}(\sigma_{\mathbf{W}}) = 1$. Finally, to verify that $\sigma_{\mathbf{W}}$ is positive definite, note that the Schur complement of the principal submatrix of $\sigma_{\mathbf{W}}$ indexed by $S_j \cap S_{j'}$ is given by

$$\sigma_{j'} - \frac{\varepsilon^2}{4d_{j'}^2}\sigma_j^{-1} \succeq 2^{-j'-1}\,\mathrm{Id} - \frac{\varepsilon^2}{4d_{j'}^2}2^{j+1}\,\mathrm{Id},$$

which is positive definite provided that $\varepsilon \le d_{j'} \cdot 2^{-j/2-j'/2}$. It follows by Fact 11.3.4 that $\sigma_{\mathbf{W}}$ is positive definite as claimed. □

The objective of this subsection is to show the following lower bound:

Lemma 11.5.19. *Fix any $j, j' \in \mathcal{J}^*$ satisfying $d_j \ge d_{j'}$. If $d_j > 1$, then we can optionally take $j = j'$. Suppose $\varepsilon \le d_{j'} \cdot 2^{-j/2-j'/2}$. Let $\sigma \in \mathbb{C}^{d\times d}$ be a diagonal density matrix. Distinguishing between whether $\rho = \sigma$ or $\rho = \sigma_{\mathbf{W}}$ for $\mathbf{W} \in \mathbb{C}^{d_j \times d_{j'}}$ consisting of Haar-random orthonormal columns, using nonadaptive unentangled measurements, has copy complexity at least*

$$\Omega\left(\frac{\sqrt{d_j} \cdot d_{j'}^2 \cdot 2^{-j'}}{\varepsilon^2}\right).$$

Note that a random $\mathbf{W}$ is equivalent to $\mathbf{U}\Pi$ for $\mathbf{U} \sim \mathcal{D}$, where $\mathcal{D}$ is the Haar measure over $U(d_j)$, and

$$\Pi \triangleq (\underset{d_{j'}}{\mathrm{Id}}\,|\mathbf{0}_{d_j-d_{j'}})^\top,$$

so we can just as well parametrize $\{\sigma_{\mathbf{W}}\}$ as $\{\sigma_{\mathbf{U}}\}$, which we will do in the sequel.

Take any single-copy sub-problem $\mathcal{P} = (\mathcal{M}, \sigma, \{\sigma_{\mathbf{U}}\}_{\mathbf{U}\sim\mathcal{D}})$ where POVM $\mathcal{M}$ consists of elements $\{M_z\}$. Analogously to Lemma 11.5.6, we may without loss of generality assume that one of the POVM elements is the projector to the coordinates outside of $S_j \cup S_{j'}$, and the remaining POVM elements are rank-1 matrices $M_z = \lambda_z v_z v_z^\dagger$ where the $\lambda_z \le 1$ satisfy

$$\sum \lambda_z = d_j + d_{j'} < 2d_j \tag{11.16}$$

and the vectors v_z are unit vectors supported on $S_j \cap S_{j'}$. Let v_z^j and $v_z^{j'}$ denote the d_j- and

$d_{j'}$-dimensional components of v_z indexed by S_j and $S_{j'}$. Note that for these z,

$$g^{\mathbf{U}}_{\mathcal{P}}(z) = \frac{\langle M_z, D_{\mathbf{W}}\rangle}{\langle M_z, \sigma\rangle} = \frac{\varepsilon}{d_{j'}} \cdot \frac{\mathrm{Re}((v^j_z)^\dagger(\mathbf{U}\Pi)v^{j'}_z)}{v^\dagger_z \sigma v_z}. \tag{11.17}$$

while for the index z corresponding to the projector to $(S_j \cup S_{j'})^c$, $g^{\mathbf{U}}_{\mathcal{P}}(z) = 0$.

We now verify that Conditions 1, 2 and 3 hold.

Lemma 11.5.20. *For any z, $\mathbb{E}_{\mathbf{U}}[g^{\mathbf{U}}_{\mathcal{P}}(z)] = 0$.*

Proof. Clearly $\mathrm{Tr}(D_{\mathbf{W}}) = 0$, so by Fact 1.3.48, $\mathbb{E}_{\mathbf{W}}[g^{\mathbf{W}}(z)] = 0$. □

Lemma 11.5.21. $\mathbb{E}_{z,\mathbf{U}}[g^{\mathbf{U}}_{\mathcal{P}}(z)^2] \le O\left(\frac{\varepsilon^2}{d^2_{j'}2^{-j'}}\right)$*, where as usual, expectation is with respect to measurement outcomes when measuring the null hypothesis σ with $\mathcal{M}$.*

Proof. From (11.17) we have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{z,\mathbf{U}}\left[g^{\mathbf{U}}_{\mathcal{P}}(z)^2\right] &= \frac{\varepsilon^2}{d^2_{j'}} \mathop{\mathbb{E}}_{\mathbf{U}}\left[\sum_z \lambda_z v^\dagger_z \sigma v_z \left(\frac{\mathrm{Re}((v^j_z)^\dagger(\mathbf{U}\Pi)v^{j'}_z)}{v^\dagger_z\sigma v_z}\right)^2\right] \\
&= \frac{\varepsilon^2}{d^2_{j'}} \sum_z \frac{\lambda_z}{v^\dagger_z \sigma v_z} \mathop{\mathbb{E}}_{\mathbf{U}}\left[\left(\mathrm{Re}((v^j_z)^\dagger(\mathbf{U}\Pi)v^{j'}_z)\right)^2\right] \\
&= \frac{\varepsilon^2}{d^2_{j'}} \sum_z \frac{\lambda_z}{v^\dagger_z \sigma v_z} \cdot \frac{\|v^j_z\|^2 \|v^{j'}_z\|^2}{d_j}, \qquad (11.18)
\end{aligned}$$

As v_z is supported on $S_j \cup S_{j'}$, the supports of v^j_z and $v^{j'}_z$ are disjoint, and the diagonal entries of σ indexed by $S_{j'}$ are at least 2^{-j-1}, we have that $v^\dagger_z \sigma v_z \ge 2^{-j'-1}\|v^{j'}_z\|^2$ and $\|v^j_z\|^2_2 \le 1$, so we can further bound (11.18) by

$$= \frac{\varepsilon^2 2^{j'+1}}{d^2_{j'} d_j} \sum_z \lambda_z \le O\left(\frac{\varepsilon^2}{d^2_{j'}2^{-j'}}\right),$$

where the last step follows by (11.16). □

Lemma 11.5.22. $\mathbb{E}_z[(g^{\mathbf{U}_1}_{\mathcal{P}}(z) - g^{\mathbf{U}_2}_{\mathcal{P}}(z))^2] \le O\left(\frac{\varepsilon^2}{d^2_{j'}2^{-j}}\right) \cdot \|\mathbf{U}_1 - \mathbf{U}_2\|^2_F$ *for any $\mathbf{U}_1, \mathbf{U}_2 \in U(d_j)$.*

Proof. Define the matrix

$$\mathbf{D} = \begin{pmatrix} \mathbf{0}_{d_j} & (\varepsilon/2d_{j'}) \cdot (\mathbf{U}_1\Pi - \mathbf{U}_2\Pi) \\ (\varepsilon/2d_{j'}) \cdot (\mathbf{U}_1\Pi - \mathbf{U}_2\Pi)^\dagger & \mathbf{0}_{d_{j'}} \end{pmatrix}$$

Note that for any POVM element M_z,

$$\langle M_z, \mathbf{D}\rangle^2 = \frac{\lambda_z^2\varepsilon^2}{d_{j'}^2}\mathrm{Re}\left((v_z^j)^\dagger(\mathbf{U}_1 - \mathbf{U}_2)\Pi v_z^{j'}\right)^2 \le \frac{\lambda_z^2\varepsilon^2}{d_{j'}^2} \cdot \|v_z^j(\mathbf{U}_1 - \mathbf{U}_2)\|^2 \cdot \|v_z^{j'}\|_2^2 \qquad (11.19)$$

We can then write

$$\begin{aligned} \mathop{\mathbb{E}}_z[(g^{\mathbf{U}}_{\mathcal{P}}(z) - g^{\mathbf{V}}_{\mathcal{P}}(z))^2] &= \sum_z \frac{\langle M_z, \mathbf{D}\rangle^2}{\langle M_z, \sigma\rangle} \\ &\le \frac{\varepsilon^2}{d_{j'}^2}\sum_z \frac{\lambda_z\|v_z^j(\mathbf{U}_1 - \mathbf{U}_2)\|^2 \cdot \|v_z^{j'}\|^2}{2^{-j'-1}\|v_z^{j'}\|^2} \\ &\le O\left(\frac{\varepsilon^2 2^{j'}}{d_{j'}^2}\right) \cdot \sum_z \lambda_z\|v_z^j(\mathbf{U}_1 - \mathbf{U}_2)\|^2 \\ &= O\left(\frac{\varepsilon^2 2^{j'}}{d_{j'}^2}\right) \cdot \left\langle (\mathbf{U}_1 - \mathbf{U}_2)(\mathbf{U}_1 - \mathbf{U}_2)^\dagger, \sum_z \lambda_z v_z^j(v_z^j)^\dagger \right\rangle \\ &= O\left(\frac{\varepsilon^2}{d_{j'}^2 2^{-j'}}\right) \cdot \|\mathbf{U}_1 - \mathbf{U}_2\|_F^2, \end{aligned}$$

where in the second step we used (11.19) and the fact that $\langle M_z, \sigma\rangle = \lambda_z v_z^\dagger \sigma v_z \ge \lambda_z 2^{-j'-1}\|v_z^{j'}\|^2$, and in the fifth step we used that $\sum_z \lambda_z v_z^j(v_z^j)^\dagger = \mathrm{Id}_{d_j}$. □

Proof of Lemma 11.5.19. This follows from Lemma 11.4.9 with $L, \varsigma = O\left(\frac{\varepsilon}{d_{j'}2^{-j'/2}}\right)$. □

11.5.4 Lower Bound Instance III: Corner Case

We will also need the a lower bound instance that will yield an $\Omega(1/\varepsilon^2)$ lower bound for state certification with respect to any σ with maximum entry at least 1/2. We will not use anything about bucketing in this warmup result.

Let i_1 be the index of the largest entry of σ, and let i_2 be the index of the second-largest (breaking ties arbitrarily). For any $u \in \{\pm 1\}$, consider the state σ^u which agrees with σ

everywhere except in the principal submatrix indexed by $\{i_1, i_2\}$. Within that submatrix, define $\sigma^u_{i_1,i_1} = \sigma_i - \varepsilon^2/4$, $\sigma^u_{i_2,i_2} = \sigma_{i_2} + \varepsilon^2/4$, and $\sigma^u_{i_1,i_2} = {\sigma^u}^\dagger_{i_2,i_1} = (\varepsilon/2)u$.

Lemma 11.5.23. *If the maximum entry of σ is at least $3/4$, then for any $\varepsilon \leq 1/2$, $\|\sigma - \sigma^u\|_1 \geq \varepsilon$ and σ^u is a density matrix.*

Proof. Note that for $\varepsilon < 1/2$,

$$\|\sigma - \sigma^u\|_1 = \left\| \begin{pmatrix} -\varepsilon^2 & (\varepsilon/2)u \\ (\varepsilon/2)\overline{u} & \varepsilon^2 \end{pmatrix} \right\|_1 = 2\sqrt{\varepsilon^4/16 + \varepsilon^2/4} \geq \varepsilon.$$

For the second part of the lemma, clearly $\mathrm{Tr}(\sigma^u) = 1$. To verify that σ^u is psd, first note that because $\sigma_{i_1,i_1} \geq 3/4$ and $\sigma_{i_2,i_2} \leq 1/2$, and $\varepsilon^2/4 \leq 1/4$, every diagonal entry of σ^u is nonnegative. On the other hand, the principal submatrix indexed by $\{i_1, i_2\}$ has determinant $(\sigma_{i_1,i_1} - \varepsilon^2)(\sigma_{i_2,i_2} + \varepsilon^2) - \varepsilon^2/4 \geq (3/4 - \varepsilon^2)\varepsilon^2 - \varepsilon^2/4 \geq 0$, so σ^u is psd as claimed. □

The objective of this subsection is to show the following lower bound:

Lemma 11.5.24. *Let $\varepsilon \leq 1/2$. If the maximum entry of σ is at least $3/4$, then distinguishing between whether $\rho = \sigma$ or $\rho = \sigma^u$ for $u \sim \{\pm 1\}$, using nonadaptive unentangled measurements, has copy complexity at least $\Omega(1/\varepsilon^2)$. In fact, this holds even for adaptive unentangled measurements.*

Because we have no a priori bound on σ_{i_2,i_2}, the KL divergence between the distribution over outcomes from measuring N copies of σ^u for random $u \in \{\pm 1\}$ and the distribution from measuring N copies of σ may be arbitrarily large, so we cannot implement the strategy in Section 11.4. Instead, we will directly upper bound the total variation between these two distributions using the following basic fact:

Fact 11.5.25. *Given distributions p, q over a discrete domain S, if likelihood ratio $p(x)/q(x) \geq 1 - \nu$, then $d_{\mathrm{TV}}(p, q) \leq \nu$.*

Proof. We can write

$$d_{\mathrm{TV}}(p, q) = \sum_{x: p(x) \leq q(x)} |p(x) - q(x)| = \sum_{x: p(x) \leq q(x)} q(x) \cdot |p(x)/q(x) - 1| \leq \nu$$

as claimed. □

Proof of Lemma 11.5.24. Let $\mathcal{D}$ be the uniform distribution over $\{\pm 1\}$, and fix an arbitrary unentangled POVM schedule $\mathcal{S}$. Let p_0 denote the distribution over transcripts $z_{\le t}$ of outcomes upon measuring N copies of σ with $\mathcal{S}$, and let p_1 denote the distribution upon measuring N copies of σ^u, where $u \sim \mathcal{D}$. We will lower bound the likelihood ratio $p_1(z_{\le N})/p_0(z_{\le N})$ for *any* transcript $z_{\le N}$. Let $\mathcal{M}^{(1)}, \dots, \mathcal{M}^{(N)}$ denote the (possibly adaptively chosen) POVMs that were used in the course of generating $z_{\le N}$.

For any $t \in [N]$, suppose $\mathcal{M}^{(t)}$ consists of elements $\{M^{(t)}_z\}$. Analogously to Lemma 11.5.6, we may without loss of generality assume that one element of $\mathcal{M}^{(t)}$ is the projector to the coordinates outside of $\{i_1, i_2\}$, and the remaining elements are rank-1 matrices $M^{(t)}_z = \lambda^{(t)}_z v^{(t)}_z (v^{(t)}_z)^\dagger$ where the $\lambda^{(t)}_z \le 1$ satisfy $\sum \lambda^{(t)}_z = 2$ and the vectors $v^{(t)}_z$ are unit vectors supported on $\{i_1, i_2\}$. Let $v^{(t)}_{z_t,1}$ and $v^{(t)}_{z_t,2}$ denote the coordinates of $v^{(t)}_z$ indexed by i_1 and i_2.

Note that for any $u \in \{\pm 1\}$ and $t \in [N]$, if z_t does not correspond to the projector to the coordinates outside of $\{i_1, i_2\}$, we can write

$$\Delta^u_t(z_t) \triangleq \frac{\langle M^{(t)}_{z_t}, \sigma^u\rangle}{\langle M^{(t)}_{z_t}, \sigma\rangle} = 1 + \frac{\varepsilon u \mathrm{Re}\left(\overline{v^{(t)}_{z_t,1}} v^{(t)}_{z_t,2}\right) - \varepsilon^2\left(\left|v^{(t)}_{z_t,1}\right|^2 - \left|v^{(t)}_{z_t,2}\right|^2\right)}{v^{(t)\dagger}_{z_t} \sigma v^{(t)}_{z_t}}$$

and if z_t does correspond to the projector, then $\Delta^u_t(z_t) = 1$.

Denoting the t-th entry of $z_{\le N}$ by z_t, we can use AM-GM to bound the likelihood ratio by

$$\begin{aligned}\frac{p_1(z_{\le N})}{p_0(z_{\le N})} &= \mathop{\mathbb{E}}_u\left[\prod_{t=1}^N \Delta^u_t(z_t)\right] \\ &\ge \left(\prod_{t=1}^N \Delta^{+1}_t(z_t)\Delta^{-1}_t(z_t)\right)^{1/2} \qquad (11.20)\end{aligned}$$

To prove the lemma, we will lower bound this by $1 - o(1)$. Because $\Delta_t^u(z_t) = 1$ if z_t corresponds to the projector to the coordinates outside of $\{i_1, i_2\}$, we may assume without loss of generality that this is not the case for any $t \in [N]$. We can then further bound (11.20) by

$$\geq \prod_{t=1}^{N} \left\{ \left(1 - \frac{\varepsilon^2 \left(\left| v_{z_t,1}^{(t)} \right|^2 - \left| v_{z_t,2}^{(t)} \right|^2 \right)}{v_{z_t}^{(t)\dagger} \sigma v_{z_t}^{(t)}} \right)^2 - \frac{\varepsilon^2 \mathrm{Re} \left(\overline{v_{z_t,1}^{(t)}} v_{z_t,2}^{(t)} \right)^2}{\left(v_{z_t}^{(t)\dagger} \sigma v_{z_t}^{(t)} \right)^2} \right\}^{1/2}. \quad (11.21)$$

For any $v \in \mathbb{C}^d$ which has entries v_1 and v_2 in coordinates i_1 and i_2 and is zero elsewhere, we have that

$$\frac{|v_1|^2 - |v_2|^2}{v^\dagger \sigma v} \leq \frac{|v_1|^2}{\sigma_{i_1,i_1} |v_1|^2} \leq 4/3 \qquad \frac{\mathrm{Re}(\overline{v_1} v_2)^2}{v^\dagger \sigma v} \leq \frac{\mathrm{Re}(\overline{v_1} v_2)^2}{\sigma_{i_1,i_1} |v_1|^2} \leq 4/3,$$

where the last step for both estimates follows by the assumed lower bound on σ_{i_1,i_1}. By (11.21) we have that

$$\frac{p_1(z_{\leq N})}{p_0(z_{\leq N})} \geq ((1 - 4\varepsilon^2/3)^2 - 4\varepsilon^2/3)^{N/2} \geq (1 - 32\varepsilon^2/9)^{N/2}.$$

In particular, for $N = o(1/\varepsilon^2)$, the likelihood ratio is at least $1 - o(1)$ as desired. □

11.5.5 Putting Everything Together

We are now ready to conclude the proof of Theorem 11.5.1.

Proof of Theorem 11.5.1. We proceed by casework depending on whether or not $d_j = 1$ for all $j \in \mathcal{J}^*$.

Case 3. *$d_j = 1$ for all $j \in \mathcal{J}^*$.*

There are two possibilities. If there is a single bucket $j = j(i)$ for which $i \notin S_{\mathsf{tail}} \cup S_{\mathsf{light}}$, then $d_{\mathsf{eff}} = 1$ and $\|\sigma^{**}\|_{1/2} = O(1)$. For ε smaller than some absolute constant, we know that $\sigma_{i,i} \geq 3/4$ and can apply Lemma 11.5.24 to conclude a lower bound of $\Omega(1/\varepsilon^2)$ as desired. Otherwise, let j' be the smallest index for which $j' = j(i')$ for some $i' \in \mathcal{J}^*$, and let $j > j'$

be the next smallest index for which $j = j(i)$ for some $i \in \mathcal{J}^*$. Consider the lower bound instance in Section 11.5.3 applied to this choice of j, j'. Provided that $\varepsilon \le 2^{-j/2-j'/2}$, we would obtain a copy complexity lower bound of $\Omega(2^{-j'}/\varepsilon^2) \ge \Omega(\|\sigma^*\|_{1/2}/(\varepsilon^2 \log(d/\varepsilon)))$, where the inequality is by Fact 11.3.1, and we would be done. On the other hand, if $\varepsilon \ge 2^{-j/2-j'/2}$, then because $2^{-j'} > 2^{-j}$, we would conclude that $2^{-j} \le \varepsilon$. In particular, this implies that $\sum_{j'' \in \mathcal{J}^*, i \in S_{j''}: j'' \neq j'} \lambda_i \le 2\varepsilon$, so after removing at most an additional 2ε mass from σ^*, we get a matrix σ^{**} (see Definition 11.5.2) with a single nonzero entry. Again, $d_{\mathsf{eff}} = 1$ and $\|\sigma^{**}\|_{1/2} = O(1)$, and if ε is smaller than some absolute constant, we conclude that that single nonzero entry is at least $3/4$ and can apply Lemma 11.5.24 to conclude a lower bound of $\Omega(1/\varepsilon^2)$ as desired.

Case 4. *$d_j > 1$ for some $j \in \mathcal{J}^*$.*

Let $j_* \triangleq \arg\max_{j \in \mathcal{J}^*} d_j$ and $j'_* \triangleq \arg\max_{j \in \mathcal{J}^*} d_j^2 2^{-j}$. By Lemma 11.5.19, we have a lower bound of $\Omega\left(\sqrt{d_{j_*}} \cdot d_{j'_*}^2 \cdot 2^{-j'_*}/\varepsilon^2\right)$ as long as ε satisfies the bound

$$\varepsilon \le d_{j'_*} \cdot 2^{-j_*/2-j'_*/2}. \tag{11.22}$$

Note that because $d_{j_*} > 1$ as we are in Case 2, we do not constrain j_*, j'_* to be distinct necessarily. We would now like to argue that this lower bound, up to log factors, holds even if the bound on ε in (11.22) does not hold. In the following, assume that (11.22) does not hold.

To this end, we will also use the lower bound from Lemma 11.5.5 of $\Omega(\|\sigma'\|_{2/5}/(\varepsilon^2 \log(d/\varepsilon)))$. We would first like to relate $\|\sigma'\|_{2/5}$ to $\|\sigma^*\|_{2/5}$.

Lemma 11.5.26. *Either $\|\sigma'\|_{2/5} \ge \Omega(\|\sigma^*\|_{2/5})$, or the following holds. Let j° be the index maximizing $d_j^{5/2} 2^{-j}$. Then 1) $j^\circ = \min_{j \in \mathcal{J}^*} j$, 2) $d_{j^\circ} = 1$, and 3) $j^\circ = 0$.*

Proof. We will assume that $\|\sigma'\|_{2/5} = o(\|\sigma^*\|_{2/5})$ and show that 1), 2), and 3) must hold. Let j° be the index maximizing $d_j^{5/2} 2^{-j}$, and let $i_{\max}$ be the index of the top entry of σ^*. Let σ'' denote the matrix obtained by zeroing out the top entry of σ^*. Note that the nonzero entries of σ' comprise a superset of those of σ'', so

$$\frac{\|\sigma^*\|_{2/5}^{2/5}}{\|\sigma'\|_{2/5}^{2/5}} \le \frac{\|\sigma^*\|_{2/5}^{2/5}}{\|\sigma''\|_{2/5}^{2/5}} = \frac{\sum_{i\in\mathcal{J}^*}\sigma_i^{2/5}}{\sum_{i\in\mathcal{J}^*\backslash\{i_{\max}\}}\sigma_i^{2/5}}.$$

Suppose 1) does not hold. Then

$$\frac{\sum_{i\in\mathcal{J}^*}\sigma_i^{2/5}}{\sum_{i\in\mathcal{J}^*\backslash\{i_{\max}\}}\sigma_i^{2/5}} \le \frac{\sigma_{i_{\max}}^{2/5} + \sum_{i\in S_{j^\circ}}\sigma_i^{2/5}}{\sum_{i\in S_{j^\circ}}\sigma_i^{2/5}} \le 2,$$

where the first inequality follows by the elementary fact that for positive integers a, b, c, $\frac{a+c}{b+c} \le \frac{a}{b}$, and the second inequality follows by the definition of j°.

Next, suppose 1) holds but 2) does not hold. Then

$$\frac{\sum_{i\in\mathcal{J}^*}\lambda_i^{2/5}}{\sum_{i\in\mathcal{J}^*\backslash\{i_{\max}\}}\lambda_i^{2/5}} \le \frac{\sum_{i\in S_{j^\circ}}\lambda_i^{2/5}}{\sum_{i\in S_{j^\circ}\backslash\{i_{\max}\}}\lambda_i^{2/5}} \le O(1),$$

where the first inequality again uses the above elementary fact, the second inequality follows by our assumption that 2) does not hold. This yields a contradiction.

Finally suppose 1) and 2) hold, but 3) does not, so that $\|\sigma^*\|_\infty \le 1/2$. Let σ'' denote the matrix obtained by zeroing out the top entry of σ^*. We would have

$$\|\sigma''\|_{2/5} \ge \|\sigma''\| \ge 1/2 - O(\varepsilon),$$

so for ε smaller than a sufficiently large absolute constant, we would have that $\|\sigma''\|_{2/5}^{2/5} \ge \Omega(\|\sigma^*\|_\infty^{2/5})$ and therefore $\|\sigma'\|_{2/5} \ge \|\sigma''\|_{2/5} \ge \Omega(\|\sigma^*\|_{2/5})$, a contradiction. □

Suppose the latter scenario in Lemma 11.5.26 happens but the former does not. In this case, because $d_{j^\circ} = 1$, we also have that $j'_* = \arg\max_{j\in\mathcal{J}^*} d_j^2 2^{-j}$, i.e. $j'_* = j^\circ$. In particular,

$$1 \ge d_{j'_*}^2 2^{-j'_*} \ge d_{j_*}^2 2^{-j_*} \ge \Omega(d_{j_*}^{3/2}\varepsilon/\log(d/\varepsilon)), \tag{11.23}$$

where the last inequality follows by the fact that $d_j 2^{-j} \ge \Omega(\varepsilon/\log(d/\varepsilon))$ for all $j \in \mathcal{J}^*$ by design. We conclude that $\varepsilon \le O(d_{j_*}^{-3/2}\log(d/\varepsilon))$. But recall that we are assuming that

(11.22) is violated, i.e. that

$$\varepsilon > d_{j'_*} \cdot 2^{-j_*/2-j'_*/2} = 2^{-j_*/2-j'_*/2} \geq \Omega(\varepsilon/(d_{j_*}\log(d/\varepsilon)))^{1/2}, \tag{11.24}$$

where the last step is by 3) in Lemma 11.5.26 and the fact that $d_j 2^{-j} \geq \Omega(\varepsilon/\log(d/\varepsilon))$ for all $j \in \mathcal{J}^*$. Combining (11.23) and (11.24), we get a contradiction of the assumption that the former scenario in Lemma 11.5.26 does not hold, unless $d_{j_*} \leq \mathrm{polylog}(d/\varepsilon)$. But if $d_{j_*} \leq \mathrm{polylog}(d/\varepsilon)$, then the lower bound claimed in Theorem 11.5.1 still holds as $d_{\mathsf{eff}} \leq O(\log(d/\varepsilon)\cdot d_{j_*}) \leq \mathrm{polylog}(d/\varepsilon)$.

Finally, suppose instead that the former scenario in Lemma 11.5.26 happens, so that Lemma 11.5.5 gives a lower bound of $\Omega(\|\sigma^*\|_{2/5}/(\varepsilon^2\log(d/\varepsilon)))$. Let j° still be as defined in Lemma 11.5.26.

Now we would certainly be done if this lower bound were, up to log factors, larger than the one guaranteed by Lemma 11.5.19 to begin with. So suppose to the contrary. We would get that

$$d_{j_*}^{5/2}2^{-j_*} \leq d_{j^\circ}^{5/2}2^{-j^\circ} \leq \frac{1}{\log^2(d/\varepsilon)}\sqrt{d_{j_*}}d_{j'_*}^2 \cdot 2^{-j'_*},$$

implying that

$$d_j^2 2^{-j_*} \leq \frac{1}{\log^2(d/\varepsilon)} d_{j'_*}^2 2^{-j'_*}. \tag{11.25}$$

If (11.22) does not hold, then

$$\frac{1}{\log(d/\varepsilon)} \cdot d_{j'_*} \cdot 2^{-j_*/2-j'_*/2} \leq \frac{\varepsilon}{\log(d/\varepsilon)} \leq d_j 2^{-j},$$

where in the last step we again used the fact that $d_j 2^{-j} > \varepsilon/\log(d/\varepsilon)$ for all $j \in \mathcal{J}^*$, yielding the desired contradiction with (11.25) upon rearranging.

Having lifted the constraint (11.22), we finally note that by Fact 11.3.1,

$$\Omega\left(\sqrt{d_{j_*}} \cdot d_{j'_*}^2 \cdot 2^{-j'_*}/\varepsilon^2\right) \geq \Omega\left(\sqrt{d_{\mathsf{eff}}} \cdot \|\sigma^*\|_{1/2}/(\varepsilon^2\,\mathrm{polylog}(d/\varepsilon))\right).$$

The proof is complete upon invoking Fact 11.5.27 below. □

Fact 11.5.27. *Given psd matrix $\sigma \in \mathbb{C}^{d\times d}$, let $\widehat{\sigma} \triangleq \sigma/\,Tr(\sigma)$. Then $\|\sigma\|_{1/2} = d\,Tr(\sigma)^2 \cdot$*

$F(\widehat{\sigma}, \rho_{\mathsf{mm}})$.

Proof. We may assumed without loss of generality that σ is diagonal. By definition

$$F(\widehat{\sigma}, \rho_{\mathsf{mm}}) = \left(\operatorname{Tr}\sqrt{\sqrt{\widehat{\sigma}}(\operatorname{Id}/d)\sqrt{\widehat{\sigma}}}\right)^2 = \left(\frac{1}{\sqrt{d}\operatorname{Tr}(\sigma)} \cdot \operatorname{Tr}(\sqrt{\sigma})\right)^2 = \frac{1}{d\operatorname{Tr}(\sigma)^2} \cdot \|\sigma\|_{1/2},$$

from which the claim follows. □

11.6 State Certification Algorithm

In this section we prove the following upper bound on state certification that nearly matches the lower bound proven in Section 11.5:

Theorem 11.6.1. *Fix $\varepsilon, \delta > 0$. Let $\rho \in \mathbb{C}^{d\times d}$ be an unknown mixed state, and let $\sigma \in \mathbb{C}^{d\times d}$ be a diagonal density matrix. Let σ' be the matrix given by zeroing out the bottom $O(\varepsilon^2)$ mass in σ (see Definition 11.6.5 below). Let $\widehat{\sigma}' \triangleq \sigma'/\operatorname{Tr}(\sigma')$ and let d_{eff} be the number of nonzero entries of σ'.*

Given an explicit description of σ and copy access to ρ, Certify *takes*

$$N = O(d\sqrt{d_{\mathsf{eff}}} \cdot F(\widehat{\sigma}', \rho_{\mathsf{mm}}) \operatorname{polylog}(d/\varepsilon) \log(1/\delta)/\varepsilon^2)$$

copies of ρ and, using unentangled nonadaptive measurements, distinguishes between $\rho = \sigma$ and $\|\rho - \sigma\|_1 > \varepsilon$ with probability at least $1 - \delta$.

First, in Section 11.6.1 we give a generic algorithm for state certification based on measuring in a Haar-random basis and applying classical identity testing. In Section 11.6.2, we describe a bucketing scheme that will be essential to the core of our analysis in Section 11.6.3, where we use this tool to obtain the algorithm in Theorem 11.6.1.

11.6.1 Generic Certification

The main result of this section is a basic state certification algorithm that will be invoked as a subroutine in our instance-near-optimal certification algorithm:

Lemma 11.6.2. *Fix $\varepsilon, \delta > 0$. Let $\rho, \sigma \in \mathbb{C}^{d\times d}$ be two mixed states. Given access to an explicit description of σ and copy access to ρ,* BasicCertify *takes $N = O(\sqrt{d}\log(1/\delta)/\varepsilon^2)$ copies of ρ and, using unentangled nonadaptive measurements, distinguishes between $\rho = \sigma$ and $\|\rho - \sigma\|_F > \varepsilon$ with probability at least $1 - \delta$.*

Algorithm 47: BasicCertify$(\rho, \sigma, \varepsilon, \delta)$

```
Input: Copy access to $\rho$, diagonal density matrix $\sigma$, error $\varepsilon$, failure probability $\delta$
Output: YES if $\rho = \sigma$, NO if $\|\rho - \sigma\|_F > \varepsilon$, with probability $1 - \delta$
$N \leftarrow O(\sqrt{d}/\varepsilon^2)$.
for $T = 1, \ldots, O(\log(1/\delta))$ do
    Sample a Haar-random unitary matrix $\mathbf{U}$.
    Form the POVM $\mathcal{M}$ consisting of $\{|\mathbf{U}_1\rangle\langle\mathbf{U}_1|, \ldots, |\mathbf{U}_d\rangle\langle\mathbf{U}_d|\}$.
    Measure each copy of $\rho$ with $\mathcal{M}$, yielding outcomes $z_1, \ldots, z_N$.
    Let $q \in \Delta^d$ denote the distribution over outcomes from measuring $\sigma$ with $\mathcal{M}$.
    Draw i.i.d. samples $z'_1, \ldots, z'_N$ from $q$.
    $b_i \leftarrow$L2Tester$(\{z_i\}, \{z'_i\})$.
return majority among $b_1, \ldots, b_T$.
```

To prove Lemma 11.6.2, we will need the following result from classical distribution testing.

Lemma 11.6.3 (Lemma 2.3 from [DK16])**.** *Let p, q be two unknown distributions on $[d]$ for which $\|p\|_2 \wedge \|q\|_2 \le b$ for some $b > 0$. There exists an algorithm* L2Tester *that takes $N = O(b\log(1/\delta)/\varepsilon^2)$ samples from each of p and q and distinguishes between $p = q$ and $\|p - q\|_2 > \varepsilon$ with probability at least $1 - \delta$.*[4]

We will also need the following moment calculations:

Lemma 11.6.4. *For any Hermitian $\mathbf{M} \in \mathbb{C}^{d\times d}$ and Haar-random $\mathbf{U} \in U(d)$, let Z denote the random variable $\sum_{i=1}^d \left(\mathbf{U}_i^\dagger \mathbf{M}\mathbf{U}_i\right)^2$. Then*

$$\mathbb{E}[Z] = \frac{1}{d+1}\left(Tr(\mathbf{M})^2 + \|\mathbf{M}\|_F^2\right).$$

[4] Note that Lemma 2.3 in [DK16] only gives a constant probability guarantee, but the version we state follows by a standard amplification argument.

If in addition we have that Tr$(\mathbf{M}) = 0$*, then*

$$\mathbb{E}[Z^2] \le \frac{1+o(1)}{d^2}\|\mathbf{M}\|_F^4.$$

Proof. By symmetry $\mathbb{E}[Z] = d\,\mathbb{E}[(\mathbf{U}_1\mathbf{M}\mathbf{U}_1)^2]$, and by Lemma 1.3.49, if Π denotes the projector to the first coordinate,

$$\mathbb{E}[(\mathbf{U}_1\mathbf{M}\mathbf{U}_1)^2] = \sum_{\pi,\tau\in\mathcal{S}_2} \mathrm{Wg}(\pi\tau^{-1},d)\langle\Pi\rangle_\pi\langle\mathbf{M}\rangle_\tau = \frac{1}{d(d+1)}(\mathrm{Tr}(\mathbf{M})^2 + \mathrm{Tr}(\mathbf{M}^2)),$$

from which the first part of the lemma follows.

For the second part, let $\mathcal{S}_4^* \subset \mathcal{S}_4$ denote the set of permutations π for which $\pi(1),\pi(2) \in \{1,2\}$ and $\pi(3),\pi(4) \in \{3,4\}$. Note that

$$\mathbb{E}[Z^2] = d\cdot\mathbb{E}\Big[(\mathbf{U}_1^\dagger\mathbf{M}\mathbf{U}_1)^4\Big] + (d^2-d)\cdot\mathbb{E}\Big[(\mathbf{U}_1^\dagger\mathbf{M}\mathbf{U}_1)^2(\mathbf{U}_2^\dagger\mathbf{M}\mathbf{U}_2)^2\Big]. \tag{11.26}$$

For the first term, by Lemma 1.3.49 we have

$$\begin{aligned}\mathbb{E}[(\mathbf{U}_1^\dagger\mathbf{M}\mathbf{U}_1)^4] &= \sum_{\pi,\tau\in\mathcal{S}_4}\mathrm{Wg}(\pi\tau^{-1},d)\langle\mathbf{M}\rangle_\tau \\ &= \frac{1}{d(d+1)(d+2)(d+3)}\sum_\tau\langle\mathbf{M}\rangle_\tau \\ &= \frac{1}{d(d+1)(d+2)(d+3)}\sum_{\tau\text{ derangement}}\langle\mathbf{M}\rangle_\tau \\ &\le \frac{O(\|\mathbf{M}\|_F^4)}{d(d+1)(d+2)(d+3)},\end{aligned}$$

where the third step follows by the fact that $\mathrm{Tr}(\mathbf{M}) = 0$, and the fourth by the fact that for any derangement $\tau \in \mathcal{S}_4$, either $\langle\mathbf{M}\rangle_\tau = \mathrm{Tr}(\mathbf{M}^2)^2 = \|\mathbf{M}\|_F^4$, or $\langle\mathbf{M}\rangle_\tau = \mathrm{Tr}(\mathbf{M}^4) \le \|\mathbf{M}\|_F^4$. Similarly,

$$\begin{aligned}\mathbb{E}[(\mathbf{U}_1^\dagger\mathbf{M}\mathbf{U}_1)^2(\mathbf{U}_2^\dagger\mathbf{M}\mathbf{U}_2)^2] &= \sum_{\pi\in\mathcal{S}_4^*,\tau\in\mathcal{S}_4}\mathrm{Wg}(\pi\tau^{-1},d)\langle\mathbf{M}\rangle_\tau \\ &= \sum_{\tau\in\mathcal{S}_4^*}\mathrm{Wg}(e,d)\langle\mathbf{M}\rangle_\tau + \sum_{\pi\in\mathcal{S}_4^*,\tau\in\mathcal{S}_4:\tau\neq\pi}\mathrm{Wg}(\pi\tau^{-1},d)\langle\mathbf{M}\rangle_\tau\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{Wg}(e,d)\|\mathbf{M}\|_F^4 + \sum_{\pi\in\mathcal{S}_4^*,\tau\in\mathcal{S}_4:\tau\neq\pi} \mathrm{Wg}(\pi\tau^{-1},d)\langle\mathbf{M}\rangle_\tau \\
&\le \frac{d^4-8d^2+6}{d^2(d^6-14d^4+49d^2-36)}\|\mathbf{M}\|_F^4 + O(1/d^5)\cdot\|\mathbf{M}\|_F^4 \\
&= \frac{1+o(1)}{d^4}\|\mathbf{M}\|_F^4,
\end{aligned}
$$

where in the second step $\mathrm{Wg}(e,d)$ denotes the Weingarten function corresponding to the identity permutation, in the third step we used the fact that the only $\tau\in\mathcal{S}_4^*$ which is a derangement is the permutation that interchanges 1 with 2, and 3 with 4, and in the fourth step we used the form of $\mathrm{Wg}(e,d)$, the fact that $|\mathrm{Wg}(\pi\tau^{-1},d)| = O(1/d^5)$ for $\pi\neq\tau$, and the fact that $\langle\mathbf{M}\rangle_\tau \le \|\mathbf{M}\|_F^4$. The second part of the lemma follows from (11.26). □

We can now complete the proof of Lemma 11.6.2.

Proof of Lemma 11.6.2. Let p and q be the distribution over d outcomes when measuring ρ and σ respectively using the POVM defined in a single iteration of the main loop of BasicCertify. Applying both parts of Lemma 11.6.4 to $\mathbf{M}=\rho-\sigma$, for which the random variable Z is $\|p-q\|_2^2$, we conclude that for some sufficiently small absolute constant $c>0$, $\Pr[\|p-q\|_2 \ge c\|\mathbf{M}\|_F/\sqrt{d}] \ge 5/6$. Applying the first part of Lemma 11.6.4 to $\mathbf{M}=\rho$ and $\mathbf{M}=\sigma$, for which the random variable Z is $\|p\|_2^2$ and $\|q\|_2^2$ respectively, we have that $\mathbb{E}[\|p\|_2^2], \mathbb{E}[\|q\|_2^2] \le 2/d$, so by Markov's, for some absolute constant $c'>0$, $\|p\|, \|q\|_2 \le c'/\sqrt{d}$ with probability at least 5/6. We can substitute these bounds for $\|p\|_2, \|q\|_2, \|p-q\|_2$ into Lemma 11.6.3 to conclude that the output of L2Tester is correct with some constant advantage. Repeating this $O(\log(1/\delta))$ times and taking the majority among all the outputs from L2Tester gives the desired high-probability guarantee. □

11.6.2 Bucketing and Mass Removal

We may without loss of generality assume that σ is the diagonal matrix $\mathrm{diag}(\lambda_1,\dots,\lambda_d)$, where $\lambda_1\le\cdots\le\lambda_d$.

We will use the bucketing procedure outlined in Section 11.5.1. The way that we remove a small amount of mass from the spectrum of σ slightly differs from that outlined in Definition 11.5.2 for our lower bound. Our bucketing and mass removal procedure is as follows:

Definition 11.6.5 (Removing low-probability elements- upper bound)**.** *Let $d' \le d$ denote the largest index for which $\sum_{i=1}^{d'} \lambda'_i \le \varepsilon^2/20$,[5] and let $S_{\mathsf{tail}} \triangleq [d']$. Let σ' denote the matrix given by zeroing out the diagonal entries of σ indexed by S_{tail}. For $j \in \mathbf{Z}_{\ge 0}$, let S_j denote the indices $i \notin S_{\mathsf{tail}}$ for which $\lambda_i \in [2^{-j-1}, 2^{-j}]$, and denote $|S_j|$ by d_j. Let $\mathcal{J}$ denote the set of j for which $S_j \neq \emptyset$.*

As in the proofs of our lower bounds, we use the following basic consequence of bucketing:

Fact 11.6.6. *There are at most* $\log(10d/\varepsilon^2)$ *indices $j \in \mathcal{J}$.*

Proof. The largest element among $\{\lambda_i\}_{i \in S_{\mathsf{tail}}}$ is at least $\varepsilon^2/10d$, from which the claim follows. □

We now introduce some notation. Let $m \triangleq \log(10d/\varepsilon^2)$ denote this upper bound on the number of buckets in $\mathcal{J}$. For $j \in \mathcal{J}$, let $\rho[j,j], \sigma[j,j] \in \mathbb{C}^{d\times d}$ denote the Hermitian matrices given by zeroing out entries of ρ, σ outside of the principal submatrix indexed by S_j. For distinct $j, j' \in \mathcal{J}$, let $\rho[j,j'] \in \mathbb{C}^{d\times d}$ denote the Hermitian matrix given by zeroing out entries of ρ outside of the two non-principal submatrices with rows and columns indexed by S_i and S_j, and by S_j and S_i. Lastly, let $\widehat{\rho}[j,j], \widehat{\sigma}[j,j], \widehat{\rho}[j,j'], \widehat{\sigma}[j,j']$ denote these same matrices but with trace normalized to 1.

Let $\rho^{\mathsf{diag}}_{\mathsf{junk}} \in \mathbb{C}^{d\times d}$ be the principal submatrix of ρ indexed by S_{tail}, and let $\rho^{\mathsf{off}}_{\mathsf{junk}} \in \mathbb{C}^{d\times d}$ be the matrix given by zeroing out the principal submatrices indexed by S_{tail} and by $[d]\backslash S_{\mathsf{tail}}$.

Lastly, we will need the following basic fact:

Fact 11.6.7. *Given two psd matrices ρ, σ, if $|Tr(\rho) - Tr(\sigma)| \le \varepsilon/2$ and $\|\rho - \sigma\| \ge \varepsilon$, then*

$$\|\rho/\ Tr(\rho) - \sigma/\ Tr(\sigma)\|_1 \ge \varepsilon/2\ Tr(\rho).$$

Proof. Note that

$$\|\sigma/\operatorname{Tr}(\rho) - \sigma/\operatorname{Tr}(\sigma)\|_1 = \left|\frac{\operatorname{Tr}(\sigma)}{\operatorname{Tr}(\rho)} - 1\right| \le \frac{\varepsilon}{2\operatorname{Tr}(\rho)},$$

[5] We made no effort to optimize this constant factor.

so by triangle inequality,

$$\|\rho/\operatorname{Tr}(\rho)-\sigma/\operatorname{Tr}(\sigma)\|_1 \geq \frac{1}{\operatorname{Tr}(\rho)}\|\rho-\sigma\|_1 - \|\sigma/\operatorname{Tr}(\rho)-\sigma/\operatorname{Tr}(\sigma)\|_1 \geq \frac{\varepsilon}{2\operatorname{Tr}(\rho)}.$$

□

11.6.3 Instance-Near-Optimal Certification

We are ready to prove Theorem 11.6.1.

Proof of Theorem 11.6.1. We have that

$$\rho = \sum_{j\in\mathcal{J}}\rho[j,j] + \sum_{j\in\mathcal{J}:j\neq j'}\rho[j,j'] + \rho^{\mathsf{diag}}_{\mathsf{junk}} + \rho^{\mathsf{off}}_{\mathsf{junk}} \qquad \sigma' = \sum_{j\in\mathcal{J}}\sigma[j,j]$$

If $\|\rho-\sigma\|_1 > \varepsilon$, then by triangle inequality,

$$\left\|\sum_{j\in\mathcal{J}}(\rho[j,j]-\sigma[j,j]) + \sum_{j,j'\in\mathcal{J}:j\neq j'}\rho[j,j'] + \rho^{\mathsf{diag}}_{\mathsf{junk}} + \rho^{\mathsf{off}}_{\mathsf{junk}}\right\|_1 = \|\rho-\sigma'\|_1 \geq \varepsilon - \varepsilon^2/20 \geq 9\varepsilon/10$$

and one of four things can happen:

1. $\|\rho^{\mathsf{diag}}_{\mathsf{junk}}\|_1 \geq \varepsilon^2/8$.

2. $\|\rho^{\mathsf{off}}_{\mathsf{junk}}\|_1 \geq \varepsilon/2$,

3. There exists $j\in\mathcal{J}$ for which $\|\rho[j,j]-\sigma[j,j]\|_1 \geq \varepsilon/(10m^2)$

4. There exist distinct $j,j'\in\mathcal{J}$ for which $\|\rho[j,j']\|_1 \geq \varepsilon/(5m^2)$.

Otherwise we would have

$$\|\rho-\sigma'\|_1 \leq m\cdot\frac{\varepsilon}{10m^2} + \binom{m}{2}\cdot\frac{\varepsilon}{5m^2} + \frac{\varepsilon^2}{8} + \frac{\varepsilon}{2} = \frac{\varepsilon}{10m} + \frac{\varepsilon(m-1)}{10m} + \frac{3\varepsilon}{4} < 9\varepsilon/10,$$

a contradiction.

It remains to demonstrate how to test whether we are in any of Scenarios 1 to 4.

Lemma 11.6.8. *$O(\log(1/\delta)/\varepsilon^2)$ copies suffice to test whether $\rho = \sigma$ or whether Scenario 1 holds, with probability $1 - O(\delta)$.*

Proof. We can use the POVM consisting of the projector Π to the principal submatrix indexed by S_{tail}, together with $\mathrm{Id} - \Pi$, to distinguish between whether $\mathrm{Tr}(\rho^{\mathsf{diag}}_{\mathsf{junk}}) \geq \varepsilon^2/8$ or whether $\mathrm{Tr}(\rho^{\mathsf{diag}}_{\mathsf{junk}}) \leq \varepsilon^2/10$, the latter of which holds if $\rho = \sigma$ by definition of S_{tail}. For this distinguishing task, $O(\log(1/\delta)/\varepsilon^2)$ copies suffice. □

Lemma 11.6.9. *If Scenario 1 does not hold, then Scenario 2 cannot hold.*

Proof. Suppose Scenario 1 does not hold so that $\|\rho^{\mathsf{diag}}_{\mathsf{junk}}\|_1 < \varepsilon^2/4$. Then by the first part of Fact 11.3.5, $\|\rho^{\mathsf{off}}_{\mathsf{junk}}\|_1^2 < (1 - \varepsilon^2/4) \cdot \varepsilon^2/4 < \varepsilon^2/4$, a contradiction. □

Lemma 11.6.10. *$O\left(\|\sigma'\|_{2/5}\,\mathrm{polylog}(d/\varepsilon)\log(m/\delta)/\varepsilon^2\right)$ copies suffice to test whether $\rho = \sigma$ or whether Scenario 3 holds, with probability $1 - O(\delta)$.*

Proof. If $\mathrm{Tr}(\sigma[j,j]) < \varepsilon/(10m^2)$, then to test whether $\rho = \sigma$ or Scenario 3 holds, it suffices to decide whether $\mathrm{Tr}(\rho[j,j]) \geq \mathrm{Tr}(\sigma[j,j]) + \varepsilon/(10m^2)$. We can do this by measuring ρ using the POVM consisting of the projection Π_j to the principal submatrix indexed by S_j, together with $\mathrm{Id} - \Pi_j$, for which $O(m^4 \log^2(1/\delta)/\varepsilon^2)$ copies suffice to determine this with probability $1 - O(\delta)$.

Suppose now that $\mathrm{Tr}(\sigma[j,j]) \geq \varepsilon/(10m^2)$. We can use $O(\log^4(d/\varepsilon) \cdot \log(1/\delta)/\varepsilon^2)$ copies to approximate $\mathrm{Tr}(\rho[j,j])$ to additive error $\varepsilon/(40m^2)$ with probability $1 - O(\delta)$ using the same POVM.

If our estimate for $\mathrm{Tr}(\rho[j,j])$ is greater than $\varepsilon/(40m^2)$ away from $\mathrm{Tr}(\sigma[j,j])$, then $\rho \neq \sigma$.

Otherwise, $|\mathrm{Tr}(\rho[j,j]) - \mathrm{Tr}(\sigma[j,j])| \leq \varepsilon/(20m^2)$. Then by Fact 11.6.7, to determine whether we are in Scenario 3, it suffices to design a tester to distinguish whether the mixed states $\widehat{\rho}[j,j]$ and $\widehat{\sigma}[j,j]$ are equal or ε'-far in trace distance for

$$\varepsilon' \triangleq \frac{\varepsilon}{20m^2\,\mathrm{Tr}(\sigma[j,j])} = \Theta\left(\frac{\varepsilon}{20m^2 d_j 2^{-j}}\right). \tag{11.27}$$

Note that if $\widehat{\rho}[j,j]$ and $\widehat{\sigma}[j,j]$ are ε'-far in trace distance, they are at least $\varepsilon'/\sqrt{d_j}$-far in Frobenius. We conclude from Lemma 11.6.2 that we can distinguish with probability

$1 - O(\delta)$ between whether $\widehat{\rho}[j,j]$ and $\widehat{\sigma}[j,j]$ are equal or ε'-far in trace distance using $O(d_j^{3/2}\log(1/\delta)/\varepsilon'^2) = O(d_j^{7/2}2^{-2j}\log^4(d/\varepsilon)\log(1/\delta)/\varepsilon^2)$ measurements on the conditional state $\widehat{\rho}[j,j]$. Note that $\mathrm{Tr}(\rho[j,j]) \geq \Omega(\mathrm{Tr}(\sigma[j,j]))$ because $\mathrm{Tr}(\sigma[j,j]) \geq \varepsilon/(10m^2)$ by assumption, so $\mathrm{Tr}(\sigma[j,j]) \geq \Omega(d_j 2^{-j})$. As a result, this tester can make the desired number of measurements on the conditional state by using $O(d_j^{5/2}2^{-j}\log^4(d/\varepsilon)\log(1/\delta)/\varepsilon^2)$ copies of ρ and rejection sampling.

By a union bound over distinct pairs j, j', it therefore takes $O(\log(m/\delta))$ times

$$\sum_{j\in\mathcal{J}} O\left(d_j^{5/2}2^{-j}\log^4(d/\varepsilon)/\varepsilon^2\right) \leq \sum_{j\in\mathcal{J}} O\left(d_j^{5/2}\lambda_j\log^4(d/\varepsilon)/\varepsilon^2\right) \leq O\left(\|\sigma'\|_{2/5}\,\mathrm{polylog}(d/\varepsilon)/\varepsilon^2)\right),$$

copies to test whether Scenario 3 holds, where the last step above follows by Fact 11.3.1. □

Lemma 11.6.11. *If Scenario 3 does not hold, then* $O\left(\sqrt{d-d'}\|\sigma'\|_{1/2}\log(m/\delta)\,\mathrm{polylog}(d/\varepsilon)/\varepsilon^2\right)$ *copies suffice to test whether* $\rho = \sigma$ *or whether Scenario 4 holds, with probability* $1 - O(\delta)$.

Proof. Fix any $j \neq j' \in \mathcal{J}$ and suppose without loss of generality that $d_j \geq d_{j'}$. Let ρ^* and σ^* denote the matrices obtained by zeroing out all entries of ρ and σ except those in the principal submatrix indexed by $S_j \cup S_{j'}$. Let $\widehat{\rho}^*_{j,j'}$ and $\widehat{\sigma}^*_{j,j'}$ denote these same matrices with trace normalized to 1. For brevity, we will freely omit subscripts.

If $\mathrm{Tr}(\sigma^*) < \varepsilon/(5m^2)$, then $\|\sigma[j,j']\|_1 \leq \varepsilon/(10m^2)$ by the second part of Fact 11.3.5. If Scenario 2 holds, then $\|\rho[j,j']\|_1 \geq \varepsilon/(5m^2)$, so by another application of the second part of Fact 11.3.5, we would get that $\mathrm{Tr}(\rho^*) \geq 2\varepsilon/(5m^2)$, contradicting the fact that Scenario 1 does not hold.

Suppose now that $\mathrm{Tr}(\sigma^*) \geq \varepsilon/(5m^2)$. As in the proof of Lemma 11.6.10, we can use $O(\log^4(d/\varepsilon)\cdot\log(1/\delta)/\varepsilon^2)$ copies to approximate $\mathrm{Tr}(\rho^*)$ to within additive error $\varepsilon/(20m^2)$ with probability $1 - O(\delta)$.

If our estimate is greater than $\varepsilon/(20m^2)$ away from $\mathrm{Tr}(\sigma[j,j])$ then we know that $\rho \neq \sigma$.

Otherwise, $|\mathrm{Tr}(\rho^*) - \mathrm{Tr}(\sigma^*)| \leq \varepsilon/(10m^2)$, and in particular $\mathrm{Tr}(\rho^*) \geq \Omega(\mathrm{Tr}(\sigma^*))$ as a result. If Scenario 3 holds but Scenario 4 does not, then $\|\rho^* - \sigma^*\| \geq \varepsilon/(5m^2)$. So by Fact 11.6.7, to determine whether we are in Scenario 2, it suffices to design a tester to distinguish whether the mixed states $\widehat{\rho}^*$ and $\widehat{\sigma}^*$ are equal or ε''-far in trace distance, where

$$\varepsilon'' \triangleq \frac{\varepsilon}{10m^2 \operatorname{Tr}(\sigma^*)} = \Theta\left(\frac{\varepsilon}{10m^2} \cdot (d_j 2^{-j} + d_{j'} 2^{-j'})^{-1}\right) \tag{11.28}$$

Note that if ρ^* and σ^* are ε''-far in trace distance, they are at least $\varepsilon''/\sqrt{d_j}$-far in Frobenius, by the assumption that $d_j \geq d_{j'}$. We conclude from Lemma 11.6.2 that we can distinguish these two cases using

$$O(\sqrt{d_j} d_{j'} \log(1/\delta)/\varepsilon'^2) = O\left(\sqrt{d_j} d_{j'} (d_j 2^{-j} + d_{j'} 2^{-j'})^2 \log^4(d/\varepsilon) \log(1/\delta)/\varepsilon^2\right)$$

measurements on the conditional state $\widehat{\rho}^*$. Because $\operatorname{Tr}(\rho^*) \geq \Omega(\operatorname{Tr}(\sigma^*)) \geq \Omega(d_j 2^{-j} + d_{j'} 2^{-j'})$, this tester can make the desired number of measurements on the conditional state by using $O\left(\sqrt{d_j} d_{j'} (d_j 2^{-j} + d_{j'} 2^{-j'}) \log^4(d/\varepsilon) \log(1/\delta)/\varepsilon^2\right)$ copies of ρ and rejection sampling.

Summing over $j \neq j' \in \mathcal{J}$ for which $d_j \geq d_{j'}$, we conclude that it takes $O(\log(1/\delta))$ times

$$\begin{aligned}\sum_{j \neq j' \in \mathcal{J}: d_j \geq d_{j'}} \sqrt{d_j} d_{j'} (d_j 2^{-j} + d_{j'} 2^{-j'}) &\leq \sum_{j,j' \in \mathcal{J}: d_j \geq d_{j'}} d_j^{3/2} d_{j'} 2^{-j} + \sum_{j,j' \in \mathcal{J}: d_j \geq d_{j'}} \sqrt{d_j} d_{j'}^2 2^{-j'} \\ &\leq |\mathcal{J}| \cdot \sum_{j \in \mathcal{J}} d_j^{5/2} 2^{-j} + \left(\sum_{j \in \mathcal{J}} \sqrt{d_j}\right) \left(\sum_{j \in \mathcal{J}} d_j^2 2^{-j}\right) \\ &\leq \operatorname{polylog}(d/\varepsilon) \cdot \left(\|\sigma'\|_{2/5} + \sqrt{d - d'} \cdot \|\sigma'\|_{1/2}\right),\end{aligned}$$

copies to test whether Scenario 4 holds, where the last step above uses Fact 11.3.1.

We claim that the above bound is dominated by $O(\log(m/\delta) \operatorname{polylog}(d/\varepsilon))\sqrt{d - d'}\|\sigma'\|_{1/2}$. Indeed, note that for any vector $v \in \mathbb{R}^m$,

$$\|v\|_{2/5}^{2/5} = \sum_i v_i^{2/5} \leq \left(\sum_i (v_i^{2/5})^{5/4}\right)^{4/5} \cdot \left(\sum_i 1^5\right)^{1/5} \leq \|v\|_{1/2}^{2/5} \cdot \sqrt{m}^{2/5},$$

as desired. □

Altogether, Lemmas 11.6.8 to 11.6.11 allow us to conclude correctness of the algorithm Certify whose pseudocode is provided in Algorithm 48 below. The copy complexity guarantee follows from these lemmas together with Fact 11.5.27. □

11.7 Appendix: Adaptive Lower Bound

In this section we prove a lower bound against state certification algorithms that use adaptive, unentangled measurements.

Theorem 11.7.1. *There is an absolute constant $c > 0$ for which the following holds for any $0 < \varepsilon < c$.*[6] *Let $\sigma \in \mathbb{C}^{d\times d}$ be a diagonal density matrix. There is a matrix σ^* given by zeroing out the largest entry of σ and at most $O(\varepsilon \log(d/\varepsilon))$ additional mass from σ (see Definition 11.7.2 below), such that the following holds:*

Any algorithm for state certification to error ε with respect to σ using adaptive, unentangled measurements has copy complexity at least

$$\Omega\left(d \cdot d_{\mathsf{eff}}^{1/3} \cdot F(\widehat{\sigma}^*, \rho_{\mathsf{mm}})/(\varepsilon^2 \log(d/\varepsilon))\right).$$

The outline follows that of Section 11.5. In Section 11.7.1, we describe the procedure by which we remove mass from σ, which will be more aggressive than the one used for our nonadaptive lower bound. As a result, it will suffice to analyze the lower bound instance given Section 11.5.3, which we do in Section 11.7.2. For our analysis, we need to check some additional conditions hold for the adaptive lower bound framework of Section 11.4.3 to apply.

11.7.1 Bucketing and Mass Removal

Define $\{S_j\}, \mathcal{J}, S_{\mathsf{sing}}, S_{\mathsf{many}}$ in the same way as in Section 11.5.1. The way in which we remove mass from σ will be more aggressive than in the nonadaptive setting. We will end up removing up to $O(\varepsilon \log(d/\varepsilon))$ mass (see Fact 11.7.3) as follows:

Definition 11.7.2 (Removing low-probability elements- adaptive lower bound)**.** *Without loss of generality, suppose that $\lambda_1, \dots, \lambda_d$ are sorted in ascending order according to λ_i. Let $d' \le d$ denote the largest index for which $\sum_{i=1}^{d'} \lambda'_i \le 4\varepsilon$. Let $S_{\mathsf{tail}} \triangleq [d']$.*

Let σ^* denote the matrix given by zeroing out the largest entry of σ and the entries indexed by S_{tail}. It will be convenient to define $\mathcal{J}^*$ to be the buckets for the nonzero entries of σ^*,

[6] As presented, our analysis yields c within the vicinity of 1/3, but we made no attempt to optimize for this constant.

```
Algorithm 48: CERTIFY($\rho, \sigma, \varepsilon, \delta$)
  Input: Copy access to $\rho$, diagonal density matrix $\sigma$, error $\varepsilon$, failure probability $\delta$
  Output: YES if $\rho = \sigma$, NO if $\|\rho - \sigma\|_F > \varepsilon$, with probability $1 - \delta$.
  $m \leftarrow \log(10d/\varepsilon^2)$.
  Let $\Pi$ be the projector to the principal submatrix indexed by $S_{\mathsf{tail}}$.   // Scenario 1
  $\mathcal{M} \leftarrow \{\Pi, \mathrm{Id} - \Pi\}$.
  Measure $O(\log(1/\delta)/\varepsilon^2)$ copies of $\rho$ with the POVM $\mathcal{M}$.
  if $\geq (\varepsilon^2/5)$ fraction of outcomes observed correspond to $\Pi$ then
      return NO.
  for $j \in \mathcal{J}$ do                                               // Scenario 3
      Let $\Pi_j$ denote the projection to the principal submatrix indexed by $S_j$.
      $\mathcal{M}_j \leftarrow \{\Pi_j, \mathrm{Id} - \Pi_j\}$.
      Measure $O(\mathrm{polylog}(d/\varepsilon)\log(1/\delta)/\varepsilon^2)$ copies of $\rho$ with the POVM $\mathcal{M}_j$.
      if $\geq (Tr(\sigma[j,j]) + \varepsilon/(40m^2))$ fraction of outcomes observed correspond to $\Pi_j$
       then
          return NO.
      else
          Define $\varepsilon'$ according to (11.27).
          $b_j \leftarrow$ BASICCERTIFY($\widehat{\rho}[j,j], \widehat{\sigma}[j,j], \varepsilon', O(\delta/m)$).
          if $b_j =$ NO then
              return NO.
  for $j, j' \in \mathcal{J}$ distinct and satisfying $d_j \geq d_{j'}$ do        // Scenario 4
      Let $\Pi_{j,j'}$ denote the projection to the principal submatrix indexed by $S_j \cup S_{j'}$.
      $\mathcal{M}_{j,j'} \leftarrow \{\Pi_{j,j'}, \mathrm{Id} - \Pi_{j,j'}\}$.
      Measure $O(\mathrm{polylog}(d/\varepsilon)\log(1/\delta)/\varepsilon^2)$ copies of $\rho$ with the POVM $\mathcal{M}_{j,j'}$.
      if $\geq (Tr(\sigma^*_{j,j'}) + \varepsilon/(20m^2))$ fraction of outcomes observed correspond to $\Pi_{j,j'}$ then
          return NO.
      else
          Define $\varepsilon''$ according to (11.28).
          $b_{j,j'} \leftarrow$ BASICCERTIFY($\widehat{\rho}^*_{j,j'}, \widehat{\sigma}^*_{j,j}, \varepsilon'', O(\delta/m^2)$).
          if $b_{j,j'} =$ NO then
              return NO.
  return YES.
```

i.e. the set of $j \in \mathcal{J}$ for which S_j has nonempty intersection with $[d]\backslash S_{\mathsf{tail}}$.

Fact 11.7.3. *There are at most $O(\log(d/\varepsilon))$ indices $j \in \mathcal{J}^*$. As a consequence, $Tr(\sigma^*) \geq 1 - O(\varepsilon \log(d/\varepsilon))$.*

Proof. For any $i_1 \notin S_{\mathsf{tail}}$ and $i_2 \in S_{\mathsf{tail}}$, we have that $p_{i_1} > p_{i_2}$. In particular, summing over $i_2 \in S_{\mathsf{tail}}$, we conclude that $p_{i_1} \cdot |S_{\mathsf{tail}}| > 4\varepsilon$, so $p_{i_1} > 4\varepsilon/d$. By construction of the buckets S_j, the first part of the claim follows. As in the proof of Fact 11.5.3, the second part of the claim follows by definition of S_{light}. □

11.7.2 Analyzing Lower Bound II

We will analyze the sub-problem defined in Section 11.5.3 and prove the following lower bound:

Lemma 11.7.4. *Fix any $j, j' \in \mathcal{J}^*$ satisfying $d_j \geq d_{j'}$. If $d_j > 1$, then we can optionally take $j = j'$. Suppose $\varepsilon \leq d_{j'} \cdot 2^{-j/2-j'/2-1}$. Distinguishing between whether $\rho = \sigma$ or $\rho = \sigma_{\mathbf{W}}$ for $\mathbf{W} \in \mathbb{C}^{d_j \times d_{j'}}$ consisting of Haar-random orthonormal columns (see (11.14) and (11.15)), using adaptive unentangled measurements, has copy complexity at least*

$$\Omega\left(\frac{d_j^{1/3} \cdot d_{j'}^2 \cdot 2^{-j'}}{\varepsilon^2}\right).$$

Proof. As in Section 11.5.3, we will focus on the case where $j \neq j'$, but at the cost of some factors of two, the following arguments easily extend to the construction for $j = j'$ when $d_j > 1$ by replacing S_j and $S_{j'}$ with S_j^1, S_j^2 defined immediately before (11.15).

We have already verified in Section 11.5.3 that Conditions 1, 2, and (3) hold for $L, \varsigma = O\left(\frac{\varepsilon}{d_{j'} 2^{-j'/2}}\right)$.

For Condition 4, recall (11.17). As the diagonal entries of ρ indexed by S_j (resp. $S_{j'}$) are at least 2^{-j-1} (resp. $2^{-j'-1}$),

$$v_z^\dagger \rho v_z \geq 2^{-j-1}\|v_z^j\|^2 + 2^{-j'-1}\|v_z^{j'}\|^2 \geq 2^{-j/2-j'/2}\|v_z^j\|\|v_z^{j'}\|,$$

so

$$g^{\mathbf{U}}_{\mathcal{P}}(z) \le \frac{\varepsilon}{d_{j'}} \cdot \frac{\|v_z^j\|\|v_z^{j'}\|}{2^{-j/2-j'/2}\|v_z^j\|\|v_z^{j'}\|} \le \frac{\varepsilon}{d_{j'}2^{-j/2-j'/2}}.$$

In particular, as long as $\varepsilon \le d_{j'}2^{-j/2-j'/2-1}$, Condition 4 holds.

We can now apply Theorem 11.4.10 with $\tau = O\left(\frac{\varepsilon^2}{d_j^{1/3}d_{j'}^2 2^{-j'}}\right)$, noting that

$$\exp\left(-\Omega\left(\left\{\frac{d_j\tau^2}{L^2\varsigma^2}\wedge\frac{d\tau}{L^2}\right\}\right)\right) = \exp\left(-\Omega\left(d_j^{1/3}\right)\right),$$

to get that for any adaptive unentangled POVM schedule $\mathcal{S}$, if $p_0^{\le N}$ is the distribution over outcomes from measuring N copies of σ with $\mathcal{S}$ and $p_1^{\le N}$ is the distribution from measuring N copies of $\sigma_{\mathbf{U}}$, then

$$\mathrm{KL}\left(p_1^{\le N}\|p_0^{\le N}\right) \le \frac{N\varepsilon^2}{d_j^{1/3}d_{j'}^2 2^{-j'}} + O(N)\cdot\exp\left(-\Omega\left(d_j^{1/3} - \frac{N\varepsilon^2}{d_{j'}^2 2^{-j'}}\right)\right).$$

In particular, if $N = o\left(\frac{d_j^{1/3}d_{j'}^2 2^{-j'}}{\varepsilon^2\log(d/\varepsilon)}\right)$, then $\mathrm{KL}\left(p_1^{\le N}\|p_0^{\le N}\right) = o(1)$ and we get the desired lower bound. □

11.7.3 Putting Everything Together

Proof of Theorem 11.7.1. As in the proof of Theorem 11.5.1, we proceed by casework depending on whether $d_j = 1$ for all $j \in \mathcal{J}^*$.

Case 1. $d_j = 1$ *for all* $j \in \mathcal{J}^*$.

The analysis for this case in the nonadaptive setting completely carries over to this setting, because the lower bound from Lemma 11.5.24 holds even against adaptive POVM schedules. There are two possibilities. If there is a single bucket $j = j(i)$ for which $i \notin S_{\mathsf{tail}}$, then $d_{\mathsf{eff}} = 1$ and $\|\sigma^*\|_{1/2} = O(1)$; for ε smaller than some absolute constant, we have that $\sigma_{i,i} \ge 3/4$ and Lemma 11.5.24 gives an $\Omega(1/\varepsilon^2)$ lower bound as desired. Otherwise, let j' be the smallest index for which $j' = j(i')$ for some $i' \in \mathcal{J}^*$, and let $j > j'$ be the next smallest index for which $j = j(i)$ for some $i \in \mathcal{J}^*$. Consider the lower bound instance in Section 11.7.2 applied to this choice of j, j'. Provided that $\varepsilon \le 2^{-j/2-j'/2-1}$, we would obtain a copy complexity lower bound of $\Omega(2^{-j'}/\varepsilon^2) \geq \Omega(\|\sigma^*\|_{1/2}/(\varepsilon^2 \log(d/\varepsilon)))$, where the inequality is by Fact 11.3.1, and we would be done. On the other hand, if $\varepsilon \geq 2^{-j/2-j'/2-1}$, then because $2^{-j'} > 2^{-j}$, we would conclude that $2^{-j} \leq 2\varepsilon$. In particular, this implies that $\sum_{j'' \in \mathcal{J}^*, i \in S_{j''}: j'' \neq j'} \lambda_i \leq 4\varepsilon$, contradicting the fact that we have removed all buckets of total mass at most 4ε in defining S_{tail}.

Case 2. *$d_j > 1$ for some $j \in \mathcal{J}^*$.*

Let $j_* \triangleq \arg\max_{j \in \mathcal{J}^*} d_j$ and $j'_* \triangleq \arg\max_{j \in \mathcal{J}^*} d_j^2 2^{-j}$. By Lemma 11.5.19, as long as ε satisfies the bound

$$\varepsilon \leq d_{j'_*} \cdot 2^{-j_*/2-j'_*/2-1}, \tag{11.29}$$

we have a lower bound of

$$\Omega\left(d_{j_*}^{1/3} \cdot d_{j'_*}^2 \cdot 2^{-j'_*}/\varepsilon^2\right) \geq \Omega\left(d \cdot d_{\mathsf{eff}}^{1/3} \cdot F(\sigma^*, \rho_{\mathsf{mm}})/(\varepsilon^2 \log(d/\varepsilon))\right),$$

where the second step follows by Fact 11.3.1 and Fact 11.5.27. Note that because $d_{j_*} > 1$ as we are in Case 2, we do not constrain j_*, j'_* to be distinct necessarily.

But under our assumptions on j, j' and on $\mathcal{J}^*$, (11.29) must hold:

$$d_{j'} 2^{-j/2-j'/2-1} \geq d_j 2^{-j-1} \geq \varepsilon$$

where the first step follows by the assumption that $j' \triangleq \arg\max_{j \in \mathcal{J}^*} d_j^2 2^{-j}$, and the second by the assumption that every bucket indexed by $\mathcal{J}^*$ has total mass at least 4ε. □

11.8 Appendix: Deferred Proofs

11.8.1 Proof of Theorem 11.4.10

We will need the following helper lemmas:

Lemma 11.8.1 (Implicit in Lemma 10.6.2). *Conditions 2 and 4 together imply that for any transcript $z_{\leq t}$, $\Delta(z_{\leq t}) \geq \exp(-\varsigma^2 t)$.*

Proof. By convexity of the exponential function and the fact that $1 + g^{\mathbf{U}}_{z_{<t}}(z_t) > 0$ for all

$\mathbf{U}, t, z_t$,

$$\Delta(z_{<t}) \geq \prod_{i=1}^{t-1} \exp\left(\mathop{\mathbb{E}}_{\mathbf{U}\sim\mathcal{D}}[\ln(1 + g^{\mathbf{U}}_{z_{<i}}(z_i))]\right).$$

For any $i < t$ we have that

$$\begin{aligned}\exp\left(\mathop{\mathbb{E}}_{\mathbf{U}\sim\mathcal{D}}[\ln(1 + g^{\mathbf{U}}_{z_{<i}}(z_i))]\right) &\geq \exp\left(\mathop{\mathbb{E}}_{\mathbf{U}}[g^{\mathbf{U}}_{z_{<i}}(z_i) - g^{\mathbf{U}}_{z_{<i}}(z_i)^2]\right) \\ &\geq \exp\left(-\varsigma^2\right),\end{aligned}$$

where the first step follows by the elementary inequality $\ln(x) \geq x - x^2$ for all $x \in [-1/2, 1/2]$, and the second step follows by Conditions 1 and 2. □

Lemma 11.8.2. *Conditions 1, 2, 3, and 4 together imply that* $\mathbb{E}_{z_{<t},\mathbf{U},\mathbf{V}}\left[(\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t}})^2\right] \leq \exp(O(t\varsigma^2))$.

To prove this, it will be convenient to define the following for any ℓ-copy sub-problem corresponding to POVM $\mathcal{M}$

$$K^{\mathbf{U},\mathbf{V}}_{\mathcal{P}} \triangleq \mathop{\mathbb{E}}_{z\sim p_0(\mathcal{M})}\left[\left(g^{\mathbf{U}}_{\mathcal{P}}(z) + g^{\mathbf{V}}_{\mathcal{P}}(z)\right)^2\right]$$

and first show the following:

Lemma 11.8.3. *Under the hypothesis of Lemma 11.8.2,* $\mathbb{E}_{\mathbf{U},\mathbf{V}}\left[\left(1 + \gamma K^{\mathbf{U},\mathbf{V}}_{\mathcal{P}}\right)^t\right] \leq \exp(O(\gamma t\varsigma^2))$ *for any absolute constant* $\gamma > 0$ *and any* $t = o(d/L^2)$.

Proof. By the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have that $K^{\mathbf{U},\mathbf{V}}_{\mathcal{P}} \leq G(\mathbf{U})^2 + G(\mathbf{V})^2$. By Lemma 11.4.5, we immediately get that $\Pr_{\mathbf{U}}\left[K^{\mathbf{U},\mathbf{V}}_{\mathcal{P}} > (\mathbb{E}[G(\mathbf{U})] + s)^2\right] \leq \exp(-ds^2/L^2)$. Applying the inequality again allows us to lower bound the left-hand side by $\Pr_{\mathbf{U}}\left[K^{\mathbf{U},\mathbf{V}}_{\mathcal{P}} > 2\,\mathbb{E}[G(\mathbf{U})]^2 + 2s^2\right]$ so we conclude that

$$\Pr_{\mathbf{U}}\left[K^{\mathbf{U},\mathbf{V}}_{\mathcal{P}} > 2\,\mathbb{E}[G(\mathbf{U})]^2 + s\right] \leq \exp(-ds/2L^2).$$

We can apply Fact 1.3.30 to the random variable $Z \triangleq K^{\mathbf{U},\mathbf{V}}_{\mathcal{P}}$ and the function $f(Z) \triangleq (1+\gamma Z)^t$ to conclude that

$$\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}}\left[\left(1 + \gamma\cdot K^{\mathbf{U},\mathbf{V}}_{\mathcal{P}}\right)^t\right] \leq 2(1 + 2\gamma\,\mathbb{E}[G(\mathbf{U})]^2)^t + \int_0^\infty \gamma t(1+\gamma x)^{t-1}\cdot e^{-x\cdot d/2L^2}\,\mathrm{d}x$$

$$\begin{aligned}
&\le 2(1+2\gamma\,\mathbb{E}[G(\mathbf{U})]^2)^t + \gamma t \int_0^\infty e^{-x(d/2L^2-\gamma(t-1))}\,\mathrm{d}x \\
&\le 2(1+2\gamma\,\mathbb{E}[G(\mathbf{U})]^2)^t + \frac{\gamma t}{d/2L^2-\gamma(t-1)} \le \exp(O(t\gamma\,\mathbb{E}[G(\mathbf{U})]^2)),
\end{aligned}$$

where in the last two steps we used that $t = o(d/L^2)$ to ensure that the integral is bounded and that the second term in the final expression is negligible. □

We can now prove Lemma 11.8.2:

Proof of Lemma 11.8.2. By the second part of Condition 4, we know that for any constant $c \ge 2$,

$$\mathop{\mathbb{E}}_{z\sim\Omega(\mathcal{M}^{z_{<t-1}})}\left[g^{\mathbf{U}}_{z_{<t-1}}(z)^a \cdot g^{\mathbf{V}}_{z_{<t-1}}(z_t)^b\right] \le \frac{1}{4}\mathop{\mathbb{E}}_{z}[g^{\mathbf{U}}_{z_{<t-1}}(z)^2],$$

so we conclude that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{z\sim p_0(\mathcal{M}^{z_{<t-1}})}\left[(1+g^{\mathbf{U}}_{z_{<t-1}}(z))^c(1+g^{\mathbf{V}}_{z_{<t-1}}(z))^c\right] \\
&\le 1 + O_c\left(\mathop{\mathbb{E}}_{z}[g^{\mathbf{U}}_{z_{<t-1}}(z)^2]\right) + O_c\left(\mathop{\mathbb{E}}_{z}[g^{\mathbf{V}}_{z_{<t-1}}(z)^2]\right) + O_c\left(\phi^{\mathbf{U},\mathbf{V}}_{z_{<t-1}}\right) \le 1 + C(c)\cdot K^{\mathbf{U},\mathbf{V}}_{z_{<t-1}} \qquad (11.30)
\end{aligned}$$

for some absolute constant $C(c) > 0$, where the last step follows by AM-GM. For $\alpha_i \triangleq 2\cdot\left(\frac{t-1}{t-2}\right)^i$, we have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{z_{<t},\mathbf{U},\mathbf{V}}\left[\left(\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t}}\right)^{\alpha_i}\right] \\
&\quad\le \mathop{\mathbb{E}}_{z_{<t-1},\mathbf{U},\mathbf{V}}\left[\left(\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t-1}}\right)^{\alpha_i}\cdot\left(1+C(\alpha_i)\cdot K^{\mathbf{U},\mathbf{V}}_{z_{<t-1}}\right)\right] \qquad (11.31) \\
&\quad\le \mathop{\mathbb{E}}_{z_{<t-1},\mathbf{U},\mathbf{V}}\left[\left(\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t-1}}\right)^{\alpha_i(t-1)/(t-2)}\right]^{(t-2)/(t-1)}\cdot\mathop{\mathbb{E}}_{z_{<t-1},\mathbf{U},\mathbf{V}}\left[\left(1+C(\alpha_i)\cdot K^{\mathbf{U},\mathbf{V}}_{z_{<t-1}}\right)^{t-1}\right]^{1/(t-1)} \qquad (11.32) \\
&\quad\le \mathop{\mathbb{E}}_{z_{<t-1},\mathbf{U},\mathbf{V}}\left[\left(\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t-1}}\right)^{\alpha_{i+1}(t-1)/(t-2)}\right]\cdot\mathop{\mathbb{E}}_{z_{<t-1},\mathbf{U},\mathbf{V}}\left[\left(1+C(\alpha_i)\cdot K^{\mathbf{U},\mathbf{V}}_{z_{<t-1}}\right)^{t-1}\right]^{1/(t-1)}.
\end{aligned}$$

where (11.31) follows by (11.30), and (11.32) follows by Holder's. Unrolling this recurrence, we conclude that

$$\mathop{\mathbb{E}}_{z_{<t},\mathbf{U},\mathbf{V}}\left[\left(\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t}}\right)^2\right] \le \prod_{i=1}^{t-1}\mathop{\mathbb{E}}_{z_{<i},\mathbf{U},\mathbf{V}}\left[\left(1+C(\alpha_{t-1-i})\cdot K^{\mathbf{U},\mathbf{V}}_{z_{<i}}\right)^{t-1}\right]^{1/(t-1)}$$

$$
\begin{aligned}
&\le \prod_{i=1}^{t-1} \mathop{\mathbb{E}}_{z_{<i},\mathbf{U},\mathbf{V}}\left[\left(1 + C(2e)\cdot K^{\mathbf{U},\mathbf{V}}_{z_{<i}}\right)^{t-1}\right]^{1/(t-1)}, \qquad (11.33)\\
&\le \sup_{\mathcal{M}} \mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}}\left[\left(1 + O(K^{\mathbf{U},\mathbf{V}}_{\mathcal{M}})\right)^{t-1}\right]
\end{aligned}
$$

where (11.33) follows by the fact that for $1 \le i \le t-1$, $\alpha_{t-1-i} \le 2\left(1+\frac{1}{t-2}\right)^{t-2} \le 2e$, and the supremum in the last step is over all POVMs $\mathcal{M}$. The lemma then follows from Lemma 11.8.3. □

We now have all the ingredients to complete the proof of Theorem 11.4.10.

Proof of Theorem 11.4.10. Given transcript $z_{<t}$ and $\mathbf{U},\mathbf{V} \sim \mathcal{D}$, let $\mathbb{1}\left[\mathcal{E}^{\mathbf{U},\mathbf{V}}_{z_{<t}}(\tau)\right]$ denote the indicator of whether $\left|\phi^{\mathbf{U},\mathbf{V}}_{z_{<t}}\right| > \tau$; note that by Lemma 11.4.6, this event happens with probability at most $\xi(\tau)$, where

$$
\xi(s) \triangleq \exp\left(-\Omega\left(\frac{ds^2}{L^2\varsigma^2} \wedge \frac{ds}{L^2}\right)\right).
$$

We have that

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}}\left[\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t}} \cdot \phi^{\mathbf{U},\mathbf{V}}_{z_{<t}}\right] &= \mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}}\left[\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t}} \cdot \phi^{\mathbf{U},\mathbf{V}}_{z_{<t}} \cdot \left(\mathbb{1}[\mathcal{E}^{\mathbf{U},\mathbf{V}}_{z_{<t}}(\tau)] + \mathbb{1}[\mathcal{E}^{\mathbf{U},\mathbf{V}}_{z_{<t}}(\tau)^c]\right)\right]\\
&\le \frac{1}{4}\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}}\left[\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t}} \cdot \mathbb{1}[\mathcal{E}^{\mathbf{U},\mathbf{V}}_{z_{<t}}(\tau)]\right] + \tau \cdot \mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}}\left[\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t}} \cdot \mathbb{1}[\mathcal{E}^{\mathbf{U},\mathbf{V}}_{z_{<t}}(\tau)^c]\right]\\
&\le \frac{1}{4} \cdot \underbrace{\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}}\left[\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t}} \cdot \mathbb{1}[\mathcal{E}^{\mathbf{U},\mathbf{V}}_{z_{<t}}(\tau)]\right]}_{\text{Ⓑ}_{z_{<t}}} + \tau \cdot \underbrace{\mathop{\mathbb{E}}_{\mathbf{U},\mathbf{V}}\left[\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t}}\right]}_{\text{Ⓖ}_{z_{<t}}},
\end{aligned}
$$

where in the second step we used Condition 4 to conclude that $\phi^{\mathbf{U},\mathbf{V}}_{z_{<t}} \le 1/4$. Note that for any transcript $z_{<t}$, $\Delta(z_{<t})^2 = \mathbb{E}_{\mathbf{U},\mathbf{V}}[\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t}}] = \text{Ⓖ}_{z_{<t}}$, so by this and the fact that the likelihood ratio between two distributions always integrates to 1,

$$
\mathop{\mathbb{E}}_{z_{<t}\sim p_0^{\le t-1}}\left[\frac{1}{\Delta^{(t-1)}(z_{<t})} \cdot \text{Ⓖ}_{z_{<t}}\right] = \mathop{\mathbb{E}}_{z_{<t}\sim p_0^{\le t-1}}[\Delta^{(t-1)}(z_{<t})] = 1. \qquad (11.34)
$$

Recalling the definition of Z_t in Lemma 10.6.1, we conclude that

$$\begin{aligned} Z_t &\le \frac{1}{4} \mathop{\mathbb{E}}_{z_{<t}\sim p_0^{\le t-1}}\left[\frac{1}{\Delta^{(t-1)}(z_{<t})}\cdot \textcircled{B}_{z_{<t}}\right] + \tau \cdot \mathop{\mathbb{E}}_{z_{<t}\sim p_0^{\le t-1}}\left[\frac{1}{\Delta^{(t-1)}(z_{<t})}\cdot \textcircled{G}_{z_{<t}}\right] \\ &\le \frac{1}{4}\exp(t\varsigma^2) \mathop{\mathbb{E}}_{z_{<t}\sim p_0^{\le t-1}}\left[\textcircled{B}_{z_{<t}}\right] + \tau, \end{aligned}$$

where the second step follows by Lemma 11.8.1 and (11.34).

To upper bound $\mathbb{E}_{z_{<t}\sim p_0^{\le t-1}}[\textcircled{B}_{z_{<t}}]$, apply Cauchy-Schwarz to get

$$\begin{aligned} \mathop{\mathbb{E}}_{z_{<t}\sim p_0^{\le t-1}}\left[\textcircled{B}_{z_{<t}}\right] &\le \mathop{\mathbb{E}}_{z_{<t}\sim p_0^{\le t-1},\mathbf{U},\mathbf{V}}\left[\left(\Psi^{\mathbf{U},\mathbf{V}}_{z_{<t}}\right)^2\right]^{1/2} \cdot \mathop{\Pr}_{z_{<t}\sim p_0^{\le t-1},\mathbf{U},\mathbf{V}}\left[\mathcal{E}^{\mathbf{U},\mathbf{V}}_{z_{<t}}(\tau)\right]^{1/2} \\ &\le \exp(O(t\varsigma^2))\cdot \xi(\tau), \end{aligned}$$

where the second step follows by Lemma 11.4.6 and Lemma 11.8.2. Invoking Lemma 10.6.1 concludes the proof. □

11.8.2 Proof of Fact 11.5.16

Proof. We may assume $s < m+n$ (otherwise obviously $b = n$). Assume to the contrary that $\sum_{i=1}^{b+1} v_i d_i \le \varepsilon$. We proceed by casework based on whether $w_{s'+1} = u_{a+1}$ or $w_{s'+1} = v_{b+1}$.

If $w_{s'+1} = u_{a+1}$, then

$$3\varepsilon < \sum_{i=1}^{s+1} w_i d_i^* = \sum_{i=1}^{a+1} u_i + \sum_{i=1}^{b} v_i d_i \le \sum_{i=1}^{a+1} v_{b+1}\cdot 2^{1-i} + \sum_{i=1}^{b} v_i \le 2\varepsilon + \sum_{i=1}^{b} v_i d_i,$$

where in the first step we used maximality of s, in the third step we used that $u_{a+1} \le v_{b+1}$ and that $u_{i+1} \ge 2u_i$ for all i, and in the last step we used that $v_{b+1} \le \sum_{i=1}^{b+1} v_i d_i \le \varepsilon$. From this we conclude that $\sum_{i=1}^{b} v_i d_i > \varepsilon$, a contradiction.

If $w_{s'+1} = v_{b+1}$, the argument is nearly identical. We have

$$3\varepsilon < \sum_{i=1}^{s+1} w_i d_i^* = \sum_{i=1}^{a} u_i + \sum_{i=1}^{b+1} v_i d_i \le \sum_{i=1}^{a} v_{b+1}\cdot 2^{1-i} + \sum_{i=1}^{b+1} v_i d_i \le 2\varepsilon + \sum_{i=1}^{b+1} v_i,$$

where in the first step we again used maximality of s, in the third step we used that $u_a \le v_{b+1}$

and $u_{i+1} \geq 2u_i$ for all i, and in the last step we used that $v_{b+1} \leq \sum_{i=1}^{b+1} v_i d_i \leq \varepsilon$. From this we conclude that $\sum_{i=1}^{b} v_i d_i > \varepsilon$, a contradiction. □

Bibliography

[Abb73] Ernst Abbe. Beiträge zur theorie des mikroskops und der mikroskopischen wahrnehmung. *Archiv für mikroskopische Anatomie*, 9(1):413–418, 1873. 28, 610, 663

[ABM19] Jason Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best arm identification for contaminated bandits. *J. Mach. Learn. Res.*, 20(91):1–39, 2019. 342

[AC16] Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 116–120, 2016. 342

[ACBFS02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002. 67, 68, 341, 682, 691

[ACK14] Jayadev Acharya, Clément L Canonne, and Gautam Kamath. A chasm between identity and equivalence testing with conditional queries. *arXiv preprint arXiv:1411.7346*, 2014. 685

[ADH$^+$15] J. Acharya, I. Diakonikolas, C. Hegde, J. Li, and L. Schmidt. Fast and Near-Optimal Algorithms for Approximating Distributions by Histograms. In *PODS*, 2015. 229

[ADJ$^+$11] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Competitive closeness testing. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 47–68. JMLR Workshop and Conference Proceedings, 2011. 70, 718

[ADJ$^+$12] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Suresh. Competitive classification and closeness testing. In *Conference on Learning Theory*, pages 22–1. JMLR Workshop and Conference Proceedings, 2012. 70, 718

[ADLS16] Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Fast algorithms for segmented regression. In *International Conference on Machine Learning*, pages 2878–2886, 2016. 181

[ADLS17] Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1278–1289. SIAM, 2017. 46, 226, 227, 229, 235, 252, 253, 255, 291, 499, 503, 508, 516

[AGJ14] Anima Anandkumar, Rong Ge, and Majid Janzamin. Analyzing tensor power method dynamics: Applications to learning overcomplete latent variable models. *arXiv preprint arXiv:1411.1488*, 2014. 98

[AGJ15] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Learning overcomplete latent variable models through tensor methods. In *Conference on Learning Theory*, pages 36–112, 2015. 98

[AGKE15] Leandro Aolita, Christian Gogolin, Martin Kliesch, and Jens Eisert. Reliable quantum certification of photonic state preparations. *Nature communications*, 6(1):1–8, 2015. 29, 684

[AGKS21] Pranjal Awasthi, Sreenivas Gollapudi, Kostas Kollias, and Apaar Sadhwani. Online learning under adversarial corruptions, 2021. 342

[AGZ10] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*, volume 118. Cambridge university press, 2010. 93, 706, 722

[AH97] Carmen O Acuna and Joseph Horowitz. A statistical approach to the resolution of point sources. *Journal of Applied Statistics*, 24(4):421–436, 1997. 655

[Air35] George Biddell Airy. On the diffraction of an object-glass with circular aperture. *Transactions of the Cambridge Philosophical Society*, 5:283, 1835. 28, 609, 658

[AK01] S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd Symposium on Theory of Computing*, pages 247–257, 2001. 51, 229, 617

[AK08] Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008. 428, 455, 456

[AL99] N. Abe and Philip M. Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, 1999. 400

[ALPV12] Albert Ai, Alex Lapanowski, Yaniv Plan, and Roman Vershynin. One-bit compressed sensing with non-gaussian measurements. *arXiv preprint arXiv:1208.6279*, 2012. 102

[AM91] William Aiello and Milena Mihail. Learning the fourier spectrum of probabilistic lists and trees. In *Proceedings of the Second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '91, page 291?299, USA, 1991. Society for Industrial and Applied Mathematics. 56, 58, 412

[AM05] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the Eighteenth Annual Conference on Learning Theory (COLT)*, pages 458–469, 2005. 51, 229, 617

[AMS09] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009. 101, 109

[AN72] EA Ash and G Nicholls. Super-resolution aperture scanning microscope. *Nature*, 237(5357):510, 1972. 658

[AN04] Noga Alon and Assaf Naor. Approximating the cut-norm via grothendieck's inequality. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 72–80. ACM, 2004. 279, 282

[Ans60] Frank J Anscombe. Rejection of outliers. *Technometrics*, 2(2):123–146, 1960. 229

[ANSV08] Koenraad MR Audenaert, Michael Nussbaum, Arleta Szkoła, and Frank Verstraete. Asymptotic error rates in quantum hypothesis testing. *Communications in Mathematical Physics*, 279(1):251–283, 2008. 684

[AOK09] Vladimir Al Osipov and Eugene Kanzieper. Statistics of thermal to shot noise crossover in chaotic cavities. *Journal of Physics A: Mathematical and Theoretical*, 42(47):475101, 2009. 683

[App12] Benny Applebaum. Pseudorandom generators with long stretch and low locality from random local one-way functions. In *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pages 805–816, New York, NY, USA, 2012. Association for Computing Machinery. 182

[APVZ14] Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning sparse polynomial functions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 500–510. SIAM, 2014. 95

[Aus19] Tim Austin. The structure of low-complexity gibbs measures on product spaces. *The Annals of Probability*, 47(6):4002–4023, 2019. 51

[Aus20] Tim Austin. Multi-variate correlation and mixtures of product measures. *Kybernetika*, 56(3):459–499, 2020. 51

[AW01] Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001. 326, 343

[Axe81] Daniel Axelrod. Cell-substrate contacts illuminated by total internal reflection fluorescence. *The Journal of Cell Biology*, 89(1):141–145, 1981. 658

[AZGL$^+$18] Zeyuan Allen-Zhu, Ankit Garg, Yuanzhi Li, Rafael Oliveira, and Avi Wigderson. Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 172–181, 2018. 101

[AZL16] Zeyuan Allen-Zhu and Yuanzhi Li. Lazysvd: Even faster svd decomposition yet without agonizing pain. In *Advances in Neural Information Processing Systems*, pages 974–982, 2016. 75

[AZLL19] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pages 6158–6169, 2019. 34, 178, 179

[Bak11] Laurent Bako. Identification of switched linear systems via sparse optimization. *Automatica*, 47(4):668–677, 2011. 502

[BB96] PW Brouwer and CWJ Beenakker. Diagrammatic method of integration over the unitary group, with applications to quantum transport in mesoscopic systems. *Journal of Mathematical Physics*, 37(10):4904–4934, 1996. 683

[BB00] Ya M Blanter and Markus Büttiker. Shot noise in mesoscopic conductors. *Physics reports*, 336(1-2):1–166, 2000. 683

[BB16] Aleksandrs Belovs and Eric Blais. A polynomial lower bound for testing monotonicity. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1021–1032, 2016. 685

[BB$^+$18] Dmitry Babichev, Francis Bach, et al. Slice inverse regression with score functions. *Electronic Journal of Statistics*, 12(1):1507–1543, 2018. 33, 102, 104, 181

[BBBB72] Richard E Barlow, David J Bartholomew, James M Bremner, and H Daniel Brunk. Statistical inference under order restrictions: The theory and application of isotonic regression. Technical report, Wiley New York, 1972. 228

[BBM$^+$05] Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. 373

[BC09] Stephen M Barnett and Sarah Croke. Quantum state discrimination. *Advances in Optics and Photonics*, 1(2):238–278, 2009. 684

[BC18] Rishiraj Bhattacharyya and Sourav Chakraborty. Property testing of joint distributions using conditional samples. *ACM Transactions on Computation Theory (TOCT)*, 10(4):1–20, 2018. 685

[BCB12] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012. 68, 682, 691

[BCBL13] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013. 342

[BCE06] Radu Balan, Pete Casazza, and Dan Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006. 53, 497

[BCG19] Eric Blais, Clément L Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. *ACM Transactions on Computation Theory (TOCT)*, 11(2):1–37, 2019. 718

[BCP$^+$17] Roksana Baleshzar, Deeparnab Chakrabarty, Ramesh Krishnan S Pallavoor, Sofya Raskhodnikova, and C Seshadhri. Optimal unateness testers for real-valued functions: Adaptivity helps. *arXiv preprint arXiv:1703.05199*, 2017. 685

[BDJ$^+$20] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M. Kane, Pravesh K. Kothari, and Santosh S. Vempala. Robustly learning mixtures of k arbitrary gaussians. *CoRR*, abs/2012.02119, 2020. 39, 57

[BDLS17] Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212, 2017. 229

[Bee97] Carlo WJ Beenakker. Random-matrix theory of quantum transport. *Reviews of modern physics*, 69(3):731, 1997. 683

[BEH07] Stefan Bretschneider, Christian Eggeling, and Stefan W Hell. Breaking the diffraction barrier in fluorescence microscopy by optical shelving. *Physical Review Letters*, 98(21):218103, 2007. 658

[Bel18] Aleksandrs Belovs. Adaptive lower bound for testing monotonicity on the line. *arXiv preprint arXiv:1801.08709*, 2018. 685

[BEMGS17] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 118–128, 2017. 26

[Ben03] V Bentkus. An inequality for tail probabilities of martingales with differences bounded from one side. *Journal of Theoretical Probability*, 16(1):161–173, 2003. 84, 85, 152

[BEZ$^+$97] LCEO Brand, C Eggeling, C Zander, KH Drexhage, and CAM Seidel. Single-molecule identification of coumarin-120 by time-resolved fluorescence detection: Comparison of one-and two-photon excitation in solution. *The Journal of Physical Chemistry A*, 101(24):4313–4321, 1997. 658

[BFJ$^+$94] Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 253–262. ACM, 1994. 413, 444, 495

[BFKV98] Avrim Blum, Alan Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1):35–52, 1998. 25

[BG17] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 605–614, 2017. 34, 178

[BG18] Sébastien Bubeck and Shirshendu Ganguly. Entropic clt and phase transition in high-dimensional wishart matrices. *International Mathematics Research Notices*, 2018(2):588–606, 2018. 68

[Bir40] Garrett Birkhoff. *Lattice theory*, volume 25. American Mathematical Soc., 1940. 196

[Bir87a] L. Birgé. Estimating a density under order restrictions: Nonasymptotic minimax risk. *Annals of Statistics*, 15(3):995–1012, 1987. 228

[Bir87b] L. Birgé. On the risk of histograms for estimating decreasing densities. *Annals of Statistics*, 15(3):1013–1022, 1987. 228

[BJK78] Gilbert Bassett Jr and Roger Koenker. Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, 73(363):618–622, 1978. 340

[BJK15] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015. 41, 341

[BJKK17] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems*, pages 2110–2119, 2017. 41, 341

[BJW18] Ainesh Bakshi, Rajesh Jayaram, and David P Woodruff. Learning two layer rectified neural networks in polynomial time. *arXiv preprint arXiv:1811.01885*, 2018. 34, 36, 102, 178, 179

[BK94] AL Blum and Ravi Kannan. Learning an intersection of k halfspaces over a uniform distribution. In *Theoretical Advances in Neural Computation and Learning*, pages 337–356. Springer, 1994. 181

[BK20] Ainesh Bakshi and Pravesh Kothari. Outlier-robust clustering of non-spherical mixtures. *arXiv preprint arXiv:2005.02970*, 2020. 39, 325

[BKS14] Boaz Barak, Jonathan A Kelner, and David Steurer. Rounding sum-of-squares relaxations. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 31–40, 2014. 90

[Blu92] Avrim Blum. Rank-r decision trees are a subclass of r-decision lists. *Information Processing Letters*, 42(4):183–185, 1992. 408, 412

[BNR10] Laura Balzano, Robert Nowak, and Benjamin Recht. Online identification and tracking of subspaces from highly incomplete information. In *2010 48th Annual allerton conference on communication, control, and computing (Allerton)*, pages 704–711. IEEE, 2010. 101, 102

[Bos57] Roger Joseph Boscovich. De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impessa. *Bononiensi Scientiarum et Artum Instuto Atque Academia Commentarii*, 4:353–396, 1757. 41, 332

[Bou14] Nicolas Boumal. *Optimization and estimation on manifolds.* PhD thesis, Catholic University of Louvain, Louvain-la-Neuve, Belgium, 2014. 101

[BOW19] Costin Bădescu, Ryan O'Donnell, and John Wright. Quantum state certification. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 503–514, 2019. 66, 680, 684

[BP20] Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. *arXiv preprint arXiv:2007.01394*, 2020. 41, 47, 325, 339, 340

[BPS$^+$06] Eric Betzig, George H Patterson, Rachid Sougrat, O Wolf Lindwasser, Scott Olenych, Juan S Bonifacino, Michael W Davidson, Jennifer Lippincott-Schwartz, and Harald F Hess. Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313(5793):1642–1645, 2006. 658

[BR89] Avrim Blum and Ronald L Rivest. Training a 3-node neural network is np-complete. In *Advances in neural information processing systems*, pages 494–501, 1989. 25, 180

[BR19] Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*, 2019. 347

[Bri12] David R Brillinger. A generalized linear model with “gaussian” regressor variables. In *Selected Works of David Brillinger*, pages 589–606. Springer, 2012. 33, 36, 101, 180

[BRST21] Joan Bruna, Oded Regev, Min Jae Song, and Yi Tang. Continuous lwe. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 694–707, 2021. 60

[Bru55] Hugh D Brunk. Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, pages 607–616, 1955. 226

[Bru58] H. D. Brunk. On the estimation of parameters restricted by inequalities. *The Annals of Mathematical Statistics*, 29(2):pp. 437–454, 1958. 229

[BRW09] F. Balabdaoui, K. Rufibach, and J. A. Wellner. Limit distribution theory for maximum likelihood estimation of a log-concave density. *The Annals of Statistics*, 37(3):pp. 1299–1331, 2009. 229

[BS12] Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1, 2012. 342

[BS14] Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *arXiv preprint arXiv:1404.5236*, 2014. 87

[BS15] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. *SIAM Journal on Computing*, 44(4):889–911, 2015. 51, 617

[BT03] Brendan O Bradley and Murad S Taqqu. Financial risk and heavy tails. In *Handbook of heavy tailed distributions in finance*, pages 35–103. Elsevier, 2003. 27

[Bub14] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014. 355, 389

[Bux37] A Buxton. Xli. note on optical resolution. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 23(154):440–442, 1937. 610, 665

[BV08] S Charles Brubaker and Santosh S Vempala. Isotropic pca and affine-invariant clustering. In *Building Bridges*, pages 241–281. Springer, 2008. 51, 617

[BVDDD$^+$99] E Bettens, D Van Dyck, AJ Den Dekker, J Sijbers, and A Van den Bos. Model-based two-object resolution from observations having counting statistics. *Ultramicroscopy*, 77(1-2):37–48, 1999. 655

[BW07] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: Limit distribution theory and the spline connection. *The Annals of Statistics*, 35(6):pp. 2536–2564, 2007. 229

[BW10] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds. *Statistica Neerlandica*, 64(1):45–70, 2010. 229

[BW13] Max Born and Emil Wolf. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*. Elsevier, 2013. 666

[BWY17] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017. 53, 497, 501

[Byl94] Tom Bylander. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 340–347, 1994. 25

[Can20] Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020. 64, 685

[CAT$^+$20] Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I Jordan, Nicolas Flammarion, and Peter L Bartlett. Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08137*, 2020. 41, 47, 325, 329, 339

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006. 325, 343

[CCLM17] Emanuel Carneiro, Vorrapan Chandee, Friedrich Littmann, and Micah B Milinovich. Hilbert spaces and the pair correlation of zeros of the riemann zeta-function. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 2017(725):143–182, 2017. 63, 615, 641, 644

[CDGR18] Clément L Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. *Theory of Computing Systems*, 62(1):4–62, 2018. 683

[CDGS20] Yu Cheng, Ilias Diakonikolas, Rong Ge, and Mahdi Soltanolkotabi. High-dimensional robust mean estimation via gradient descent. In *International Conference on Machine Learning*, pages 1768–1778. PMLR, 2020. 45

[CDKS18] Yu Cheng, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Robust learning of fixed-structure bayesian networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, 2018. 39

[CDS10] Giulio Chiribella, Giacomo Mauro D'Ariano, and Dirk Schlingemann. Barycentric decomposition of quantum measurements in finite dimensions. *Journal of mathematical physics*, 51(2):022111, 2010. 91

[CDSS13] Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1380–1394. Society for Industrial and Applied Mathematics, 2013. 226, 227, 229

[CDSS14a] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *NIPS*, pages 1844–1852, 2014. 229

[CDSS14b] Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 2014. 226, 227, 229, 499, 503, 508

[CDVV14] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM, 2014. 683

[CEHV15] Aldo Conca, Dan Edidin, Milena Hering, and Cynthia Vinzant. An algebraic characterization of injectivity in phase retrieval. *Applied and Computational Harmonic Analysis*, 38(2):346–356, 2015. 33, 96, 98, 102

[CFG13] Emmanuel J Candès and Carlos Fernandez-Granda. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013. 617

[CFG14] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, 67(6):906–956, 2014. 617

[CFGM16] Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. *SIAM Journal on Computing*, 45(4):1261–1296, 2016. 685

[CGB$^+$11] Chao Chen, Kay Grennan, Judith Badner, Dandan Zhang, Elliot Gershon, Li Jin, and Chunyu Liu. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one*, 6(2):e17238, 2011. 28

[Che00] Anthony Chefles. Quantum state discrimination. *Contemporary Physics*, 41(6):401–424, 2000. 684

[Chi20] Geoffrey Chinot. Erm and rerm are optimal estimators for regression problems when malicious outliers corrupt the labels, 2020. 41, 340, 343

[CHRZ07] Kamalika Chaudhuri, Eran Halperin, Satish Rao, and Shuheng Zhou. A rigorous analysis of population stratification with limited data. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1046–1055. Society for Industrial and Applied Mathematics, 2007. 28, 51

[CK20] Michael Chmielewski and Sarah C Kucker. An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473, 2020. 26

[CKMY20] Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Classification under misspecification: Halfspaces, generalized linear models, and connections to evolvability. *arXiv preprint arXiv:2006.04787*, 2020. 341

[CKPS16] Xue Chen, Daniel M Kane, Eric Price, and Zhao Song. Fourier-sparse interpolation without a frequency gap. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 741–750. IEEE, 2016. 503

[CL13] Arun Tejasvi Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning*, pages 1040–1048, 2013. 53, 497, 498, 501

[CLL$^+$17] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 26

[CLS15] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015. 33, 36, 96, 98, 102, 106, 107

[CM20] Sitan Chen and Raghu Meka. Learning polynomials of few relevant dimensions. In *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, pages 1161–1227. PMLR, 2020. 219

[Col03] Benoît Collins. Moments and cumulants of polynomial random variables on unitarygroups, the itzykson-zuber integral, and free probability. *International Mathematics Research Notices*, 2003(17):953–982, 2003. 707

[CR14] Richard Cole and Tim Roughgarden. The sample complexity of revenue maximization. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 243–252. ACM, 2014. 229

[CRS14] Clément Canonne, Dana Ron, and Rocco A Servedio. Testing equivalence between distributions using conditional samples. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1174–1192. SIAM, 2014. 685

[CRS15] Clément L Canonne, Dana Ron, and Rocco A Servedio. Testing probability distributions using conditional samples. *SIAM Journal on Computing*, 44(3):540–616, 2015. 685

[CSV13] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013. 33, 53, 96, 98, 102, 497, 503

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60. ACM, 2017. 227, 229, 291, 325, 502

[CT04] K.S. Chan and H. Tong. Testing for multimodality with dependent data. *Biometrika*, 91(1):113–123, 2004. 229

[CW20] Jordan Cotler and Frank Wilczek. Quantum overlapping tomography. *Physical Review Letters*, 124(10):100401, 2020. 680

[CWO16] Jerry Chao, E Sally Ward, and Raimund J Ober. Fisher information theory for parameter estimation in single molecule microscopy: tutorial. *JOSA A*, 33(7):B36–B57, 2016. 655, 656

[CWX17a] Xi Chen, Erik Waingarten, and Jinyu Xie. Beyond talagrand functions: new lower bounds for testing monotonicity and unateness. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 523–536, 2017. 685

[CWX17b] Xi Chen, Erik Waingarten, and Jinyu Xie. Boolean unateness testing with $\widetilde{O}(n^{3/4})$ adaptive queries. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 868–879. IEEE, 2017. 685

[CYC13] Yudong Chen, Xinyang Yi, and Constantine Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. *arXiv preprint arXiv:1312.7006*, 2013. 53, 497, 501

[Dan16] Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 105–117, 2016. 25

[Dan17] Amit Daniely. Sgd learns the conjugate kernel class of the network. *CoRR*, abs/1702.08503, 2017. 178

[Das99] Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science*, pages 634–644. IEEE, 1999. 51, 229, 617

[Daw67] William Rutter Dawes. *Catalogue of micrometrical measurements of double stars.* Royal Astronomical Society, 1867. 666

[DD96] Arnold J Den Dekker. Model-based optical resolution. In *Quality Measurement: The Indispensable Bridge between Theory and Reality (No Measurements? No Science!) Joint Conference-1996: IEEE Instrumentation and*

Measurement Technology Conference and IMEKO Tec, volume 1, pages 441–446. IEEE, 1996. 655

[DDKT16] Constantinos Daskalakis, Anindya De, Gautam Kamath, and Christos Tzamos. A size-free clt for poisson multinomials and its applications. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1074–1086. ACM, 2016. 229

[DDO$^+$13] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R.A. Servedio, and L. Tan. Learning Sums of Independent Integer Random Variables. In *FOCS*, pages 217–226, 2013. 229

[DDS12a] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In *SODA*, pages 1371–1385, 2012. 229

[DDS12b] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In *STOC*, pages 709–728, 2012. 229

[DDS14] Anindya De, Ilias Diakonikolas, and Rocco A Servedio. Learning from satisfying assignments. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 478–497. SIAM, 2014. 412

[DDVdB97] Arnold Jan Den Dekker and A Van den Bos. Resolution: a survey. *JOSA A*, 14(3):547–557, 1997. 665, 668, 670

[Den98] François Denis. Pac learning from positive statistical queries. In *International Conference on Algorithmic Learning Theory*, pages 112–126. Springer, 1998. 412

[DF52] G Toraldo Di Francia. Super-gain antennas and optical resolving power. *Il Nuovo Cimento (1943-1954)*, 9:426–438, 1952. 658

[DF55] G Toraldo Di Francia. Resolving power and information. *Josa*, 45(7):497–501, 1955. 611, 669, 672

[DGK$^+$20] Ilias Diakonikolas, Surbhi Goel, Sushrut Karmalkar, Adam R Klivans, and Mahdi Soltanolkotabi. Approximation schemes for relu regression. In *Conference on Learning Theory*, 2020. 34, 178, 179

[DGT19] Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent pac learning of halfspaces with massart noise. In *Advances in Neural Information Processing Systems*, pages 4749–4760, 2019. 341

[DGZ12] Vacha Dave, Saikat Guha, and Yin Zhang. Measuring and fingerprinting click-spam in ad networks. In *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*, pages 175–186, 2012. 27

[DH18] Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In *Conference On Learning Theory*, pages 1887–1930, 2018. 33, 97, 101, 102, 180

[DHKK20] Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, and Sushrut Karmalkar. Robustly learning any clusterable mixture of gaussians. *arXiv preprint arXiv:2005.06417*, 2020. 39, 325

[DHL19] Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In *Advances in Neural Information Processing Systems*, pages 6065–6075, 2019. 279

[Dia16] Ilias Diakonikolas. Learning structured distributions. *Handbook of Big Data*, 267, 2016. 229, 235, 503

[DJ98] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921, 1998. 229

[DJKP95] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia. *Journal of the Royal Statistical Society, Ser. B*, pages 371–394, 1995. 229

[DJKP96] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996. 229

[DJS08] Arnak S Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9(Aug):1647–1678, 2008. 33, 102

[DK16] Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016. 718, 724, 755

[DK19] Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019. 339

[DK20] Ilias Diakonikolas and Daniel M. Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 184–195, 2020. 34, 36, 61, 176, 178

[DKK$^+$17] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 999–1008. JMLR. org, 2017. 101, 227, 229, 279, 325

[DKK$^+$18] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. SIAM, 2018. 345

[DKK$^+$19a] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019. 39, 43, 45, 101, 227, 229, 231, 279, 325

[DKK$^+$19b] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606, 2019. 101, 279, 325, 339

[DKK$^+$20] Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. A polynomial time algorithm for learning halfspaces with tsybakov noise. *arXiv preprint arXiv:2010.01705*, 2020. 341

[DKKZ20] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In *Conference on Learning Theory*, pages 1514–1539, 2020. 22, 36, 37, 39, 178, 179, 180, 181

[DKN14] Ilias Diakonikolas, Daniel M Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 1841–1854. SIAM, 2014. 683

[DKS16a] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. The fourier transform of poisson multinomial distributions and its algorithmic applications. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1060–1073. ACM, 2016. 229, 503

[DKS16b] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Optimal learning via the fourier transform for sums of independent integer random variables. In *Conference on Learning Theory*, pages 831–849, 2016. 229, 503

[DKS16c] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Properly learning poisson binomial distributions in almost polynomial time. In *Conference on Learning Theory*, pages 850–878, 2016. 229, 503

[DKS17] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017. 57, 60, 413, 426, 445, 677

[DKS18a] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1061–1073, 2018. 102

[DKS18b] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060, 2018. 51, 57, 227, 231, 617

[DKS19] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019. 339

[DKSS21] Ilias Diakonikolas, Daniel M Kane, Alistair Stewart, and Yuxin Sun. Outlier-robust learning of ising models under dobrushin's condition. *arXiv preprint arXiv:2102.02171*, 2021. 39

[DKTZ20] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning halfspaces with massart noise under structured distributions. *arXiv preprint arXiv:2002.05632*, 2020. 341

[DL01] Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer Science & Business Media, 2001. 228, 252, 453, 466

[DLT18] Simon S Du, Jason D Lee, and Yuandong Tian. When is a convolutional filter easy to learn? In *6th International Conference on Learning Representations, ICLR 2018*, 2018. 34, 178

[DMN19] Anindya De, Elchanan Mossel, and Joe Neeman. Is your function low dimensional? In *Conference on Learning Theory*, pages 979–993, 2019. 102, 180

[DMN20] Anindya De, Elchanan Mossel, and Joe Neeman. Robust testing of low-dimensional functions. *arXiv preprint arXiv:2004.11642*, 2020. 180

[dNS20] Tommaso d'Orsi, Gleb Novikov, and David Steurer. Regress consistently when oblivious outliers overwhelm, 2020. 340

[Don92] David L Donoho. Superresolution via sparsity constraints. *SIAM journal on mathematical analysis*, 23(5):1309–1331, 1992. 62, 617

[DR09] L. Dumbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68, 2009. 229

[DS79] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28, 1979. 51

[DS89] David L Donoho and Philip B Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, 49(3):906–931, 1989. 617

[DS00] S. Dasgupta and L. Schulman. A two-round variant of EM for Gaussian mixtures. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 143–151, 2000. 51, 229, 617

[dSLCP11] Marcus P da Silva, Olivier Landon-Cardinal, and David Poulin. Practical characterization of quantum devices without tomography. *Physical Review Letters*, 107(21):210404, 2011. 29, 684

[DT19] Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using l1-penalized huber's m-estimator. In *Advances in Neural Information Processing Systems*, pages 13188–13198, 2019. 41, 340

[Dur19] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019. 80

[DV89] Richard D De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, 1989. 501

[DV20] Amit Daniely and Gal Vardi. Hardness of learning neural networks with natural weights. *arXiv preprint arXiv:2006.03177*, 2020. 180

[DV21] Amit Daniely and Gal Vardi. From local pseudorandom generators to hardness of learning. *arXiv preprint arXiv:2101.08303*, 2021. 182

[DWSD15] Justin Demmerle, Eva Wegel, Lothar Schermelleh, and Ian M Dobbie. Assessing resolution in super-resolution imaging. *Methods*, 88:3–10, 2015. 658, 665, 669, 671

[DZM$^+$14] Hendrik Deschout, Francesca Cella Zanacchi, Michael Mlodzianoski, Alberto Diaspro, Joerg Bewersdorf, Samuel T Hess, and Kevin Braeckmans. Precisely and accurately localizing single emitters in fluorescence microscopy. *Nature Methods*, 11(3):253, 2014. 656

[EAS98] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. 101, 127

[EG18] Ronen Eldan and Renan Gross. Decomposition of mean-field gibbs distributions into product measures. *Electronic Journal of Probability*, 23:1–24, 2018. 51

[EH89] Andrzej Ehrenfeucht and David Haussler. Learning decision trees from random examples. *Information and Computation*, 82(3):231–246, 1989. 408, 412

[EV13] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013. 502, 503

[Fal67] Oscar Falconi. Limits to which double lines, double stars, and disks can be resolved and measured. *JOSA*, 57(8):987–993, 1967. 612, 656

[Far66] Edward J Farrell. Information content of photoelectric star images. *JOSA*, 56(5):578–587, 1966. 655

[FB81] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 502

[FB12] Eric D Feigelson and G Jogesh Babu. *Modern statistical methods for astronomy: with R applications*. Cambridge University Press, 2012. 656

[FG13] Carlos Fernandez-Granda. Support detection in super-resolution. *arXiv preprint arXiv:1302.3921*, 2013. 617

[FG16] Carlos Fernandez-Granda. Super-resolution of point sources via convex programming. *Information and Inference: A Journal of the IMA*, 5(3):251–303, 2016. 617

[FGLE12] Steven T Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012. 684

[FGR$^+$17] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 64(2):8, 2017. 443, 444

[FL11] Steven T Flammia and Yi-Kai Liu. Direct fidelity estimation from few pauli measurements. *Physical review letters*, 106(23):230501, 2011. 684

[FLS11] Richard P Feynman, Robert B Leighton, and Matthew Sands. *The Feynman lectures on physics, Vol. I: The new millennium edition: mainly mechanics, radiation, and heat*, volume 1. Basic books, 2011. 610, 667

[FM99] Yoav Freund and Yishay Mansour. Estimating a mixture of two product distributions. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 53–62. ACM, 1999. 408

[FN71] D Kh Fuk and Sergey V Nagaev. Probability inequalities for sums of independent random variables. *Theory of Probability & Its Applications*, 16(4):643–660, 1971. 349

[FOS05] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *FOCS 2005*, pages 501–510, 2005. 59, 229, 408, 413, 415, 445, 454, 464

[Fou97] A.-L. Fougères. Estimation de densités unimodales. *Canadian Journal of Statistics*, 25:375–387, 1997. 229

[Fow89] Grant R Fowles. *Introduction to Modern Optics.* Courier Corporation, 1989. 658

[FR20] Dylan J Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. *arXiv preprint arXiv:2002.04926*, 2020. 331, 333, 342, 347, 348, 375, 395, 400, 401, 403

[FS10] Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010. 53, 497, 501

[GGI$^+$02] Anna C Gilbert, Sudipto Guha, Piotr Indyk, Shanmugavelayutham Muthukrishnan, and Martin Strauss. Near-optimal sparse fourier representations via sampling. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 152–161, 2002. 617

[GGJ$^+$20] Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. *arXiv preprint arXiv:2006.12011*, 2020. 22, 36, 37, 39, 179, 180, 181

[GHK15] Rong Ge, Qingqing Huang, and Sham M Kakade. Learning mixtures of gaussians in high dimensions. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 761–770, 2015. 51, 617

[GIIS14] Anna C Gilbert, Piotr Indyk, Mark Iwen, and Ludwig Schmidt. Recent developments in the sparse fourier transform: A compressed fourier transform for big data. *IEEE Signal Processing Magazine*, 31(5):91–100, 2014. 617

[GK19] Surbhi Goel and Adam R Klivans. Learning neural networks with two nonlinear layers in polynomial time. In *Conference on Learning Theory*, pages 1470–1499, 2019. 34, 102, 178

[GKK19] Surbhi Goel, Sushrut Karmalkar, and Adam Klivans. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In *Advances in Neural Information Processing Systems*, pages 8584–8593, 2019. 180

[GKKT17] Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the relu in polynomial time. In *Conference on Learning Theory*, pages 1004–1042. PMLR, 2017. 34, 102, 178, 179, 180

[GKLW18] Rong Ge, Rohith Kuditipudi, Zhize Li, and Xiang Wang. Learning two-layer neural networks with symmetric inputs. *arXiv preprint arXiv:1810.06793*, 2018. 34, 102, 178, 179

[GKM18] Surbhi Goel, Adam R. Klivans, and Raghu Meka. Learning one convolutional layer with overlapping patches. In Jennifer G. Dy and Andreas Krause 0001, editors, *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1778–1786. PMLR, 2018. 34, 178

[GKT19] Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578, 2019. 342

[GLM17] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017. 34, 102, 178, 179

[GLS81] M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, Jun 1981. 89

[GM15] Rong Ge and Tengyu Ma. Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2015)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2015. 98

[GMOV18] Weihao Gao, Ashok Vardhan Makkuva, Sewoong Oh, and Pramod Viswanath. Learning one-hidden-layer neural networks under general input distributions. *CoRR*, abs/1810.04133, 2018. 34, 178

[GMS05] Anna C Gilbert, Shan Muthukrishnan, and Martin Strauss. Improved time bounds for near-optimal sparse fourier representations. In *Wavelets XI*, volume 5914, page 59141A. International Society for Optics and Photonics, 2005. 617

[GNJN13] Sivakant Gopi, Praneeth Netrapalli, Prateek Jain, and Aditya Nori. One-bit compressed sensing: Provable support and vector recovery. In *International Conference on Machine Learning*, pages 154–162, 2013. 102

[Gol17] Oded Goldreich. *Introduction to property testing*. Cambridge University Press, 2017. 685

[Gon18] Felipe Gonçalves. A note on band-limited minorants of an euclidean ball. *Proceedings of the American Mathematical Society*, 146(5):2063–2068, 2018. 63, 615, 641, 644

[Goo05] Joseph W Goodman. *Introduction to Fourier Optics*. Roberts and Company Publishers, 2005. 658

[Goo15] Joseph W Goodman. *Statistical optics*. John Wiley &amp; Sons, 2015. 657, 658, 671

[Gre56] U. Grenander. On the theory of mortality measurement. *Skand. Aktuarietidskr.*, 39:125–153, 1956. 228

[Gro85] P. Groeneboom. Estimating a monotone density. In *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pages 539–555, 1985. 228

[GRS18] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018. 177

[GS99] Scott Gaffney and Padhraic Smyth. Trajectory clustering with mixtures of regression models. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 63–72. ACM, 1999. 53, 497

[GS19] Navin Goyal and Abhishek Shetty. Non-gaussian component analysis using entropy methods. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 840–851, 2019. 181

[Gus99] Mats GL Gustafsson. Extended resolution fluorescence microscopy. *Current opinion in structural biology*, 9(5):627–628, 1999. 658

[GW09] F. Gao and J. A. Wellner. On the rate of convergence of the maximum likelihood estimator of a k-monotone density. *Science in China Series A: Mathematics*, 52:1525–1538, 2009. 229

[Har64] James L Harris. Resolving power and decision theory. *JOSA*, 54(5):606–611, 1964. 655

[HATC$^+$19] Shivayogi V Hiremath, Amir Mohammad Amiri, Binod Thapa-Chhetry, Gretchen Snethen, Mary Schmidt-Read, Marlyn Ramos-Lamboy, Donna L Coffman, and Stephen S Intille. Mobile health-based physical activity intervention for individuals with spinal cord injury in the community: A pilot study. *PloS one*, 14(10):e0223762, 2019. 27

[Haz19] Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019. 393

[HBZ10] Bo Huang, Hazen Babcock, and Xiaowei Zhuang. Breaking the diffraction barrier: super-resolution imaging of cells. *Cell*, 143(7):1047–1058, 2010. 657, 673

[Hec15] Eugene Hecht. *Optics*. Pearson, 2015. 24, 29, 658, 660, 664, 667

[Hel64] C Helstrom. The detection and resolution of optical signals. *IEEE Transactions on Information Theory*, 10(4):275–287, 1964. 655, 662

[Hel69] Carl W Helstrom. Detection and resolution of incoherent objects by a background-limited optical system. *JOSA*, 59(2):164–175, 1969. 655

[Hel70] Carl W Helstrom. Resolvability of objects from the standpoint of statistical parameter estimation. *JOSA*, 60(5):659–666, 1970. 655

[Hel04] Stefan W Hell. Strategy for far-field optical imaging and writing without diffraction limit. *Physics Letters A*, 326(1-2):140–145, 2004. 658

[Hel07] Stefan W Hell. Far-field optical nanoscopy. *Science*, 316(5828):1153–1158, 2007. 657

[Hel09] Stefan W Hell. Microscopy and its focal switch. *Nature methods*, 6(1):24, 2009. 657

[HG09] Rainer Heintzmann and Mats GL Gustafsson. Subdiffraction resolution in continuous samples. *Nature Photonics*, 3(7):362, 2009. 657

[HGM06] Samuel T Hess, Thanu PK Girirajan, and Michael D Mason. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophysical Journal*, 91(11):4258–4272, 2006. 658

[HHJ$^+$17] Jeongwan Haah, Aram W Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Transactions on Information Theory*, 63(9):5628–5641, 2017. 66, 680, 682, 685, 694

[HHP$^+$16] Roarke Horstmeyer, Rainer Heintzmann, Gabriel Popescu, Laura Waller, and Changhuei Yang. Standardizing the resolution claims for coherent microscopy. *Nature Photonics*, 10(2):68, 2016. 658

[HIKP12] Haitham Hassanieh, Piotr Indyk, Dina Katabi, and Eric Price. Nearly optimal sparse fourier transform. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 563–578, 2012. 503, 617

[Hil54] Clifford Hildreth. Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, 49(267):598–619, 1954. 226

[HJP$^+$01] Marian Hristache, Anatoli Juditsky, Jörg Polzehl, Vladimir Spokoiny, et al. Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566, 2001. 33, 102

[HJS01] Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001. 33, 102

[HK95] Stefan W Hell and Matthias Kroug. Ground-state-depletion fluorscence microscopy: A concept for breaking the diffraction resolution limit. *Applied Physics B*, 60(5):495–497, 1995. 658

[HK13] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20, 2013. 51, 617

[HK15] Qingqing Huang and Sham M Kakade. Super-resolution off the grid. In *Advances in Neural Information Processing Systems*, pages 2665–2673, 2015. 63, 615, 618, 641, 642, 643, 673

[HKY17] Elad Hazan, Adam Klivans, and Yang Yuan. Hyperparameter optimization: A spectral approach. *arXiv preprint arXiv:1706.00764*, 2017. 412

[HKZ+12] Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012. 82

[HL18] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018. 51, 57, 89, 227, 228, 229, 231, 232, 258, 325, 382, 617

[HL19] Samuel B Hopkins and Jerry Li. How hard is robust mean estimation? In *Conference on Learning Theory*, pages 1649–1682, 2019. 345

[HLZ20] Samuel B. Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 45

[HMR18] Zhiyi Huang, Yishay Mansour, and Tim Roughgarden. Making the most of your samples. *SIAM Journal on Computing*, 47(3):651–674, 2018. 229

[Hop18] Samuel Hopkins. *Statistical inference and the sum of squares method*. PhD thesis, Cornell University, 2018. 57

[Hou27] William V Houston. A compound interferometer for fine structure work. *Physical Review*, 29(3):478, 1927. 610

[HP76] D. L. Hanson and G. Pledger. Consistency in concave regression. *The Annals of Statistics*, 4(6):pp. 1038–1050, 1976. 229

[HP15] Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two Gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 753–760. ACM, 2015. 51, 591, 616, 617, 677

[HPS98] Gordon B Hazen, James M Pellissier, and Jayavel Sounderpandian. Stochastic-tree models in medical decision making. *Interfaces*, 28(4):64–80, 1998. 411

[HPZ18] Nina Holden, Yuval Peres, and Alex Zhai. Gravitational allocation on the sphere. *Proceedings of the National Academy of Sciences*, 115(39):9666–9671, 2018. 512

[HS65] Richard F Hespos and Paul A Strassmann. Stochastic decision trees for the analysis of investment decisions. *Management Science*, 11(10):B–244, 1965. 411

[HS92] Stefan Hell and Ernst HK Stelzer. Properties of a 4pi confocal fluorescence microscope. *JOSA A*, 9(12):2159–2166, 1992. 658

[HS15] Reshad Hosseini and Suvrit Sra. Matrix manifold optimization for gaussian mixtures. In *Advances in Neural Information Processing Systems*, pages 910–918, 2015. 101

[HS16] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016. 336, 375

[HSLC94] Stefan W Hell, Ernst HK Stelzer, Steffen Lindek, and Christoph Cremer. Confocal microscopy with an increased detection aperture: type-b 4pi confocal microscopy. *Optics Letters*, 19(3):222–224, 1994. 658

[HSS15] Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pages 956–1006, 2015. 98

[HSSS16] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191, 2016. 98

[Hub64] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964. 41, 326, 340

[Hub73] Peter J Huber. Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics*, pages 799–821, 1973. 41, 340

[Hub92] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992. 229

[HV$^+$96] Jeffrey J Holt, Jeffrey D Vaaler, et al. The beurling-selberg extremal functions for a ball in euclidean space. *Duke Mathematical Journal*, 83(1):203–248, 1996. 63, 615, 641, 644

[HW94] Stefan W Hell and Jan Wichmann. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics Letters*, 19(11):780–782, 1994. 658

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 24

[IAVHDL11] Mariya Ishteva, P-A Absil, Sabine Van Huffel, and Lieven De Lathauwer. Best low multilinear rank approximation of higher-order tensors, based on the riemannian trust-region scheme. *SIAM Journal on Matrix Analysis and Applications*, 32(1):115–135, 2011. 101

[IJMT05] Nicole Immorlica, Kamal Jain, Mohammad Mahdian, and Kunal Talwar. Click fraud resistant methods for learning click-through rates. In *International Workshop on Internet and Network Economics*, pages 34–45. Springer, 2005. 27

[IK14] Piotr Indyk and Michael Kapralov. Sample-optimal Fourier sampling in any constant dimension. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 514–523. IEEE, 2014. 503

[IKP14] Piotr Indyk, Michael Kapralov, and Eric Price. (nearly) sample-optimal sparse fourier transform. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 480–499. SIAM, 2014. 617

[IS12] Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2012. 68, 69, 689, 720, 731

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018. 180

[JHW18] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the *l*_{1} distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018. 718

[JJ94] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994. 501

[JLR07] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007. 28

[JLS20] Yaonan Jin, Daogao Liu, and Zhao Song. A robust multi-dimensional sparse fourier transform in the continuous setting. *arXiv preprint arXiv:2005.06156*, 2020. 618

[JO19] Ayush Jain and Alon Orlitsky. Robust learning of discrete distributions from batches. *arXiv preprint arXiv:1911.08532*, 2019. 226, 279, 282, 283, 288, 289, 304, 307

[JO20] Ayush Jain and Alon Orlitsky. A general method for robust learning from batches. *arXiv preprint arXiv:2002.11099*, 2020. 284

[JO21] Ayush Jain and Alon Orlitsky. Robust density estimation from batches: The best things in life are (nearly) free. In *International Conference on Machine Learning*, pages 4698–4708. PMLR, 2021. 284

[JSA15] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv*, pages arXiv–1506, 2015. 34, 36, 102, 178, 179

[JSZB08] Na Ji, Hari Shroff, Haining Zhong, and Eric Betzig. Advances in the speed and resolution of light microscopy. *Current opinion in neurobiology*, 18(6):605–616, 2008. 657

[JW37] Francis A Jenkins and Harvey E White. *Fundamentals of optics*. Tata McGraw-Hill Education, 1937. 658

[JW09] H. K. Jankowski and J. A. Wellner. Estimation of a discrete monotone density. *Electronic Journal of Statistics*, 3:1567–1605, 2009. 228

[Kan20] Daniel M. Kane. Robust learning of mixtures of gaussians. *arXiv preprint arXiv:2007.05912*, 2020. 39, 57, 325

[Kap16] Michael Kapralov. Sparse fourier transform in any constant dimension with nearly-optimal sample complexity in sublinear time. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 264–277, 2016. 503, 617

[Kap17] Michael Kapralov. Sample efficient estimation and recovery in sparse FFT via isolation on average. In *Foundations of Computer Science, 2017. FOCS'17. IEEE 58th Annual IEEE Symposium on*. https://arxiv.org/pdf/1708.04544, 2017. 503

[KC19] Jeongyeol Kwon and Constantine Caramanis. EM converges for a mixture of many linear regressions. *arXiv preprint arXiv:1905.12106*, 2019. 53, 497, 500, 501, 534, 562

[KCB$^+$20] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D Waggoner, Ryan Jewell, and Nicholas JG Winter. The shape of and solutions to the mturk quality crisis. *Political Science Research and Methods*, 8(4):614–629, 2020. 26

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998. 413, 426, 443, 445, 677

[Kee10] Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer Science & Business Media, 2010. 329

[Ken08] Ian R Kenyon. *The light fantastic: a modern introduction to classical and quantum optics*. Oxford University Press, USA, 2008. 24, 29, 658, 664

[KH99] Thomas A Klar and Stefan W Hell. Subdiffraction resolution in far-field fluorescence microscopy. *Optics letters*, 24(14):954–956, 1999. 658

[KJH95] Janos Kirz, Chris Jacobsen, and Malcolm Howells. Soft x-ray microscopes and their biological applications. *Quarterly reviews of biophysics*, 28(1):33–130, 1995. 658

[KKK19] Sushrut Karmalkar, Pravesh Kothari, and Adam Klivans. List-decodable linear regression. In *NeurIPS*. arXiv preprint arXiv:1905.05679, 2019. 229, 498, 502

[KKM18] Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430, 2018. 41, 47, 229, 325, 329, 339, 340

[KL05] Vilmos Komornik and Paola Loreti. *Fourier series in control theory*. Springer Science & Business Media, 2005. 614

[KLT09] Adam R Klivans, Philip M Long, and Alex K Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 588–600. Springer, 2009. 102, 181

[KM10] R. Koenker and I. Mizera. Quasi-concave density estimation. *Ann. Statist.*, 38(5):2998–3027, 2010. 229

[KM15] Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015. 373

[KMR$^+$94] Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 273–282. ACM, 1994. 407

[KMV10] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562, 2010. 51, 62, 229, 617

[KMY$^+$16] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. 228

[KOS04] Adam R Klivans, Ryan O'Donnell, and Rocco A Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer and System Sciences*, 68(4):808–840, 2004. 95, 102

[KOS08] Adam R Klivans, Ryan O'Donnell, and Rocco A Servedio. Learning geometric concepts via gaussian surface area. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 541–550. IEEE, 2008. 95, 181

[KP92] G. Kerkyacharian and D. Picard. Density estimation in Besov spaces. *Statistics & Probability Letters*, 13(1):15–24, 1992. 229

[KP18] Sushrut Karmalkar and Eric Price. Compressed sensing with adversarial sparse noise via l1 regression. In *2nd Symposium on Simplicity in Algorithms (SOSA 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018. 340

[KPK19] Sayash Kapoor, Kumar Kshitij Patel, and Purushottam Kar. Corruption-tolerant bandit learning. *Machine Learning*, 108(4):687–715, 2019. 342, 345

[KPRvdO16] Stefan Kunis, Thomas Peter, Tim Römer, and Ulrich von der Ohe. A multivariate generalization of prony's method. *Linear Algebra and its Applications*, 490:31–47, 2016. 617

[KPT96] G. Kerkyacharian, D. Picard, and K. Tribouley. Lp adaptive density estimation. *Bernoulli*, 2(3):pp. 229–247, 1996. 229

[KQC$^+$18] Jeongyeol Kwon, Wei Qian, Constantine Caramanis, Yudong Chen, and Damek Davis. Global convergence of EM algorithm for mixtures of two component linear regression. *arXiv preprint arXiv:1810.05752*, 2018. 53, 497, 501

[KS91] Olav Kallenberg and Rafal Sztencel. Some dimension-free features of vector-valued martingales. *Probability Theory and Related Fields*, 88(2):215–247, 1991. 80, 403, 404

[KS04] Adam R Klivans and Rocco A Servedio. Learning dnf in time 2o (n1/3). *Journal of Computer and System Sciences*, 68(2):303–318, 2004. 95

[KS08] Subhash Khot and Rishi Saket. On hardness of learning intersection of two halfspaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 345–354, 2008. 102

[KS09] Adam R Klivans and Alexander A Sherstov. Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12, 2009. 180

[KS16] Subhash Khot and Igor Shinkar. An $\widetilde{o}(n)$ queries adaptive tester for unateness. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016. 685

[KS17] Pravesh K Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. *arXiv preprint arXiv:1711.11581*, 2017. 51, 617

[KSS09] BA Khoruzhenko, DV Savin, and H-J Sommers. Systematic approach to statistics of conductance and shot-noise in chaotic cavities. *Physical Review B*, 80(12):125301, 2009. 683

[KSS18] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046. ACM, 2018. 57, 227, 228, 229, 231, 325

[KSV14] Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. Low-rank tensor completion by riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014. 101

[KT19] Gautam Kamath and Christos Tzamos. Anaconda: A non-adaptive conditional sampling algorithm for distribution testing. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 679–693. SIAM, 2019. 685

[KWB19] Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*, 2019. 57

[KYB17] Jason M Klusowski, Dana Yang, and WD Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *arXiv preprint arXiv:1704.08231*, 2017. 53, 497, 501

[L^+17] Po-Ling Loh et al. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896, 2017. 41, 340

[Lan00] LJ Landau. Bessel functions: monotonicity and bounds. *Journal of the London Mathematical Society*, 61(1):197–215, 2000. 627

[Las01] Jean B. Lasserre. *New Positive Semidefinite Relaxations for Nonconvex Quadratic Programs*, pages 319–331. Springer US, Boston, MA, 2001. 88

[Lat97] Rafał Latała. Estimation of moments of sums of independent real random variables. *The Annals of Probability*, 25(3):1502–1513, 1997. 231, 240, 371

[Lau12] Marcel A Lauterbach. Finding, defining and breaking the diffraction barrier in microscopy–a historical perspective. *Optical Nanoscopy*, 1(1):8, 2012. 657, 664

[LDG00] Fabien Letouzey, François Denis, and Rémi Gilleron. Learning from positive and unlabeled examples. In *International Conference on Algorithmic Learning Theory*, pages 71–85. Springer, 2000. 412

[LeC73] Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973. 686

[Li91] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. 33, 102, 104, 181

[Li92] Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992. 33, 36, 101, 180

[Li18a] Chris Junchi Li. A note on concentration inequality for vector-valued martingales with weak exponential-type tails. *arXiv preprint arXiv:1809.02495*, 2018. 151, 152

[Li18b] Jerry Zheng Li. *Principled approaches to robust machine learning and beyond.* PhD thesis, Massachusetts Institute of Technology, 2018. 101, 229, 339

[Lia15] Wenjing Liao. Music for multidimensional spectral estimation: stability and super-resolution. *IEEE Transactions on Signal Processing*, 63(23):6395–6406, 2015. 617

[Lin95] B. Lindsay. *Mixture models: theory, geometry and applications.* Institute for Mathematical Statistics, 1995. 229

[LL18] Yuanzhi Li and Yingyu Liang. Learning mixtures of linear regressions with nearly optimal complexity. In *Conference On Learning Theory*, pages 1125–1144, 2018. 53, 60, 497, 498, 499, 500, 501, 506, 509, 512, 514, 532, 564, 576, 578, 580, 582, 583

[LLY$^+$12] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012. 503

[LM21a] Allen Liu and Ankur Moitra. Learning gmms with nearly optimal robustness guarantees. *arXiv preprint arXiv:2104.09665*, 2021. 39

[LM21b] Allen Liu and Ankur Moitra. Settling the robust learnability of mixtures of gaussians. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 518–531, 2021. 39, 57

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *Journal of the ACM (JACM)*, 40(3):607–620, 1993. 56, 95, 408, 412

[LMPL18] Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018. 342

[LMZ$^+$12] Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan. Robust and efficient subspace segmentation via least squares regression. In *European conference on computer vision*, pages 347–360. Springer, 2012. 503

[LMZ20] Yuanzhi Li, Tengyu Ma, and Hongyang R. Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In Jacob D. Abernethy and Shivani Agarwal 0001, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 2613–2682. PMLR, 2020. 34, 178

[LRR13] Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(1):295–347, 2013. 228

[LRV16] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016. 39, 227, 229, 231, 325

[LS59] Adrien-Marie Legendre and DE Smith. On the method of least squares. *A Source Book in Mathemathics, Ed. DE Smith (originally published in 1805)*, pages 576–579, 1959. 41

[LS17] Jerry Li and Ludwig Schmidt. Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. In *Conference on Learning Theory*, pages 1302–1382, 2017. 235

[LSM09] Jennifer Lippincott-Schwartz and Suliana Manley. Putting super-resolution fluorescence microscopy to work. *Nature Methods*, 6(ARTICLE):21–23, 2009. 657

[LSSS14] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in neural information processing systems*, pages 855–863, 2014. 180

[Luc92a] Leon B Lucy. Resolution limits for deconvolved images. *The Astronomical Journal*, 104:1260–1265, 1992. 656

[Luc92b] Leon B Lucy. Statistical limits to super resolution. *Astronomy and Astrophysics*, 261:706, 1992. 656

[LWW15] Xin-Guo Liu, Xue-Feng Wang, and Wei-Guo Wang. Maximization of matrix trace function of product stiefel manifolds. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1489–1506, 2015. 101

[LY17] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 597–607, 2017. 34, 178

[M$^+$15] Stanislav Minsker et al. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015. 336, 374, 375, 376, 377

[Man59] Leonard Mandel. Fluctuations of photon beams: the distribution of the photo-electrons. *Proceedings of the Physical Society*, 74(3):233, 1959. 655, 662

[MC16] Veniamin I Morgenshtern and Emmanuel J Candes. Super-resolution of positive sources: The discrete setup. *SIAM Journal on Imaging Sciences*, 9(1):412–444, 2016. 617

[MCSF10] Kim I Mortensen, L Stirling Churchman, James A Spudich, and Henrik Flyvbjerg. Optimized localization analysis for single-molecule tracking and super-resolution microscopy. *Nature methods*, 7(5):377, 2010. 656

[MdW16] Ashley Montanaro and Ronald de Wolf. A survey of quantum property testing. *Theory of Computing*, pages 1–81, 2016. 684

[MHC13] Bill Moran, Stephen Howard, and Doug Cochran. Positive-operator-valued measures: a general setting for frames. In *Excursions in Harmonic Analysis, Volume 2*, pages 49–64. Springer, 2013. 91

[Min61] M Minsky. Microscopy apparatus us patent 3013467. *USP Office, Ed. US*, 1961. 658

[Min17] Stanislav Minsker. On some extensions of bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017. 82

[MLX$^+$18] Chenglin Miao, Qi Li, Houping Xiao, Wenjun Jiang, Mengdi Huai, and Lu Su. Towards data poisoning attacks in crowd sensing systems. In *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 111–120, 2018. 26

[MM13] Elizabeth Meckes and Mark Meckes. Spectral measures of powers of random matrices. *Electronic communications in probability*, 18, 2013. 93, 683, 706, 722

[MMBS13] Bamdev Mishra, Gilles Meyer, Francis Bach, and Rodolphe Sepulchre. Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4):2124–2149, 2013. 101

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. 180

[MMR$^+$17] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017. 228

[Moi15] Ankur Moitra. Super-resolution, extremal functions and the condition number of vandermonde matrices. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 821–830. ACM, 2015. 62, 503, 615, 616, 617, 620, 621, 632, 633

[Mon13] Ashley Montanaro. Weak multiplicativity for random quantum channels. *Communications in Mathematical Physics*, 319(2):535–555, 2013. 92

[MOS03] Elchanan Mossel, Ryan O'Donnell, and Rocco P Servedio. Learning juntas. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 206–212. ACM, 2003. 33, 97, 408

[MR18] Pasin Manurangsi and Daniel Reichman. The computational complexity of training relu (s). *arXiv preprint arXiv:1810.04207*, 2018. 34, 178

[MSS16] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 438–446. IEEE, 2016. 89, 98

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010. 51, 57, 60, 62, 229, 409, 502, 506, 513, 616, 617, 677

[MW17] AA Maznev and OB Wright. Upholding the diffraction limit in the focusing of light and sound. *Wave Motion*, 68:182–189, 2017. 657, 658, 671

[Nes00] Yurii Nesterov. *Squared Functional Systems and Optimization Problems*, pages 405–440. Springer US, Boston, MA, 2000. 88

[NJS13] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013. 33, 36, 53, 96, 98, 102, 106, 497

[NO20] Gergely Neu and Julia Olkhovskaya. Efficient and robust algorithms for adversarial linear contextual bandits. *arXiv preprint arXiv:2002.00287*, 2020. 342

[Nob14] The nobel prize in chemistry 2014. Oct 2014. 29

[NSW19] Vasileios Nakos, Zhao Song, and Zhengyu Wang. (Nearly) sample-optimal sparse Fourier transform in any dimension; RIPless and Filterless. In *FOCS*, 2019. 503

[NWL16] Matey Neykov, Zhaoran Wang, and Han Liu. Agnostic estimation for misspecified phase retrieval models. In *Advances in Neural Information Processing Systems*, pages 4089–4097, 2016. 33, 36, 101

[O'B16] Carl M O'Brien. Nonparametric estimation under shape constraints: Estimators, algorithms and asymptotics. *International Statistical Review*, 84(2):318–319, 2016. 229

[O'D14] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014. 78, 85

[Ovc02] Sergei Ovchinnikov. Max-min representation of piecewise linear functions. *Contributions to Algebra and Geometry*, 43(1):297–302, 2002. 188, 193

[OW15] Ryan O'Donnell and John Wright. Quantum spectrum testing. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 529–538, 2015. 65, 69, 679, 680, 682, 684, 694

[OW16] Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the48th Annual ACM symposium on Theory of Computing*, pages 899–912, 2016. 66, 680, 685

[OW17] Ryan O'Donnell and John Wright. Efficient quantum tomography ii. In *Proceedings of the 49th Annual ACM Symposium on Theory of Computing*, pages 962–974, 2017. 680, 685

[OZ13] Ryan O'Donnell and Yuan Zhou. Approximability and proof complexity. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1537–1556. Society for Industrial and Applied Mathematics, 2013. 87

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. 69, 681, 686, 688, 694, 716, 719

[Par00] Pablo A. Parrilo. Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization. Technical report, California Institute of Technology, 2000. 88

[PDL84] Dieter W Pohl, Winfried Denk, and Mark Lanz. Optical stethoscopy: Image recording with resolution $\lambda/20$. *Applied physics letters*, 44(7):651–653, 1984. 658

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894. 57

[Pea00] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900. 64

[PF20] Scott Pesme and Nicolas Flammarion. Online robust regression via sgd on the l1 loss. *arXiv preprint arXiv:2007.00399*, 2020. 341

[PH05] Elizabeth Purdom and Susan P Holmes. Error distribution for gene expression data. *Statistical applications in genetics and molecular biology*, 4(1), 2005. 27

[PHL04] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explorations Newsletter*, 6(1):90–105, 2004. 502

[Pin94] Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994. 80, 596

[PJL20] Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020. 339

[Pol91] David Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991. 340

[PPG18] I Wayan Pulantara, Bambang Parmanto, and Anne Germain. Development of a just-in-time adaptive mhealth intervention for insomnia: usability study. *JMIR human factors*, 5(2):e8905, 2018. 27

[PS15] Eric Price and Zhao Song. A robust sparse Fourier transform in the continuous setting. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 583–600. IEEE, 2015. 503

[PSB$^+$20] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, Pradeep Ravikumar, et al. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B*, 82(3):601–627, 2020. 339

[PSBR20] Adarsh Prasad, Vishwak Srinivasan, Sivaraman Balakrishnan, and Pradeep Ravikumar. On learning ising models under huber's contamination model. *Advances in neural information processing systems*, 33, 2020. 39

[PV13] Yaniv Plan and Roman Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013. 102

[PV16] Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016. 33, 36, 101, 180

[PVY17] Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017. 33, 36, 101

[QV17] Mingda Qiao and Gregory Valiant. Learning discrete distributions from untrusted batches. *arXiv preprint arXiv:1711.08113*, 2017. 39, 40, 223, 224, 225, 226, 227, 228

[Rac03] Svetlozar Todorov Rachev. *Handbook of Heavy Tailed Distributions in Finance: Handbooks in Finance, Book 1*. Elsevier, 2003. 27

[Rao69] B.L.S. Prakasa Rao. Estimation of a unimodal density. *Sankhya Ser. A*, 31:23–36, 1969. 229

[Ray79] Lord Rayleigh. Investigations in optics, with special reference to the spectroscope. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 8(49):261–274, 1879. 28, 610, 664, 666

[RBZ06] Michael J Rust, Mark Bates, and Xiaowei Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm). *Nature methods*, 3(10):793, 2006. 658

[RCK41] BP Ramsay, EL Cleveland, and OT Koppius. Criteria and the intensity-epoch slope. *JOSA*, 31(1):26–33, 1941. 667

[RH17] Philippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *URL http://www-math. mit. edu/˜ rigollet/PDFs/RigNotes17. pdf*, 2017. 329, 360, 371

[Ric07] James H Rice. Beyond the diffraction limit: far-field fluorescence imaging with ultrahigh resolution. *Molecular BioSystems*, 3(11):781–793, 2007. 657

[Riv87] Ronald L Rivest. Learning decision lists. *Machine learning*, 2(3):229–246, 1987. 408, 412

[Ron61] Vasco Ronchi. Resolving power of calculated and detected images. *JOSA*, 51(4):458_1–460, 1961. 670

[Ros58] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958. 25

[RSS18] Prasad Raghavendra, Tselil Schramm, and David Steurer. High-dimensional estimation via sum-of-squares proofs. *arXiv preprint arXiv:1807.11419*, 2018. 232

[RST09] Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2009. 74

[Rus34] Ernst Ruska. Über fortschritte im bau und in der leistung des magnetischen elektronenmikroskops. *Zeitschrift für Physik A Hadrons and Nuclei*, 87(9):580–602, 1934. 658

[RV17] Oded Regev and Aravindan Vijayaraghavan. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96. IEEE, 2017. 51, 612, 614, 616, 617

[RW84] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26:195–202, 1984. 229

[RWO06] Sripad Ram, E Sally Ward, and Raimund J Ober. Beyond rayleigh's criterion: a resolution measure with application to single-molecule microscopy. *Proceedings of the National Academy of Sciences*, 103(12):4457–4462, 2006. 655

[RY19] Prasad Raghavendra and Morris Yau. List decodable learning via sum of squares. *arXiv preprint arXiv:1905.04660*, 2019. 229, 498, 502

[SAT96] Nicholas J Schork, David B Allison, and Bonnie Thiel. Mixture distributions in human genetics research. *Statistical Methods in Medical Research*, 5(2):155–178, 1996. 28

[SBRJ19] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pages 2892–2897, 2019. 41, 341

[Sch04] Arthur Schuster. *An introduction to the theory of optics.* E. Arnold, 1904. 610, 665, 667

[SCV18] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018. 229, 279, 291

[SF20] Takeyuki Sasai and H. Fujisawa. Robust estimation with lasso when outputs are adversarially contaminated. *ArXiv*, abs/2004.05990, 2020. 340

[Sha18] Ohad Shamir. Distribution-specific hardness of learning neural networks. *Journal of Machine Learning Research*, 19(32):1–29, 2018. 180

[She17] Colin JR Sheppard. Resolution and super-resolution. *Microscopy research and technique*, 80(6):590–598, 2017. 658, 672

[SHKT97] C. J. Stone, M. H. Hansen, C. Kooperberg, and Y. K. Truong. Polynomial splines and their tensor products in extended linear modeling: 1994 wald memorial lecture. *Ann. Statist.*, 25(4):1371–1470, 1997. 229

[Sho87] N.Z. Shor. Quadratic optimization problems. *Soviet Journal of Computer and Systems Sciences*, 25, 11 1987. 88

[SJA16] Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*, pages 1223–1231, 2016. 53, 497, 498, 501

[SJG09] Hao Shen, Stefanie Jegelka, and Arthur Gretton. Fast kernel-based independent component analysis. *IEEE Transactions on Signal Processing*, 57(9):3498–3511, 2009. 101

[SK12] Peter Stobbe and Andreas Krause. Learning fourier sparse set functions. In *Artificial Intelligence and Statistics*, pages 1125–1133, 2012. 412

[SKL17] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3520–3532, 2017. 26

[SL17] Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 1743–1759, 2017. 342

[SLG$^+$15] Noa Slater, Yoram Louzoun, Loren Gragert, Martin Maiers, Ansu Chatterjee, and Mark Albrecht. Power laws for heavy-tailed distributions: modeling allele and haplotype diversity for the national marrow donor program. *PLoS computational biology*, 11(4):e1004204, 2015. 27

[SLX20] David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Available at SSRN*, 2020. 331, 375

[SM04] Morteza Shahram and Peyman Milanfar. Imaging below the diffraction limit: a statistical analysis. *IEEE Transactions on image processing*, 13(5):677–689, 2004. 655, 656

[SM06] Morteza Shahram and Peyman Milanfar. Statistical and information-theoretic analysis of resolution in imaging. *IEEE Transactions on information Theory*, 52(8):3411–3437, 2006. 655, 656

[Spa16] Carroll Mason Sparrow. On spectroscopic resolving power. *The Astrophysical Journal*, 44:76, 1916. 610, 665, 668

[SQW16] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2016. 101

[SRH07] Srinath Sridhar, Satish Rao, and Eran Halperin. An efficient and accurate graph-based approach to detect population substructure. In *Research in Computational Molecular Biology*, pages 503–517. Springer, 2007. 51

[SS$^+$11] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011. 337

[SS14a] Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32*, pages II–1287, 2014. 342

[SS14b] Alex Small and Shane Stahlheber. Fluorophore localization algorithms for super-resolution microscopy. *Nature methods*, 11(3):267, 2014. 656

[SS17] Tselil Schramm and David Steurer. Fast and robust tensor decomposition with applications to dictionary learning. In *Conference on Learning Theory*, pages 1760–1793, 2017. 98

[SS18] Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. *arXiv preprint arXiv:1810.11874*, 2018. 101

[SS19] Yanyao Shen and Sujay Sanghavi. Iterative least trimmed squares for mixed linear regression. In *Advances in Neural Information Processing Systems*, pages 6076–6086, 2019. 101

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. 329

[SSSS17] Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Failures of gradient-based deep learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3067–3075, 2017. 180

[SST10] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010. 336, 373, 374

[Ste18] Jacob Steinhardt. *Robust Learning: Information Theory and Algorithms*. PhD thesis, Stanford University, 2018. 229, 339

[Sto94] C. J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, 22(1):pp. 118–171, 1994. 229

[SVWX17] Le Song, Santosh Vempala, John Wilmes, and Bo Xie. On the complexity of learning neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5520–5528, 2017. 180

[Syn28] EdwardH Synge. Xxxviii. a suggested method for extending microscopic resolution into the ultra-microscopic region. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(35):356–362, 1928. 658

[TBSR13] Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013. 617, 618

[TD79] Ming-Jer Tsai and Keh-Ping Dunn. Performance limitations on parameter estimation of closely spaced optical targets using shot-noise detector model. Technical report, MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB, 1979. 655

[Tem81] Paul A Temple. Total internal reflection microscopy: a surface inspection technique. *Applied optics*, 20(15):2656–2664, 1981. 658

[THK$^+$21] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021. 24

[Tho69] Brian J Thompson. Iv image formation with partially coherent light. In *Progress in optics*, volume 7, pages 169–230. Elsevier, 1969. 668

[Tia17] Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3404–3413. PMLR, 2017. 34, 178

[Tim14] Aleksandr Filippovich Timan. *Theory of approximation of functions of a real variable*, volume 34. Elsevier, 2014. 227

[TK08] Ambuj Tewari and Sham Kakade. Lectures notes for cmsc 35900: Learning theory, 2008. 337

[TKV17] Kevin Tian, Weihao Kong, and Gregory Valiant. Learning populations of parameters. In *Advances in Neural Information Processing Systems*, pages 5778–5787, 2017. 228

[TM$^+$14] Dan-Cristian Tomozei, Laurent Massoulié, et al. Distributed user profiling via spectral methods. *Stochastic Systems*, 4(1):1–43, 2014. 51

[Tro11] Joel A Tropp. User-friendly tail bounds for matrix martingales. Technical report, CALIFORNIA INST OF TECH PASADENA, 2011. 82

[Tro12] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012. 82

[Tsa18] Mankei Tsang. Conservative classical and quantum resolution limits for incoherent imaging. *Journal of Modern Optics*, 65(11):1385–1391, 2018. 657

[Tsa19] Mankei Tsang. Resolving starlight: a quantum perspective. *arXiv preprint arXiv:1906.02064*, 2019. 656

[TSM85] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical analysis of finite mixture distributions*. Wiley & Sons, 1985. 229

[Tsy08] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008. 353

[Tuk60] John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960. 229

[Tuk75] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975. 229

[TV15] Manolis C Tsakiris and René Vidal. Dual principal component pursuit. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–18, 2015. 503

[TV17] Manolis C Tsakiris and René Vidal. Hyperplane clustering via dual principal component pursuit. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3472–3481. JMLR. org, 2017. 503

[TV18] Yan Shuo Tan and Roman Vershynin. Polynomial time and sample complexity for non-gaussian component analysis: Spectral methods. In *Conference On Learning Theory*, pages 498–534, 2018. 181

[TXSS20] Farnaz Tahmasebian, Li Xiong, Mani Sotoodeh, and Vaidy Sunderam. Crowdsourcing under data poisoning attacks: A comparative study. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 310–332. Springer, 2020. 26

[Udr94] Constantin Udriste. *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media, 1994. 101

[Vaa85] Jeffrey D Vaaler. Some extremal functions in fourier analysis. *Bulletin of the American Mathematical Society*, 12(2):183–216, 1985. 644

[Vai89] Pravin M Vaidya. A new algorithm for minimizing convex functions over convex sets. In *30th Annual Symposium on Foundations of Computer Science*, pages 338–343. IEEE Computer Society, 1989. 389, 390

[Val84] LG Valiant. A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445. ACM, 1984. 443

[Val12] Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 11–20. IEEE, 2012. 408

[Van13] Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. 101

[VC74] Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974. 228

[VDB01] Adriaan Van Den Bos. Resolution in model-based measurement. In *IMTC 2001. Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference. Rediscovering Measurement in the Age of Informatics (Cat. No. 01CH 37188)*, volume 1, pages 295–302. IEEE, 2001. 655

[VdBDD01] A Van den Bos and AJ Den Dekker. Resolution reconsidered—conventional approaches and an alternative. In *Advances in imaging and electron physics*, volume 117, pages 241–360. Elsevier, 2001. 655

[vDSM17] Alex von Diezmann, Yoav Shechtman, and WE Moerner. Three-dimensional localization of single molecules for super-resolution imaging and single-particle tracking. *Chemical reviews*, 117(11):7244–7275, 2017. 656

[Vem10a] Santosh S Vempala. Learning convex concepts from gaussian distributions with pca. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 124–130. IEEE, 2010. 95, 102, 181

[Vem10b] Santosh S Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *Journal of the ACM (JACM)*, 57(6):1–14, 2010. 102, 181

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010. 81, 84

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018. 78, 80, 83, 349, 360, 361, 371, 376, 564, 605

[VH07] Rene Vidal and Richard Hartley. Three-view multibody structure from motion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):214–227, 2007. 502

[Vid11] René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011. 502

[VMS05] Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (gpca). *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1945–1959, 2005. 503

[Vov01] Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001. 326, 343

[Vu02] Van H Vu. Concentration of non-lipschitz functions and applications. *Random Structures &amp; Algorithms*, 20(3):262–316, 2002. 152

[Vu06] VH Vu. On the infeasibility of training neural networks with small mean-squared error. *IEEE Transactions on Information Theory*, 44(7):2892–2900, 2006. 25, 180

[VV10] Bart Vandereycken and Stefan Vandewalle. A riemannian optimization approach for computing low-rank solutions of lyapunov equations. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2553–2579, 2010. 101

[VV16] Gregory Valiant and Paul Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 142–155, New York, NY, USA, 2016. Association for Computing Machinery. 70

[VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017. 70, 715, 717, 718, 719, 720, 721, 727, 740

[VW02] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*, pages 113–122, 2002. 51, 229, 617

[VW19] Santosh Vempala and John Wilmes. Gradient descent for one-hidden-layer neural networks: Polynomial convergence and sq lower bounds. In *COLT*, volume 99, 2019. 34, 178, 179, 180

[VX11] Santosh S Vempala and Ying Xiao. Structure from local optima: Learning subspace juntas via higher order pca. *arXiv preprint arXiv:1108.3329*, 2011. 102, 180, 181

[Wal09] G. Walther. Inference and modeling with log-concave distributions. *Statistical Science*, 24(3):319–327, 2009. 229

[Weg70] Edward J Wegman. Maximum likelihood estimation of a unimodal density function. *The Annals of Mathematical Statistics*, 41(2):457–471, 1970. 226, 229

[Wei78] Don Weingarten. Asymptotic behavior of group integrals in the limit of infinite rank. *Journal of Mathematical Physics*, 19(5):999–1001, 1978. 707

[WGFC14] Nathan Wiebe, Christopher Granade, Christopher Ferrie, and David G Cory. Hamiltonian learning and certification using quantum resources. *Physical review letters*, 112(19):190501, 2014. 29

[Wil50] W.E. Williams. *Applications of Interferometry*. Methuen's monographs on physical subjects. Methuen, 1950. 666

[WN07] R. Willett and R. D. Nowak. Multiscale poisson intensity and density estimation. *IEEE Transactions on Information Theory*, 53(9):3171–3187, 2007. 229

[WRH17] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 7032–7042, 2017. 27

[Wri16] John Wright. *How to learn a quantum state*. PhD thesis, Carnegie Mellon University Pittsburgh, PA, 2016. 66, 681, 682, 694

[WS15] Siegfried Weisenburger and Vahid Sandoghdar. Light microscopy: an ongoing contemporary revolution. *Contemporary Physics*, 56(2):123–143, 2015. 657, 664

[WW83] E. J. Wegman and I. W. Wright. Splines in statistics. *Journal of the American Statistical Association*, 78(382):pp. 351–365, 1983. 229

[YBL17] Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. On stein's identity and near-optimal estimation in high-dimensional index models. *arXiv preprint arXiv:1709.08795*, 2017. 33, 36, 101

[YCS14] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621. PMLR, 2014. 53, 330, 497, 501

[YCS16] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749*, 2016. 53, 497, 498, 501, 605

[YJY09] Liu Yang, Rong Jin, and Jieping Ye. Online learning by ellipsoid method. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1153–1160, 2009. 337

[Yu19] Nengkun Yu. Quantum closeness testing: A streaming algorithm and applications, 2019. 684

[ZAR14] Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing long-tail distributions of object subcategories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922, 2014. 27

[ZB16] Dejiao Zhang and Laura Balzano. Global convergence of a grassmannian gradient descent algorithm for subspace estimation. In *Artificial Intelligence and Statistics*, pages 1460–1468. PMLR, 2016. 101

[ZBFL18] Wen-Xin Zhou, Koushiki Bose, Jianqing Fan, and Han Liu. A new perspective on robust m-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Annals of statistics*, 46(5):1904, 2018. 340

[Zin03] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003. 338, 393

[ZJD16] Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Mixed linear regression with multiple components. In *Advances in neural information processing systems*, pages 2190–2198, 2016. 53, 60, 497, 498, 501

[ZJS20] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Robust estimation via generalized quasi-gradients. *arXiv preprint arXiv:2005.14073*, 2020. 41, 45, 47, 325, 339

[ZLJ16] Yuchen Zhang, Jason D Lee, and Michael I Jordan. L1-regularized neural networks are improperly learnable in polynomial time. In *33rd International Conference on Machine Learning, ICML 2016*, pages 1555–1563. International Machine Learning Society (IMLS), 2016. 34, 178, 179

[Zmu03] Jonas Zmuidzinas. Cramer–rao sensitivity limits for astronomical instruments: implications for interferometer design. *JOSA A*, 20(2):218–233, 2003. 656

[ZPS17] Qiuyi Zhang, Rina Panigrahy, and Sushant Sachdeva. Electron-proton dynamics in deep learning. *CoRR*, abs/1702.00458, 2017. 34, 178

[ZRS16] Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. Riemannian svrg: Fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 4592–4600, 2016. 101

[ZS16] Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638, 2016. 101

[ZS19] Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 467–475. PMLR, 2019. 342

[ZSJ$^+$17] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4140–4149, 2017. 34, 36, 178, 179

[ZYWG19] Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1524–1534. PMLR, 2019. 34, 178