

Efficient Algorithms for Learning Mixture Models

by

Qingqing Huang

BEng, BBA, Hong Kong University of Science and Technology (2011)
S. M, Massachusetts Institute of Technology(2013)

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

at MASSACHUSETTS INSTITUTE OF TECHNOLOGY

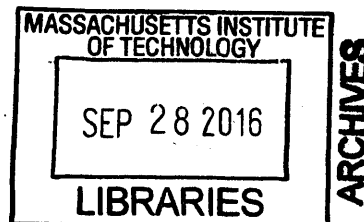
September 2016

© Massachusetts Institute of Technology 2016. All rights reserved.

Author **Signature redacted**
Department of Electrical Engineering and Computer Science
August 12, 2016

Certified by **Signature redacted**
Munther A. Dahleh
Professor of Electrical Engineering at MIT
Thesis Supervisor

Accepted by **Signature redacted**
✓ ✓ ✓ Leslie A. Kolodziejski
Chairman, Department Committee on Graduate Theses





77 Massachusetts Avenue
Cambridge, MA 02139
<http://libraries.mit.edu/ask>

DISCLAIMER NOTICE

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort possible to provide you with the best copy available.

Thank you.

The images contained in this document are of the best quality available.

Efficient Algorithms for Learning Mixture Models

by

Qingqing Huang

Submitted to Department of Electrical Engineering and Computer Science
on August 12, 2016, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

We study the statistical learning problems for a class of probabilistic models called mixture models. Mixture models are usually used to model settings where the observed data consists of different sub-populations, yet we only have access to a limited number of samples of the pooled data. It includes many widely used models such as Gaussian mixtures models, Hidden Markov Models, and topic models. We focus on parametric learning: given unlabeled data generated according to a mixture model, infer about the parameters of the underlying model. The hierarchical structure of the probabilistic model leads to non-convexity of the likelihood function in the model parameters, thus imposing great challenges in finding statistically efficient and computationally efficient solutions.

We start with a simple, yet general setup of mixture model in the first part. We study the problem of estimating a low rank $M \times M$ matrix which represents a discrete distribution over M^2 outcomes, given access to sample drawn according to the distribution. We propose a learning algorithm that accurately recovers the underlying matrix using $\Theta(M)$ number of samples, which immediately lead to improved learning algorithms for various mixture models including topic models and HMMs. We show that the linear sample complexity is actually optimal in the min-max sense.

There are “hard” mixture models for which there exist worst case lower bounds of sample complexity that scale exponentially in the model dimensions. In the second part, we study Gaussian mixture models and HMMs. We propose new learning algorithms with polynomial runtime. We leverage techniques in probabilistic analysis to prove that worst case instances are actually rare, and our algorithm can efficiently handle all the non-worst case instances. In the third part, we study the problem of super-resolution. Despite the lower bound for any deterministic algorithm, we propose a new randomized algorithm which complexity scales only quadratically in all dimensions, and show that it can handle any instance with high probability over the randomization.

Thesis Supervisor: Munther A. Dahleh

Title: Professor of Electrical Engineering at MIT

Acknowledgments

First of all, I would like to express my sincere gratitude to my advisor Munther Dahleh for his guidance and support during my five years at MIT. Munther taught me to think out of the box and always keep the big picture in mind. I am grateful to Sham Kakade for his great guidance and patience when I first stepped in the field of algorithmic statistics. I also thank Pablo Parrilo for being on my thesis committee and giving helpful comments on this work. I sincerely thank my collaborators: Rong Ge, Na Li, Gregory Valiant, Ye Yuan, Tong Zhang, who are great teachers and friends, being most valuable sources of inspiration and encouragement on my way to becoming a better researcher. I also thank Guy Bresler, Ankur Moitra, Praneeth Netrapalli, Yury Polyanskiy, Devavrat Shah, John Tsitsiklis, with whom I had numerous insightful and fun discussions on research problems.

I thank my friends in Laboratory of Information and Decision System – Elie Adam, Giancarlo Baldan, Diego Cifuentes, Hamza Fawzi, Kimon Drakopoulos, Christina Lee, Shreya Saxena, Jennifer Tang, Omer Tanovic – for making it such a great place to work. Special thanks to Shreya and Elie for being amazing coffee-mates. I thank my piano teachers Tal and Lucia, without whom I would probably have written more papers but much less fun during my Ph.D study.

I thank Mu Li, who shares my laughter and tears, joy and pain throughout the journey. Most importantly, I thank my parents, whose love have always been my stronghold of support and encouragement.

This work was supported in part by AOR with UPenn under Grant No. 6927221, and part of the work was done during internship at Microsoft Research New England.

Contents

1	Introduction	11
1.1	Background	11
1.2	Contributions	15
1.2.1	Part 1: Achieving optimal sample complexity	16
1.2.2	Part 2: Randomized Analysis to Escape Worst Cases	18
1.2.3	Part 3: Randomized Algorithm to Escape Worst Cases	20
1.3	Preliminaries	21
1.3.1	Notations	21
1.3.2	Tensor Algebra	22
1.3.3	Probabilistic analysis	26
2	Recovering structured matrices	31
2.1	Problem Statement	31
2.1.1	Formulation	34
2.1.2	Related Work	35
2.2	Main Results	38
2.2.1	Recovering Low Rank Probability Matrices	38
2.2.2	Topic Models and Hidden Markov Models	41
2.2.3	Testing vs. Learning	43
2.3	Outline of our estimation algorithm	44
2.3.1	Rank 2 algorithm	46
2.3.2	Rank R algorithm	48
2.4	Details of Rank 2 Algorithm	52

2.4.1	Binning	54
2.4.2	Estimate segments of Δ	57
2.4.3	Stitch the segments of $\hat{\Delta}$	61
2.4.4	Refinement	62
2.5	Details of Rank R Algorithm	65
2.5.1	Binning	65
2.5.2	Spectral concentration in diagonal blocks	67
2.5.3	Low rank projection	68
2.5.4	Refinement	69
2.6	Sample complexity lower bounds	70
2.7	Proofs for Chapter 2	74
2.7.1	Proofs for Rank 2 Algorithm, Phase I	74
2.7.2	Proofs for Rank 2 Algorithm Phase II	88
2.7.3	Proofs for Rank R Algorithm	94
2.7.4	Proofs for HMM testing lower bound	106
2.7.5	Analyze truncated SVD	113
2.7.6	Auxiliary Lemmas	116
3	Learning Gaussian Mixtures in High Dimensions	119
3.1	Problem Statement	119
3.1.1	Formulation	119
3.1.2	Related Work	119
3.2	Main results	122
3.3	Outline of our algorithm	128
3.3.1	Learning Mixture of Zero-Mean Gaussians	128
3.3.2	Implementing the Steps for Zero-Mean Algorithm	129
3.3.3	Learning Mixture of General Gaussians	137
3.4	Proofs for Chapter 3	139
3.4.1	Step 1 of Zero-Mean Case: Span Finding	139
3.4.2	Step 2 of Zero-Mean Case. Moments Unfolding	160

3.4.3	Step 3 of Zero-Mean Case: Tensor Decomposition	172
3.4.4	Proof of Theorem 3.5	176
3.4.5	Proofs for the General Case	179
3.4.6	Proofs for Moment Structures	191
3.4.7	Auxiliary Lemmas	196
4	Realization Problems of Hidden Markov Models	207
4.1	Problem Statement	207
4.2	Main results	211
4.2.1	Minimal Quasi-HMM Realization	212
4.2.2	Minimal HMM Realization Problem	218
4.3	Proofs for Chapter 4	222
4.3.1	Proofs	222
4.3.2	Other proofs	226
5	Super-resolution	233
5.1	Problem Statement	233
5.1.1	Formulation	233
5.1.2	Related Work	235
5.2	Main Results	238
5.2.1	Warm-up	238
5.2.2	Our Algorithm	242
5.2.3	Performance Guarantees	244
5.2.4	Key Lemmas	245
5.3	Discussions	247
5.3.1	Numerical results	247
5.3.2	Connection with learning GMMs	250
5.3.3	Open problems	252
5.4	Proofs for Chapter 5	253

List of Figures

2-1	The key algorithmic ideas of our algorithm.	45
2-2	block decomposition of the diagonal block of $B_{\hat{\mathcal{I}}_k, \hat{\mathcal{I}}_k}$ corresponding to $\hat{\mathcal{I}}_k$. 60	
2-3	block decomposition of the diagonal block of B_{N2} corresponding to $\hat{\mathcal{I}}_k$. 80	
2-4	decomposition of \tilde{E} corresponding to $\hat{\mathcal{I}}_k, \hat{\mathcal{L}}_k$ and $\hat{\mathcal{R}}_k$	101
3-1	Flow of the algorithm for learning mixture of zero-mean Gaussians. . .	137
3-2	Flow of the algorithm for learning mixtures of general Gaussians. . .	137
3-3	Structure of the matrix B_S	141
3-4	Step 1(c): Merging two subspaces.	153
3-5	Flow of the algorithm for the general case	190
4-1	Reduction of HMM to noisy parity	216
5-1	Simulation result for 2-D super-resolution	248
5-2	Cutoff frequency versus the required minimal separation	249
5-3	Number of measurements versus the required minimal separation . . .	250

Chapter 1

Introduction

1.1 Background

The era of “big data” brings us an abundance of data, and also presents a great variety of possibilities of statistical modeling, learning, and using the data, for applications in image and video signal processing, and natural language processing. In practice, we observe incomplete, noisy, uncertain data samples, which reveal information about the underlying rules according to which the data is generated. In the presence of large amount of data and complicated models, the goal is to design a statistically efficient and computationally efficient learning procedure to infer about the underlying rule according to which the noisy data is generated. Namely, we want to extract as much as possible information from available data with fast computation.

In particular, let M denote the degree-of-freedom of the underlying model, and let ϵ denote the target accuracy in estimating the model parameters. We evaluate the “efficiency” of a learning algorithm by its sample complexity, namely how the required sample size scales with M and ϵ , and its computational complexity, namely how the algorithm runtime scales with M . Moreover, in the regime where the data is scarce, or the model size is large, i.e. M is large or even scales up with the data sample size, a low sample complexity for learning is crucial. We would like to understand the information theoretical lower bound of sample complexity, below which it is impossible for any algorithm to learn, and whether we can have a fast algorithm that achieves

the optimal sample complexity.

Mixture model In this work, we focus on mixture models. Mixture models refer to a class of statistical models which includes Gaussian mixtures, Hidden Markov models and Stochastic block models that are commonly used in practice and well-studied in the literature. In abstract, we assume that there are N data samples $X = (X_1, \dots, X_N)$ generated according to some underlying probabilistic model $\Pr(X; \theta)$ specified by its model parameters θ . Mixture models impose additional structural properties on $\Pr(X; \theta)$ such that it can be parameterized with much lower degree of freedom. In particular, we assume that there exists a latent factor H , and the value it takes for each sample point, i.e. (H_1, \dots, H_N) is not observed. When conditional the value H takes, we can characterize the distribution of the data point with relatively simple probabilistic rules. Namely, the conditional distribution $\Pr(X; \theta|H)$ is some simple probabilistic model. We are interested in learning about $\Pr(X; \theta)$ without observe the latent variable H .

Compared to deep neural networks, which usually have multiple layers of latent variables, mixture models are “shallow” with only one latent layer. However, this single latent variable is already powerful in modeling different problems. Usually with the latent variable a mixture model can provide a concise description of the observed data and thus enhance the data interpretability. Modeling and using mixture models has enjoyed a great success in various machine learning applications. For example, Gaussian mixture models are used for clustering and factor analysis to identify different populations in social science, topic models are used for unsupervised document classification, and in natural language processing, Hidden Markov models are used to model language where the latent variables are the speech tags.

Approaches for learning mixture models The structures of mixture models also impose challenges that makes parametric learning fundamentally different and more difficult than that of simple statistical models. On one hand, with the latent variables, mixture models are parameterized with lower degree of freedom and

thus have lower model class complexity. However, it is not straightforward how to exploit such structured lower complexity to improve algorithmic efficiency in learning. On the other hand, the unobserved variable (H_1, \dots, H_N) makes the likelihood function $\Pr(X_1, \dots, X_N; \theta)$ a non-convex function in the parameters we are interested to learn. In this case, directly solving for the maximum likelihood estimator $\hat{\theta} = \arg \max_{\theta} \Pr(X_1, \dots, X_N; \theta)$ is computationally intractable.

There are different approaches for learning mixture models. Next, we briefly describe the two main approaches: 1. approximating the non-convex optimization for maximum likelihood estimator; 2. obtaining a consistent estimator by matching higher order moments with spectral methods.

Approximate MLE For parametric learning, the maximum likelihood estimation (MLE) and its variations are known to achieve statistical efficiency, and even survive (asymptotically) model misspecification [125]. However, in its original form, the likelihood function $\Pr(X_1, \dots, X_N; \theta)$ is a non-convex function in the parameters θ . With limited computation, one can only approximately solve the non-convex optimization.

The non-convex program can be “convexified” by techniques such as relaxing or restricting the support, change of variable, or modifying the non-convex objective functions. For special cases where strong assumptions are imposed, it is possible to rigorously bound the gap between the solution to the convex relaxation and the optimal solution, or even to show that the relaxation is exact. Examples include using nuclear norm minimization for low rank matrix sensing/ completion under RIP conditions [32, 99], using sum of square and positive semi-definite relaxations for dictionary learning [19], and using SDP relaxation for max-cut problem [6]. However in general, there is usually no theoretical guarantee on the quality of such convex relaxations. The other potential drawback is that the convexified problems may still be computationally challenging, for example consider a large size SDP from a sum-of-square approximation as an example, and this might significantly limit the applicability of such algorithms at large scale in practice.

There are also various heuristics which aim to directly tackle the non-convex op-

timizations. For example, Expectation-Maximization (EM) algorithm [86] alternates between the two steps of posteriori probabilities estimation and model parameter estimation while fixing the other until convergence. Similar heuristics include alternating minimization for matrix / tensor factorization [70] and low rank matrix completion/recovery [69], and brown clustering algorithm for feature extraction in language processing [79]. However, except for a few special cases [67] [126] with very strict assumptions on the model parameters, there is still a lack of rigorous study of the performance guarantee for such heuristics.

Spectral algorithms for moment matching Another way to estimate the model parameters is through moment matching. One starts with a set of equations that relate the exact higher order moments (namely the expected values of power of the random variables, or joint probabilities of subsets of discrete random variables) to the model parameters of interest. Then the sample data is used to estimate the higher order moments, and the equations are solved to give estimation of the parameters. For mixture models with simple conditional distribution $\Pr(X; \theta|H)$, moment matching usually specifies a system of polynomial equations linking the moments to the model parameters.

Spectral methods gain the name as it usually involve certain forms of spectral factorization of linear objects consisting of the estimated moments. It tries to exploit the structure properties of the polynomial equations to solve the equation system *algebraically*. Unlike the non-convex optimization for MLE, those linear algebra operations of the spectral methods based algorithms can usually be efficiently computed for moderate-sized problems. We refer to [98] for a general introduction of spectral methods for statistical learning and various applications.

There are two limitations of spectral methods. First, it relies on the fact that the moments are estimated sufficiently accurately and that the spectral factorizations are numerically stable to recover the parameters from the polynomial equations. The computational efficiency usually comes at the cost of a much higher sample complexity than that of MLE, namely lower tolerance of statistical noise. Second, the existing

spectral methods can only handle mixture models with very simple structures, such as spherical Gaussian mixtures, stochastic block model for community detection, and low rank matrix completion under RIP conditions. It is not clear whether there exist efficient spectral methods for learning many more general and more complicated mixture models.

Our approach In this work, we mostly adopt the approach of spectral algorithms for learning different mixture models. Our efforts are in two directions: 1. we want to improve the sample complexity of spectral methods to meet the information theoretical lower bound. 2. we want to have efficient spectral algorithms for learning the more general cases of mixture models and the cases which are not immediately modeled as a mixture models.

1.2 Contributions

The main question we are interested in is stated as follows:

“Can we have statistical and computational efficient parametric learning algorithms for learning mixture models?”

We address the above question from different perspectives in the three parts of this thesis.

In the first part, we start with a basic setting of mixture model. Despite the non-convex nature of the likelihood function, we show that it is still possible to exploit the structural properties of the underlying mixture model to efficiently estimate the model parameters. In particular, we propose a new spectral algorithm and show that it can achieve the minmax optimal sample complexity with fast and guaranteed computation, without directly solving for the maximum likelihood estimator.

Unfortunately, there are many well-studied “hard” mixture models, for which there exist sample complexity lower bounds that scale exponentially in model dimensions. Instances of model parameters have been constructed, for which it is impossible to obtain an accurate estimator with a polynomial number of samples and / or with

polynomial computation time. However, such worst case lower bounds usually do not give a full characterization of the set of the hard case, neither do they preclude algorithms that can efficiently learn the many non-worst-cases instances in the model class.

In the second part, we study Gaussian mixture models and Hidden Markov models and propose new deterministic parametric learning algorithms. We leverage recent development in probabilistic analysis to show that there are actually not too many hard instances in the model class, and our algorithms can handle all the rest of the instances with full polynomial sample complexity with polynomial runtime.

In the third part, we study the problem of super-resolution, where we explore the randomness in the learning algorithm (compared to the randomness in the analysis as in the second part) as a way to circumvent the worst case lower bounds. In particular, we show that the proposed randomized algorithm runs fast, and moreover, for every instance in the model class, the algorithm is efficient and outputs an accurate estimate, with high probability over the randomness of the algorithm.

Mixture models is a rich class and it includes many probabilistic models with distinguished structural properties. Unfortunately, we do not find a general recipe which can efficiently learn every mixture model. In order to achieve the desired learning goals, our case studies show that it is crucial to exploit the problem specific structures to obtain efficient learning algorithms.

In the rest of this chapter, we briefly introduce the formulations of the three parts of the thesis and state our main results for each problem. We provide the review of related literatures and detailed discussions in the later chapters, and for the convenience of reading, all the proofs are deferred to the end of each chapter.

1.2.1 Part 1: Achieving optimal sample complexity

In Chapter 2, we start with a basic yet general mixture model. We consider the problem of accurately recovering a matrix \mathbb{B} of size $M \times M$, which represents a probability distribution over M^2 outcomes, given access to “counts” generated by taking independent samples according to the distribution \mathbb{B} . How can structural properties of the

underlying matrix \mathbb{B} be leveraged to yield computationally efficient and information theoretically optimal reconstruction algorithms? When can accurate reconstruction be accomplished in the regime where the number of counts is relatively small compared to M ? This basic problem lies at the core of a number of questions that are currently being considered by different communities, including community detection in sparse random graphs, learning structured models such as topic models or hidden Markov models, and the efforts from the natural language processing community to compute “word embeddings”. Many aspects of this problem— in terms of both parameter estimation and property testing —remain open, on both the algorithmic and information theoretic sides.

Our results apply to the setting where \mathbb{B} has a particular low rank structure parameterized as $\mathbb{B} = PWP^\top$, where the columns of the tall matrix $P \in \mathbb{R}_+^{M \times R}$ are all supported on the standard $(M - 1)$ -simplex, and the mixing matrix $W \in \mathbb{R}_+^{R \times R}$ is a PSD matrix and $\sum_{i,j} W_{i,j} = 1$. For this setting, we propose an efficient (and practically viable) algorithm that accurately recovers the underlying $M \times M$ matrix using $\Theta(M)$ samples. Note that it is relatively easy to have algorithms whose sample complexity scales as $\Theta(M \log M)$. However, it requires extra efforts to push it all the way to the linear sample complexity that matches the information theoretical lower bound. Moreover, this extremely sparse data regime is meaningful for a lot of realistic applications.

Theorem 1.1. *Suppose we have access to N i.i.d. samples generated according to the a probability matrix \mathbb{B} . Fix the target accuracy $\epsilon_0 = \Omega(1)$, for any $r > 0$, with $N = \Theta(MR^2\epsilon_0^{-(4+r)})$ samples, our algorithm runs in time $O(M^3)$ and returns a rank R estimator \hat{B} such that with a large probability over the random sampling procedure, $\|\hat{B} - \mathbb{B}\|_1 \leq \epsilon_0$.*

If we further assume that the model parameters satisfy certain eigen-gap assumptions, we can sharpen the sample complexity to $N = \Theta(\max(MR^2\epsilon_0^{-(4+r)}, MR\epsilon^{-2}))$, for arbitrary target accuracy ϵ . This result easily translates to $\Theta(M)$ sample algorithms for learning topic models over dictionaries of size M , and learning hidden

Markov Models with observation distributions supported over M elements.

These linear sample complexities are optimal, up to constant factors, in an extremely strong sense: even testing basic properties of the underlying matrix, such as whether it has rank 1 or 2, requires $\Omega(M)$ samples. We provide an even stronger lower bound where distinguishing whether a sequence of observations were drawn from the uniform distribution over M observations versus being generated by an HMM with two hidden states requires $\Omega(M)$ observations. This precludes sublinear-sample hypothesis tests for basic properties as well as precludes sublinear sample estimators for quantities such as the HMM entropy rate.

1.2.2 Part 2: Randomized Analysis to Escape Worst Cases

There are mixture models for which parametric learning is usually deemed “hard” due to the existing lower bound results, which essentially precludes any upperbound algorithms that can efficiently learn *every* instance in the model class. In Chapter 3 and Chapter 4, we examine two such classes of “hard” mixture models separately: Gaussian mixture models (GMMs) and Hidden Markov models (HMMs). We make use of recent development in probabilistic analysis to show that the bad instances are actually rare, and we can have efficient learning algorithms for all the good instances.

Efficiently learning mixture of Gaussians is a fundamental problem in statistics and learning theory. Given samples coming from a random one out of k Gaussian distributions in n -dimensional space, the learning problem asks to estimate the means and the covariance matrices of these Gaussians. This problem arises in many areas ranging from the natural sciences to the social sciences, and has also found many machine learning applications. Unfortunately, learning mixture of Gaussians is an information theoretically hard problem: in the worst case, in order to learn the parameters up to a reasonable accuracy, the number of samples required scales exponentially in the number of Gaussian components.

We propose a deterministic algorithm for learning Gaussian mixture models in its most general form. The central algorithmic ideas consist of new ways to decompose the moment tensor of the Gaussian mixture model by exploiting its structural prop-

erties. The symmetries of this tensor are derived from the combinatorial structure of higher order moments of Gaussian distributions (sometimes referred to as Isserlis' theorem or Wick's theorem). We show that, provided we are in high enough dimension n for $n \geq \Omega(k^2)$, the class of Gaussian mixtures is learnable with polynomial running time and using polynomial number of samples, under a smoothed analysis framework. The key of this analysis framework is that we study how the proposed algorithm performs on an instance with randomly perturbed parameters starting from *any point*. This serves to bridge the gap between worst case analysis which analyze the performance on a worst case instance chosen adversarially, and average case analysis which analyze the performance on a predefined distribution of instances.

Theorem 1.2. *In the smoothed analysis setting, when $n \geq \Omega(k^2)$, given samples from the perturbed n -dimensional Gaussian mixture model with k components, there is an algorithm that learns the correct parameters up to accuracy ϵ with high probability, using polynomial time and polynomial number of samples.*

In Chapter 4, we shift attention to stationary Hidden Markov models, which can be viewed as special a case of a mixture model. We study the minimal realization problems of HMMs. In particular, given access to length N segments of observable outputs supported over a size d alphabet, which are generated by a Hidden Markov model of order k , we can compute the joint probabilities of segments of the observables. The main questions we attempt to dress is that, given such joint probabilities of segments, how to construct a model of minimal order that can generate the same output process, and how large is the required window size N .

Despite the known worst case construction where N is lower bounded by $\Omega(k)$ and the worst case computational complexity scales exponentially in k , we use generic analysis and show that all the hard cases lie in a measure zero set in the parameter space. Moreover, for all the non-degenerate instances, the required window size is only in the order of $O(\log_d(k))$. In other words, the minimal HMM realization problem actually can be solved with polynomial complexity for almost all cases.

1.2.3 Part 3: Randomized Algorithm to Escape Worst Cases

In Chapter 5, we study the problem of super-resolution and explore the randomness in the learning algorithm to circumvent the worst case lower bounds.

Super-resolution is the problem of recovering a superposition of point sources using bandlimited measurements, which may be corrupted with noise. To view it in the Fourier domain, the learning task is to disentangle a mixture of noisy complex sinusoids. This signal processing problem arises in numerous imaging problems, ranging from astronomy to biology to spectroscopy, where it is common to take coarse Fourier measurements of an object. Of particular interest is in obtaining estimation procedures which are robust to noise, with the following desirable statistical and computational properties: we seek to use coarse Fourier measurements (bounded by some *cutoff frequency*); we hope to take a small number of measurements; we desire our algorithm to run quickly.

Suppose we have k point sources in d dimensions, where the points are separated by at least Δ from each other. We provide a randomized algorithm that uses Fourier measurements at random frequencies, as opposed to taking an exponential number of measurements on the hyper-grid in the previous algorithms. We show that the required bandwidth of frequencies is bounded by $\tilde{O}(1/\Delta)$, while previous algorithms require a cutoff frequency as large as $\Omega(\sqrt{d}/\Delta)$. Moreover, the number of measurements taken by and the computational complexity of our algorithm are bounded by a polynomial in both the number of points k and the dimension d , with *no* dependence on the separation Δ . In contrast, previous algorithms depended inverse polynomially on the minimal separation and exponentially on the dimension for both of these quantities.

Theorem 1.3. *For a fixed probability of error, the proposed algorithm achieves stable recovery with a number of measurements and with computational runtime that are both on the order of $\tilde{O}((k + d)^2)$. Furthermore, the algorithm makes measurements which are bounded in frequency by $\tilde{O}(1/\Delta)$ (ignoring log factors).*

1.3 Preliminaries

1.3.1 Notations

We utilize the standard $O(\cdot)$, $\Theta(\cdot)$, $\Omega(\cdot)$ notation to hide constants, and $\tilde{O}(\cdot)$, $\tilde{\Theta}(\cdot)$, $\tilde{\Omega}(\cdot)$ to hide constants and logarithmic factors. We use \mathbb{R} , \mathbb{C} , and \mathbb{Z} to denote real, complex, and natural numbers. For $d \in \mathbb{Z}$, we use $[d]$ to denote the set $[d] = \{1, \dots, d\}$. For a set \mathcal{S} , $|\mathcal{S}|$ to denote its cardinality. We use \oplus to denote the direct sum of sets, namely $\mathcal{S}_1 \oplus \mathcal{S}_2 = \{(a + b) : a \in \mathcal{S}_1, b \in \mathcal{S}_2\}$.

Vectors and Matrices In the vector space \mathbb{R}^n , let $\langle \cdot, \cdot \rangle$ denote the inner product of two vectors, and $\|\cdot\|$ to denote the Euclidean norm. Let e_i to denote the i -th standard basis vector in \mathbb{R}^n , for $i \in [n]$.

Let I_n be the identity matrix of dimension $n \times n$. For a tall matrix $A \in \mathbb{R}^{m \times n}$, let $A_{[:,j]}$ denote its j -th column vector, let A^\top denote its transpose, $A^\dagger = (A^\top A)^{-1} A^\top$ denote the pseudoinverse. Let $\sigma_k(A)$ denote its k -th singular value. Define the condition number of a matrix $X \in \mathbb{R}^{m \times n}$ to be $\text{cond}_2(X) = \sigma_{\max}(X)/\sigma_{\min}(X)$, where $\sigma_{\max}(X)$ and $\sigma_{\min}(X)$ are the maximal and minimal singular values of X . The spectral norm of a matrix is denoted as $\|A\|$, and the Frobenius norm is denoted as $\|A\|_F$. We use $A \succeq 0$ for positive semidefinite matrix A .

When converting between vectors and matrices, let $\text{vec}(A) \in \mathbb{R}^{mn}$ denote the vector obtained by stacking all the columns of A . For a vector $x \in \mathbb{R}^{m^2}$, let $\text{mat}(x) \in \mathbb{R}^{m \times m}$ denote the inverse mapping such that $\text{vec}(\text{mat}(x)) = x$.

Linear subspaces We represent a linear subspace $\mathcal{S} \in \mathbb{R}^n$ of dimension d by a matrix $S \in \mathbb{R}^{n \times d}$, whose columns of S form an (arbitrary) orthonormal basis of the subspace. The projection matrix onto the subspace \mathcal{S} is denoted by $\text{Proj}_{\mathcal{S}} = SS^\top$, and the projection onto the orthogonal subspace \mathcal{S}^\perp is denoted by $\text{Proj}_{\mathcal{S}^\perp} = I_n - SS^\top$. When we talk about the linear span of several matrices, we mean the space spanned by their vectorization.

Matrix Products We use \otimes to denote tensor product, \odot to denote column wise Katri-Rao product, and \otimes_{kr} to denote Kronecker product. As an example, for matrices $A \in \mathbb{R}^{m_A \times n}$, $B \in \mathbb{R}^{m_B \times n}$, $C \in \mathbb{R}^{m_C \times n}$:

$$A \otimes B \otimes C \in \mathbb{R}^{m_A \times m_B \times m_C}, \quad [A \otimes B \otimes C]_{j_1, j_2, j_3} = \sum_{i=1}^n A_{j_1, i} B_{j_2, i} C_{j_3, i},$$

$$A \otimes_{kr} B \in \mathbb{R}^{m_A m_B \times n^2}, \quad A \otimes_{kr} B = \begin{bmatrix} A_{1,1}B & \cdots & A_{1,n}B \\ \vdots & \ddots & \vdots \\ A_{m_A,1}B & \cdots & A_{m_A,n}B \end{bmatrix},$$

$$A \odot B \in \mathbb{R}^{m_A m_B \times n}, \quad [A \odot B]_{[:,j]} = A_{[:,j]} \otimes_{kr} B_{[:,j]}.$$

Symmetric matrices We use $\mathbb{R}_{sym}^{n \times n}$ to denote the space of all $n \times n$ symmetric matrices, which linear subspace has dimension $\binom{n+1}{2}$. Since we will frequently use $n \times n$ and $k \times k$ symmetric matrices, we denote their dimensions by the constants $n_2 = \binom{n+1}{2}$ and $k_2 = \binom{k+1}{2}$. Similarly, we use $\mathbb{R}_{sym}^{n \times \dots \times n}$ to denote the symmetric k -dimensional multi-arrays (tensors), which subspace has dimension $\binom{n+k-1}{k}$. If a k -th order tensor $X \in \mathbb{R}_{sym}^{n \times \dots \times n}$, then for any permutation π over $[k]$, we have $X_{n_1, \dots, n_k} = X_{n_{\pi(1)}, \dots, n_{\pi(k)}}$.

1.3.2 Tensor Algebra

Our learning algorithms are mostly based on spectral methods for moment matching, which involve spectral factorization of matrices and tensors in different steps. Next, we introduce some basics of tensor algebra. A more detailed introduction to tensor algebra can be found in [70] and the references therein.

Definitions A tensor is a multi-dimensional array. Tensor notations are useful for handling higher order moments. Consider vectors $a, b, c \in \mathbb{R}^n$, define $T = a \otimes b \otimes c \in \mathbb{R}^{n \times n \times n}$ to be the rank one tensor such that $T_{i_1, i_2, i_3} = a_{i_1} b_{i_2} c_{i_3}$. For a vector $x \in \mathbb{R}^n$, let the t -fold tensor product $x \otimes^t$ denote the t -th order rank one tensor $(x \otimes^t)_{i_1, i_2, \dots, i_t} = \prod_{j=1}^t x_{i_j}$. We write the tensor product of matrices as $A \otimes B \otimes C = \sum_{i=1}^k A_{[:,i]} \otimes B_{[:,i]} \otimes C_{[:,i]}$.

Every tensor also defines a multilinear mapping. Consider a 3-rd order tensor $X \in \mathbb{R}^{n_A \times n_B \times n_C}$. For given dimension m_A, m_B, m_C , it defines a multi-linear mapping $X(\cdot, \cdot, \cdot) : \mathbb{R}^{n_A \times m_A} \times \mathbb{R}^{n_B \times m_B} \times \mathbb{R}^{n_C \times m_C} \rightarrow \mathbb{R}^{m_A \times m_B \times m_C}$ defined as below: ($\forall j_1 \in [m_A], j_2 \in [m_B], j_3 \in [m_C]$)

$$[X(V_1, V_2, V_3)]_{j_1, j_2, j_3} = \sum_{i_1 \in [n_A], i_2 \in [n_B], i_3 \in [n_C]} X_{i_1, i_2, i_3} [V_1]_{j_1, i_1} [V_2]_{j_2, i_2} [V_3]_{j_3, i_3}. \quad (1.1)$$

If X admits a decomposition $X = \sum_{i=1}^k A_{[:,i]} \otimes B_{[:,i]} \otimes C_{[:,i]}$ for $A \in \mathbb{R}^{n_A \times k}, B \in \mathbb{R}^{n_B \times k}, C \in \mathbb{R}^{n_C \times k}$, the multi-linear mapping has the form

$$X(V_1, V_2, V_3) = \sum_{i=1}^k (V_1^\top A_{[:,i]}) \otimes (V_2^\top B_{[:,i]}) \otimes (V_3^\top C_{[:,i]}).$$

In particular, the vector given by $X(\mathbf{e}_i, \mathbf{e}_j, I)$ is the one-dimensional slice of the 3-way array, with the index for the first dimension to be i and the second dimension to be j . Note that X can have different forms of decompositions, yet the mappings defined in (1.1) are all equivalent.

Definition 1.1 (Khatri-Rao product). *For matrices $A \in \mathbb{R}^{n_A \times k}, B \in \mathbb{R}^{n_B \times k}$, the (column) Khatri-Rao product $X = A \odot B \in \mathbb{R}^{n_A n_B \times k}$ is defined as follows:*

$$X_{(j_1-1)n_B+j_2, i} = A_{j_1, i} B_{j_2, i},$$

and each column of X is a rank-1 Khatri-Rao product.

An equivalent representation of a 3rd order tensor $X \in \mathbb{R}^{n_A \times n_B \times n_C}$ is its matricization, obtained by rearranging the elements of the tensor into a matrix. For example, the matricization along the third mode gives a matrix $\overline{X}^{(3)}$ is specified as: $[\overline{X}^{(3)}]_{j_3, ((j_1-1)n_B+j_2)} = X_{j_1, j_2, j_3}$. Moreover, if the tensor admits a decomposition $X = A \otimes B \otimes C$, we can write the matricization as Khatri-Rao product of the factors: $\overline{X}^{(3)} = C(A \odot B)^\top$.

Uniqueness of tensor rank decomposition Tensor algebra has many similarities to but also many striking differences from matrix algebra. For example, tensor rank decomposition is a natural extension of matrix *rank decomposition* to higher order tensors. However, under very mild conditions, *rank decomposition* of a tensor is unique up to column scaling and permutation, which is the key property we will exploit to consistently estimate the model parameters for various mixture models. This is in sharp contrast to the matrix minimal rank factorization, where if $A = BC$ is a minimal rank k factorization, we can write $A = (BW)(W^{-1}C)$ for any full rank matrix W of size $k \times k$.

Definition 1.2 (Tensor rank decomposition). *The rank decomposition of a 3rd order tensor $X \in \mathbb{R}^{n_A \times n_B \times n_C}$ is a sum of rank-1 tensors for the smallest number of summands k :*

$$X = A \otimes B \otimes C = \sum_{i=1}^k A_{[:,i]} \otimes B_{[:,i]} \otimes C_{[:,i]},$$

where matrices $A \in \mathbb{R}^{n_A \times k}$, $B \in \mathbb{R}^{n_B \times k}$, $C \in \mathbb{R}^{n_C \times k}$. The minimal number of summands k is defined to be the rank of the tensor.

In the following, we state a set of sufficient conditions on the factors A, B, C that guarantee the uniqueness of a third order tensor decomposition $X = A \otimes B \otimes C$.

Definition 1.3 (Kruskal rank). *The Kruskal rank of a matrix $A \in \mathbb{R}^{n \times m}$ equals r if any set of r columns of A are linearly independent, and there exists a set of $(r + 1)$ columns that are linearly dependent (if $r < m$).*

Lemma 1.1 (Uniqueness of tensor decomposition ([72, 105])). *The tensor factorization $X = A \otimes B \otimes C$ is unique up to column permutation and scaling, if*

$$krank(A) + krank(B) + krank(C) \geq 2k + 2. \quad (1.2)$$

Tensor decomposition algorithms Unlike matrix singular value decomposition (SVD), in general, even if the tensor rank decomposition $X = A \otimes B \otimes C$ is unique,

finding the factors A, B, C given the rank k tensor X is a hard problem. Nevertheless, for cases where the factors satisfy certain rank conditions, there exist efficient and provable algorithms to find the unique factorization. First, if both the matrix A and B have full column rank, Algorithms 1 ([77]) can uniquely recover the factors up to common column permutation, with running time polynomial in the dimension of the tensor. Other algorithms such as tensor power method and recursive projection, which are possibly more stable in practice, also apply in this setting. Second, FOABI ([42] [63]) is another tensor decomposition algorithm that has polynomial runtime, and it applies to a subset of instances even when A and B are not of full column rank k . Instead, for this algorithm to work, it requires that the factor C and the Khatri-Rao product $A \odot B$ have full column rank k . For completeness, we list two standard tensor decomposition algorithms below.

Algorithm 1: Simultaneous diagonalization for 3rd order tensor decomposition [77]

Input: A 3rd order tensor $M \in \mathbb{R}^{d^n \times d^n \times d}$

Output: $k, A, B \in \mathbb{R}^{d^n \times k}, C \in \mathbb{R}^{d \times k}$

1. Randomly pick two unit norm vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^d$. Project M along the 3rd dimension to obtain two matrices:

$$\widetilde{M}_1 = M(I, I, \mathbf{v}_1), \quad \widetilde{M}_2 = M(I, I, \mathbf{v}_2).$$

2. Compute the eigen-decomposition of matrix $(\widetilde{M}_1 \widetilde{M}_2^{-1})$ and $(\widetilde{M}_2 \widetilde{M}_1^{-1})$, and let the columns of matrix A and B be the eigenvectors of $(\widetilde{M}_1 \widetilde{M}_2^{-1})$ and $(\widetilde{M}_2 \widetilde{M}_1^{-1})$, respectively.

Scale the columns of A and B to be stochastic, and pair the eigenvectors in A and B corresponding to the reciprocal eigenvalues, namely:

$$\widetilde{M}_1 \widetilde{M}_2^{-1} = A \Lambda A^{-1}, \quad \widetilde{M}_2 \widetilde{M}_1^{-1} = B \Lambda^{-1} B^{-1}.$$

3. Let k be the number of non-zero eigenvalues.
4. Let $\overline{M}^{(3)} \in \mathbb{R}^{d^{2n} \times d}$ be the 3rd dimension matricization of M . Set C to be:

$$C = \overline{M}^{(3)} ((A \odot B)^\dagger)^\top$$

Algorithm 2: FOOBI for 3rd order tensor decomposition

Input: Three way tensor M .

Output: Rank k and factors A, B, C .

1. Let $\overline{M}^{(3)}$ be the 3rd dimension matricization of M . Compute its SVD
 $\overline{M}^{(3)} = V_H D_H U_H^\top$.
2. Set k to be the number of non-zero singular values. Let $F = V_H D_H^{1/2}$, and
 $E = U_H D_H^{1/2}$.
3. Construct matrices $\{E^{(r)} \in \mathbb{R}^{d \times d} : r \in [k]\}$:

$$[E^{(r)}]_{i,j} = E_{(i-1)d+j,r}, \forall i, j \in [d], \forall r \in [k].$$

Construct the 4-th order tensors $\{P^{(r,s)} \in \mathbb{R}^{d \times d \times d \times d} : r, s \in [k]\}$:

$$\begin{aligned} & [P^{(r,s)}]_{i_1, i_2, j_1, j_2} \\ &= [E^{(r)}]_{i_1, j_1} [E^{(s)}]_{i_2, j_2} + [E^{(s)}]_{i_1, j_1} [E^{(r)}]_{i_2, j_2} \\ & - [E^{(r)}]_{i_1, j_2} [E^{(s)}]_{i_2, j_1} - [E^{(s)}]_{i_1, j_2} [E^{(r)}]_{i_2, j_1}. \end{aligned}$$

4. Compute a basis $\{H^{(i)} : i \in [k]\}$ of the k -dimensional kernel of
 $\{P^{(r,s)} : r, s \in [k]\}$:

$$\sum_{r,s=1}^k H_{r,s}^{(i)} P^{(r,s)} = 0, \quad \text{s.t. } H_{r,s}^{(i)} = H_{s,r}^{(i)}, \forall r, s \in [k].$$

5. Find the unique $W \in \mathbb{R}^{k \times k}$ that simultaneously diagonalizes the basis:
 $H^{(i)} = W \Lambda^{(i)} W^\top$.
 6. Let $C = F(W^{-1})^\top$ and $A \odot B = EW$. Compute the rank one decomposition of each column of $A \odot B$, with proper normalization such that A and B are column stochastic.
-

1.3.3 Probabilistic analysis

In the rest of this chapter, we review some standard results of matrix perturbation and concentration inequalities, and prove some corollaries. These will come in handy for our algorithm analysis in later chapters.

Matrix perturbation bounds Many spectral algorithms involve matrix decomposition in different forms, thus characterizing the sample complexity of the learning algorithm boils down to analyzing the stability of the matrix decompositions. Given a matrix \widehat{A} and we know that $\widehat{A} = A + E$ where E is a perturbation matrix of small magnitude, how does the singular values and singular vectors of \widehat{A} relate to that of A ? This is a well-studied matrix perturbation problem and many results can be found in Stewart and Sun [108].

Theorem 1.4 (Weyl's theorem). *Given $\widehat{A} = A + E$, we know that*

$$\sigma_k(A) - \|E\| \leq \sigma_k(\widehat{A}) \leq \sigma_k(A) + \|E\|.$$

We can also bound the ℓ_2 norm change in singular values by Mirsky's Theorem.

Lemma 1.2 (Mirsky's theorem). *Given matrices $A, E \in \mathbb{R}^{m \times n}$ with $m \geq n$, then*

$$\sqrt{\sum_{i=1}^n (\sigma_i(A + E) - \sigma_i(A))^2} \leq \|E\|_F.$$

For singular vectors, the perturbation is bounded by Wedin's Theorem:

Lemma 1.3 (Wedin's theorem; Theorem 4.1, p.260 in [109]). *Given matrices $A, E \in \mathbb{R}^{m \times n}$ with $m \geq n$. Let A have the singular value decomposition*

$$A = [U_1, U_2, U_3] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{bmatrix} [V_1, V_2]^T.$$

Let $\widehat{A} = A + E$, with analogous singular value decomposition. Let Φ be the matrix of canonical angles between the column span of U_1 and that of \widehat{U}_1 , and Θ be the matrix of canonical angles between the column span of V_1 and that of \widehat{V}_1 . Suppose that there exists a δ such that $\min_{i,j} |[\Sigma_1]_{i,i} - [\Sigma_2]_{j,j}| > \delta$, and $\min_{i,i} |[\Sigma_1]_{i,i}| > \delta$, then

$$\|\sin \Phi\|^2 + \|\sin \Theta\|^2 \leq 2 \frac{\|E\|^2}{\delta^2}.$$

We do not go into the detailed definitions of canonical angles here. The only way we will be using this lemma is by combining it with the following lemma:

Lemma 1.4 (Theorem 4.5, p.92 in [109]). *Let Φ be the matrix of canonical angles between the column span of U and that of \widehat{U} , then*

$$\|Proj_{\widehat{U}} - Proj_U\| = \|\sin \Phi\|.$$

As a corollary, we have:

Lemma 1.5. *Given matrices $A, E \in \mathbb{R}^{m \times n}$ with $m \geq n$. Suppose that A has rank k and the smallest singular value is given by $\sigma_k(A)$. Let S and \widehat{S} be the subspaces spanned by the first k eigenvectors of A and $\widehat{A} = A + E$, respectively. Then if $\|E\| \leq \sigma_k(A)/\sqrt{2}$, we have:*

$$\|\widehat{S} - S\| \leq \|Proj_{\widehat{S}} - Proj_S\| = \|Proj_{\widehat{S}^\perp} - Proj_{S^\perp}\| \leq \frac{\sqrt{2}\|E\|}{\sigma_k(A)}.$$

Proof. We first prove the first inequality:

$$\begin{aligned} \|Proj_{\widehat{S}} - Proj_S\| &= \|2S(\widehat{S} - S)^\top + (\widehat{S} - S)(\widehat{S} - S)^\top\| \\ &\geq 2\|S\|\|\widehat{S} - S\| - \|\widehat{S} - S\|^2 \\ &\geq \|S\|\|\widehat{S} - S\| = \|\widehat{S} - S\|. \end{aligned}$$

The equality is because $Proj_{S^\perp} = I - Proj_S$ so the two differences are the same. The final step follows from Wedin's Theorem and Lemma 1.4. \square

We often need to bound the perturbation of a product of perturbed matrices, where we apply the following lemma:

Lemma 1.6. *Consider a product of matrices $A_1 \cdots A_k$, and consider any sub-multiplicative norm on matrix $\|\cdot\|$. Given $\widehat{A}_1, \dots, \widehat{A}_k$ and assume that $\|\widehat{A}_i - A_i\| \leq \|A_i\|$, then we*

have:

$$\|\widehat{A}_1 \cdots \widehat{A}_k - A_1 \cdots A_k\| \leq 2^{k-1} \prod_{i=1}^k \|A_i\| \sum_{i=1}^k \frac{\|\widehat{A}_i - A_i\|}{\|A_i\|}.$$

The proof of this lemma is straightforward by induction.

Next theorem bounds the perturbation on the pseudo-inverse of a matrix, provided that the smallest singular value of the matrix is lower bounded.

Theorem 1.5 (Theorem 3.4 in [108]). *Consider the perturbation of a matrix $A \in \mathbb{R}^{m \times n}$: $B = A + E$. Assume that $\text{rank}(A) = \text{rank}(B) = n$, then*

$$\|B^\dagger - A^\dagger\| \leq \sqrt{2} \|A^\dagger\| \|B^\dagger\| \|E\|.$$

As a corollary, we often use:

Lemma 1.7. *Consider the perturbation of a matrix $A \in \mathbb{R}^{m \times n}$: $B = A + E$ where $\|E\| \leq \sigma_{\min}(A)/2$. Assume that $\text{rank}(A) = \text{rank}(B) = n$, then*

$$\|B^\dagger - A^\dagger\| \leq 2\sqrt{2} \|E\| / \sigma_{\min}(A)^2.$$

Proof. We first apply Theorem 1.5, and then bound $\|A^\dagger\|$ and $\|B^\dagger\|$. By definition we know $\|A^\dagger\| = 1/\sigma_{\min}(A)$. By Weyl's theorem $\sigma_{\min}(B) \geq \sigma_{\min}(A) - \|E\| \geq \sigma_{\min}(A)/2$, hence $\|B^\dagger\| = \sigma_{\min}(B)^{-1} \leq 2\sigma_{\min}(A)^{-1}$. \square

Concentration inequalities In our probabilistic analysis, we usually need to characterize the behavior of a random variable that depend on a large number of independent random variables, in particular, how much it deviates from its expected value as the number of independent random variables increases. Next we review some concentration inequalities that bound such deviation.

Lemma 1.8 (Hoeffding's inequality). *Let X_1, \dots, X_n be independent random variables. Assume that X_i 's are bounded almost surely, namely $\Pr[X_i \in [a_i, b_i]] = 1$.*

Define the empirical mean of these variables $\bar{X} = (X_1 + \dots + X_n)/n$. We have

$$\Pr[|\bar{X} - \mathbb{E}[\bar{X}]| \geq t] \leq \exp\left(-\frac{2n^2t}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma 1.9 (Multiplicative Chernoff bound). *Suppose X_1, \dots, X_n are independent random variables with Bernoulli distribution, and $\mathbb{P}(X_i = 1) = \mu$. Then for any $\delta > 1$:*

$$\mathbb{P}\left(\sum_{i=1}^n X_i > \delta n\mu\right) < \left(\frac{e}{\delta}\right)^{\delta n\mu}.$$

Lemma 1.10 (Matrix Bernstein). *Consider a sequence of N random matrix $\{X_k\}$ of dimension $M \times M$ which are independent, self-adjoint. Assume that $\mathbb{E}[X_k] = 0$ and $\lambda_{\max}(X_k) \leq R$ almost surely. Denote the total variance by $\sigma^2 = \|\sum_{k=1}^N \mathbb{E}[X_k^2]\|$. Then the following inequality holds for all $t > 0$:*

$$\Pr\left(\left\|\sum_{k=1}^N X_k\right\| \geq t\right) \leq Me^{-\frac{t^2}{\sigma^2 + Rt/3}} \leq \begin{cases} Me^{-\frac{3t^2}{8\sigma^2}}, & \text{for } t \leq \sigma^2/R; \\ Me^{-\frac{3t}{8R}}, & \text{for } t \geq \sigma^2/R. \end{cases}$$

Lemma 1.11 (High dimensional sphere projection (Johnson Lindenstrauss lemma)). *Let the random vector $\mathbf{u} \in \mathbb{R}^d$ be uniformly distributed on the surface of the d -dimensional unit sphere, i.e. uniform distribution in the set: $\{\sum_{i=1}^d u_i^2 = 1\}$. Denote its projection onto the first dimension to be $|u_1|$. We have:*

$$\mathbb{P}(|u_1| > t) < \frac{4}{\sqrt{d-2}} e^{-\frac{d-2}{2}t^2}.$$

Theorem 1.6 (Gershgorin's theorem). *Given a symmetric matrix $X \in \mathbb{R}^{k \times k}$, a lower bound on the smallest eigenvalue is given by:*

$$\sigma_{\min}(X) \geq \min_{i \in [k]} \left\{ X_{i,i} - \sum_{j \in [k], j \neq i} |X_{i,j}| \right\}.$$

Chapter 2

Recovering structured matrices

2.1 Problem Statement

Consider an unknown $M \times M$ probability matrix \mathbb{B} , satisfying $\mathbb{B}_{i,j} \geq 0$ and $\sum_{i,j} \mathbb{B}_{i,j} = 1$. Suppose one is given N independently drawn (i, j) -pairs, sampled according to the distribution defined by \mathbb{B} . How many draws are necessary to accurately recover \mathbb{B} ? What can one infer about the underlying matrix based on these samples? How can one accurately test whether the underlying matrix possesses certain properties of interest? How do structural assumptions on \mathbb{B} — for example, the assumption that \mathbb{B} has low rank — affect the information theoretic or computational complexity of these questions? For the majority of these tasks, we currently lack both a basic understanding of the computational and information theoretic lay of the land, as well as algorithms that seem capable of achieving the information theoretic or computational limits.

This general question of making accurate inferences about a matrix of probabilities, given a matrix of observed “counts” of discrete outcomes, lies at the core of a number of problems that disparate communities have been tackling independently. On the theoretical side, these problems include both work on community detection in stochastic block models (where the goal is to infer the community memberships from an adjacency matrix of a graph that has been drawn according to an underlying matrix of probabilities expressing the community structure) as well as the line of work on

recovering topic models, hidden Markov models (HMMs), and richer structured probabilistic models (where the model parameters can often be recovered using observed count data). On the practical side, these problems include work on computing low-rank approximations to sparsely sampled data, which arise in collaborative filtering and recommendation systems, as well as the recent work from the natural language processing community on understanding matrices of word co-occurrence counts for the purpose of constructing good “word embeddings”. Additionally, work on latent semantic analysis and non-negative matrix factorization can also be recast in this setting.

In this part, we start this line of inquiry by focusing on the estimation problem where the probability matrix \mathbb{B} possesses a particular low rank structure. While this estimation problem is rather specific, it generalizes the basic community detection problem and also encompasses the underlying problem behind learning HMMs and topic models. Furthermore, this low rank case also provides a means to study how property testing and estimation problems are different in this structured setting, as opposed to the simpler rank 1 setting that is equivalent to the standard setting of independent draws from a distribution supported on M elements.

We focus on the estimation of a low rank probability matrix \mathbb{B} in the sparse data regime, near the information theoretic limit. In many practical scenarios involving sample counts, we seek algorithms capable of extracting the underlying structure in the sparsely sampled regime. To give two motivating examples, consider forming the matrix of word co-occurrences—the matrix whose rows and columns are indexed by the set of words, and whose (i, j) -th element consists of the number of times the i -th word follows the j -th word in a large corpus of text. In the context of recommendation system, one could consider a low rank matrix model, where the rows are indexed by customers, and the columns are indexed by products, with the (i, j) -th entry corresponding to the number of times the i -th customer has purchased the j -th product. In both settings, the structure of the probability matrix underlying these observed counts contains insights into the two domains, and in both domains we only have relatively sparse data. This is inherent in many other natural scenarios

involving heavy-tailed distributions where, regardless of how much data one collects, a significant fraction of items (e.g. words, products purchased, genetic mutations, etc.) will only be observed a few times.

Such estimation questions have been actively studied in the community detection literature, where the objective is to accurately recover the communities in the regime where the average degree (e.g. the row sums of the adjacency matrix) are constant. In contrast, the recent line of works for recovering highly structured models (such as topic models, HMMs, etc.) are only applicable to the *over-sampled* regime where the amount of data is well beyond the information theoretic limits. In these cases, achieving the information theoretic limits remains a widely open question. This work begins to bridge the divide between these recent algorithmic advances in both communities. We hope that the low rank probability matrix setting that studied here serves as a jumping-off point for the more general questions of developing information theoretically optimal algorithms for estimating structured matrices and tensors in general, or recovering low-rank approximations to arbitrary probability matrices, in the sparse data regime. While the general settings are more challenging, we believe that some of our algorithmic techniques can be fruitfully extended.

In addition to developing algorithmic tools which we hope are applicable to a wider class of problems, a second motivation for considering this particular low rank case is that, with respect to distribution learning and property testing, the entire lay-of-the-land seems to change completely when the probability matrix \mathbb{B} has rank larger than 1. In the rank 1 setting — where a sample consists of 2 *independent* draws from a distribution supported on $\{1, \dots, M\}$ — the distribution can be learned using $\Theta(M)$ draws. Nevertheless, many properties of interest can be tested or estimated using a sample size that is *sublinear* in M^1 . However, even just in the case where the probability matrix is of rank 2, although the underlying matrix \mathbb{B} can be represented with $O(M)$ parameters (and, as we show, it can also be accurately and efficiently re-

¹Distinguishing whether a distribution is uniform versus far from uniform can be accomplished using only $O(\sqrt{M})$ draws, testing whether two sets of samples were drawn from similar distributions can be done with $O(M^{2/3})$ draws, estimating the entropy of the distribution to within an additive ϵ can be done with $O(\frac{M}{\epsilon \log M})$ draws, etc.

covered with $O(M)$ sample counts), sublinear sample property testing and estimation is generally impossible. This result begs a more general question: *what conditions must be true of a structured statistical setting in order for property testing to be easier than learning?*

2.1.1 Formulation

Assume our vocabulary is the index set $\mathcal{M} = \{1, \dots, M\}$ of M words and that there is an underlying low rank probability matrix \mathbb{B} , of size $M \times M$, with the following structure:

$$\mathbb{B} = PWP^\top, \text{ where matrix } P = [p^{(1)}, \dots, p^{(R)}]. \quad (2.1)$$

Here the matrix P is of dimension $M \times R$, and the columns are supported on the standard $(M - 1)$ -simplex. Also, $W \in \mathbb{R}_+^{R \times R}$ is the *mixing matrix*, which is a probability matrix satisfying $\sum_{i,j} W_{i,j} = 1$.

In the case where $R = 2$, we denote $w_p = W_{1,1} + W_{1,2}$ and $w_q = W_{2,1} + W_{2,2}$. Note that $\sum_k \mathbb{B}_{i,k} = w_p p + w_q q$. Define the *covariance matrix* of any probability matrix P as:

$$[\text{Cov}(P)]_{i,j} := P_{i,j} - \left(\sum_k P_{i,k}\right)\left(\sum_k P_{k,j}\right).$$

Note that $\text{Cov}(P)\vec{1} = \vec{0}$ and $\vec{1}^\top \text{Cov}(P) = \vec{0}$ (where $\vec{1}$ and $\vec{0}$ are the all ones and zeros vectors, respectively). This implies that, without loss of generality, the covariance of the mixing matrix, $\text{Cov}(W)$, can be expressed as: $\text{Cov}(W) = [w_L, -w_L]^\top [w_R, -w_R]$ for some real numbers $w_L, w_R \in [-1, 1]$. For ease of exposition, we restrict to the symmetric case where $w_L = w_R = w$, though our results hold more generally.

Suppose we obtain N , i.i.d. sample counts from \mathbb{B} of the form $\{(i_1, j_1) (i_2, j_2), \dots (i_N, j_N)\}$, where each sample $(i_n, j_n) \in \mathcal{M} \times \mathcal{M}$. The probability of obtaining a count (i, j) in a sample is $\mathbb{B}_{i,j}$. Moreover, assume that the number of samples follows a Poisson distribution: $N \sim \text{Poi}(\mathbb{N})$. The Poisson assumption on the number of samples is made only for the convenience of analysis: so that the counts of observing (i, j) follows

a Poisson distribution $\text{Poi}(\mathbb{N}\mathbb{B}_{i,j})$ and is independent from the counts of observing (i', j') for $(i', j') \neq (i, j)$. As M is asymptotically large, with high probability, N and \mathbb{N} are within a subconstant factor of each other and both upper and lower bounds translate between the Poissonized setting, and the setting of fixed N . Throughout, our sample complexity results are stated in terms of N .

Notation We use the following standard shorthand notations throughout this chapter. We denote $[n] \triangleq \{1, \dots, n\}$. Let \mathcal{I} denote a subset of indices in \mathcal{M} . For a M -dimensional vector x , we use vector $x_{\mathcal{I}}$ to denote the elements of x restricting to the indices in \mathcal{I} ; for two index sets \mathcal{I}, \mathcal{J} , and a $M \times M$ dimensional matrix X , we use $X_{\mathcal{I} \times \mathcal{J}}$ denote the submatrix of X with rows restricting to indices in \mathcal{I} and columns restricting to indices in \mathcal{J} .

We use $\text{Poi}(\lambda)$ to denote a Poisson distribution with rate λ ; we use $\text{Ber}(\lambda)$ to denote a Bernoulli random variable with success probability λ ; and we use $\text{Mul}(x; \lambda)$ to denote a multinomial distribution over M outcomes with λ number of trials and event probability vector $x \in \mathbb{R}_+^M$ such that $\sum_i x_i = 1$.

All of our order notations are with respect to the vocabulary size M , which is asymptotically large. Also, we say that a statement is true “with high probability” if the failure probability of the statement is inverse poly in M ; and we say a statement is true “with large probability” if the failure probability is of some small constant δ , which can be easily boosted via repetition.

2.1.2 Related Work

As mentioned earlier, the general problem of reconstructing an underlying matrix of probabilities given access to a count matrix drawn according to the corresponding distribution, lies at the core of questions that are being actively pursued by several different communities. We briefly describe these questions, and their relation to the present work.

Community Detection. With the increasing prevalence of large scale social networks, there has been a flurry of activity from the algorithms and probability com-

munities to both model structured random graphs, and understand how (and when it is possible) to examine a graph and infer the underlying structures that might have given rise to the observed graph. One of the most well studied community models is the *stochastic block model* [56]. In its most basic form, this model is parameterized by a number of individuals, M , and two probabilities, α, β . The model posits that the M individuals are divided into two equal-sized “communities”, and such a partition defines the following random graph model: for each pair of individuals in the same community, the edge between them is present with probability α (independently of all other edges); for a pair of individuals in different communities, the edge between them is present with probability $\beta < \alpha$. Phrased in the notation of our setting, the adjacency matrix of the graph is generated by including each potential edge (i, j) independently, with probability $\mathbb{B}_{i,j}$, with $\mathbb{B}_{i,j} = \alpha$ or β according to whether i and j are in the same community. Note that \mathbb{B} has rank 2 and is expressible in the form of Equation 2.1 as $\mathbb{B} = PWP^\top$ where $P = [p, q]$ for vectors $p = \frac{2}{M}I_1$ and $q = \frac{2}{M}I_2$ where I_1 is the indicator vector for membership in the first community, and I_2 is defined analogously, and W is the 2×2 matrix with $\alpha \frac{M^2}{4}$ on the diagonal and $\beta \frac{M^2}{4}$ on the off-diagonal.

What values of α, β , and M enable the community affiliations of all individuals to be accurately recovered with high probability? What values of α, β , and M allow for the graph to be distinguished from an Erdos-Renyi random graph (that has no community structure)? The crucial regime is where $\alpha, \beta = O(\frac{1}{M})$, and hence each person has a constant, or logarithmic expected degree. The naive spectral approaches will fail in this regime, as there will likely be at least one node with degree $\approx \log M / \log \log M$, which will ruin the top eigenvector. Nevertheless, in a sequence of works sparked by the paper of Friedman, and Szemerédi [47], the following punchline has emerged: the naive spectral approach will work, even in the constant expected degree setting, provided one first either removes, or at least diminishes the weight of these high-degree problem vertices (e.g. [44, 68, 87, 73, 75]). In the past year, for both the *exact* recovery problem and the detection problem, the exact tradeoffs between α, β , and M were established, down to subconstant factors [88, 1, 81]. More recently, there has

been further research investigating more complex stochastic block models, consisting of three or more components, components of unequal sizes, etc. (see e.g. [36, 2]).

Word Embeddings. On the more applied side, some of the most impactful advances in natural language processing over the past two years has been work on “word embeddings” [83, 78, 111, 15]. The main idea is to map every word w to a vector $v_w \in \mathbb{R}^d$ (typically $d \approx 500$) in such a way that the geometry of the vectors captures the semantics of the word.² One of the main constructions for such embeddings is to form the $M \times M$ matrix whose rows/columns are indexed by words, with (i, j) -th entry corresponding to the total number of times the i -th and j -th word occur next to (or near) each other in a large corpus of text (e.g. wikipedia). The word embedding is then computed as the rows of the singular vectors corresponding to the top rank d approximation to this empirical count matrix.³ These embeddings have proved to be extremely effective, particularly when used as a way to map text to features that can then be trained in downstream applications. Despite their successes, current embeddings seem to suffer from sampling noise in the count matrix (where many transformations of the count data are employed, e.g. see [110])—this is especially noticeable in the relatively poor quality of the embeddings for relatively rare words. The recent theoretical work [16] sheds some light on why current approaches are so successful, yet the following question largely remains: Is there a more accurate way to recover the best rank- d approximation of the underlying matrix than simply computing the best rank- d approximation for the (noisy) matrix of empirical counts?

Efficient Algorithms for Latent Variable Models. There is a growing body of work from the algorithmic side (as opposed to information theoretic) on how to recover the structure underlying various structured statistical settings. This body of work includes work on learning HMMs [58, 90, 34], recovering low-rank structure [14, 13, 24], and learning or clustering various structured distributions such as Gaussian

²The goal of word embeddings is not just to cluster similar words, but to have semantic notions encoded in the geometry of the points: the example usually given is that the direction representing the difference between the vectors corresponding to “king” and “queen” should be similar to the difference between the vectors corresponding to “man” and “woman”, or “uncle” and “aunt”, etc.

³A number of pre-processing steps have been considered, including taking the element-wise square roots of the entries, or logarithms of the entries, prior to computing the SVD.

mixture models [37, 121, 85, 23, 57, 64, 48] and latent dirichlet allocation (a very popular topic model) [10]. A number of these methods essentially can be phrased as solving an inverse moments problem, and the work in [7] provides a unifying viewpoint for computationally efficient estimation for many of these models under a tensor decomposition perspective. In general, this body of work has focussed on the computational issues and has considered these questions in the regime in which the amount of data is plentiful—well above the information theoretic limits.

Sublinear Sample Testing and Estimation. In contrast to the work described in the previous section on efforts to devise computationally efficient algorithms for tackling complex structural settings in the “over-sampled” regime, there is also significant work establishing information theoretically optimal algorithms and (matching) lower bounds for estimation and distributional hypothesis testing in the most basic setting of independent samples drawn from (unstructured) distributions. This work includes algorithms for estimating basic statistical properties such as entropy [93, 52, 116, 118], support size [97, 116], distance between distributions [116, 118, 117], and various hypothesis tests, such as whether two distributions are very similar, versus significantly different [50, 20, 92, 119, 27], etc. While many of these results are optimal in a worst-case (“minimax”) sense, there has also been recent progress on instance optimal (or “competitive”) estimation and testing, e.g. [3, 4, 119], with stronger information theoretic optimality guarantees. There has also been a long line of work beginning with [28, 21] on these tasks in “simply structured” settings, e.g. where the domain of the distribution has a total ordering or where the distribution is monotonic or unimodal.

2.2 Main Results

2.2.1 Recovering Low Rank Probability Matrices

For rank $R = 2$, it is possible to recover the dictionary $P = [p, q]$ uniquely up to column permutation. Assume that W is symmetric, where $w_L = w_R = w$ (all our

results extend to the asymmetric case). Define the *marginal probability* vector, ρ and the *dictionary separation* vector as:

$$\rho_i := \sum_k \mathbb{B}_{i,k}, \quad \Delta := w(p - q). \quad (2.2)$$

Observe that in this rank 2 case, the matrix $\text{Cov}(\mathbb{B})$ admits a unique rank-1 decomposition, which implies that:

$$\mathbb{B} = \rho\rho^\top + \Delta\Delta^\top. \quad (2.3)$$

We focus on a class of model parameters where p and q are well separated, which assumption guarantees that the rank 2 matrix \mathbb{B} is well-conditioned. This assumption also has natural interpretations in different applications including community detection, topic modeling, and HMMs.

Assumption 1 (Separation). *Assume that w_p and w_q are lower bounded by some constant $C_w = \Omega(1)$, and assume that the ℓ_1 -norm of the dictionary separation is lower bounded by $\|\Delta\|_1 \geq C_\Delta = \Omega(1)$.*

Theorem 2.1 (Upper bound for rank 2 matrices). *Suppose we have access to N i.i.d. samples generated according to the a rank 2 symmetric probability matrix \mathbb{B} parameterized as (2.1), and suppose the true matrix satisfies Assumption 1. For $\epsilon > 0$, with $N = \Theta(M/\epsilon^2)$ samples, our algorithm runs in time $\text{poly}(M)$ and returns estimators \widehat{B} , $\widehat{\rho}$, $\widehat{\Delta}$, such that with large probability:*

$$\|\widehat{B} - \mathbb{B}\|_1 \leq \epsilon, \quad \|\widehat{\rho} - \rho\|_1 \leq \epsilon, \quad \|\widehat{\Delta} - \Delta\|_1 \leq \epsilon.$$

(here, the ℓ_1 -norm of an $M \times M$ matrix P is simply defined as $\|P\|_1 = \sum_{i,j} |P_{i,j}|$).

Note that for $R > 2$, the dictionary matrix P and the mixing matrix W are not uniquely identifiable. We only focus on obtaining a low rank estimator for the underlying probability matrix \mathbb{B} .

Theorem 2.2 (Upper bound for rank R , constant accuracy). *Suppose we have access to N i.i.d. samples generated according to the a probability matrix \mathbb{B} parameterized as (2.1). Assume the mixing matrix W is PSD with row sums bounded by $\sum_j W_{i,j} \geq w_{min}$. Fix constant accuracy $\epsilon_0 > 0$ and $\epsilon_0 = \Omega(1)$, for any $r > 0$, with $N = \Theta(\frac{MR^2}{w_{min}^2 \epsilon_0^{4+r}})$ samples, our algorithm runs in time $\text{poly}(M)$ and returns a rank R estimator \widehat{B} such that with large probability:*

$$\|\widehat{B} - \mathbb{B}\|_1 \leq \epsilon_0. \quad (2.4)$$

Compared to the sample complexity result $N = \Theta(MR^2)$ for the community detection problem with R communities as in [36], in the more general parameterization, we incur an extra w_{min}^{-2} dependence, which can be easily removed in the special setup of community detection to recover the result in the community detection problem.

Assumption 2 (Well separated dictionary). *We assume that the minimal singular value of \mathbb{B}^{sqr} scaled with the inverse square root of the exact marginal probabilities is lower bounded.*

$$\sigma_R(\text{Diag}(\rho_i)^{-1/2} \mathbb{B} \text{Diag}(\rho_i)^{-1/2}) \geq \sigma_{min}. \quad (2.5)$$

Note that in the ideal case where the support of the dictionaries are non-overlapping, and the mixing matrix W is diagonal, we have

$$\sigma_1(\text{Diag}(\rho_i)^{-1/2} \mathbb{B} \text{Diag}(\rho_i)^{-1/2}) = \sigma_R(\text{Diag}(\rho_i)^{-1/2} \mathbb{B} \text{Diag}(\rho_i)^{-1/2}) = 1.$$

Under the well-separation assumption for the dictionary, we can sharpen the error bound.

Theorem 2.3. (Upper bound for rank R under separation condition) *Under the conditions of Theorem 2.2, further assume that Assumption 2 is satisfied for $\sigma_{min} > \epsilon_0$, and that $N = \Omega(\frac{MR^2}{w_{min}^2 \epsilon_0^{4+r}})$ for any $r > 0$. For any $\epsilon > 0$ such that $\epsilon < \epsilon_0$, with $N = \Theta(\frac{MR}{\epsilon^2})$ samples, our algorithm runs in time $\text{poly}(M)$ and returns a rank R*

estimator \widehat{B} such that with large probability:

$$\|\widehat{B} - \mathbb{B}\|_1 \leq \epsilon. \quad (2.6)$$

Note that when the marginal probabilities ρ_i are not roughly uniform, spectral error bounds in terms of $\|\widehat{B} - \mathbb{B}\|_2$ are not particularly strong. Instead, here we consider the ℓ_1 norm error bound, or equivalently the total variation distance, which is a more natural measure of estimation error for probability distributions. Moreover, note that naively estimating a distribution over M^2 outcomes requires order M^2 samples. Our algorithm utilizes the low rank structure of the underlying probability matrix to achieve a sample complexity which is *precisely* linear in the vocabulary size M .

We now turn to the implications of this theorem to testing and learning problems.

2.2.2 Topic Models and Hidden Markov Models

One of the main motivations for considering the specific low rank structure on the underlying matrix \mathbb{B} is that this structure encompasses the structure of the matrix of expected bigrams generated by both topic models and HMMs. We now make these connections explicit for the rank 2 case, and then briefly discuss the rank R case.

Definition 2.1. *A 2-topic model over a vocabulary of size M is defined by a pair of distributions, p and q supported over M words, and a pair of topic mixing weights π_p and $\pi_q = 1 - \pi_p$. The process of drawing a bigram (i, j) consists of first randomly picking one of the two “topics” according to the mixing weights, and then drawing two independent words from the word distribution corresponding to the chosen topic. Thus the probability of seeing bigram (i, j) is $(\pi_p p_i p_j + \pi_q q_i q_j)$, and so the expected bigram matrix can be written as $\mathbb{B} = PWP^\top$ with $P = [p, q]$, and $W = [\pi_p, 0; 0, \pi_q]$.*

Definition 2.2. *A hidden Markov model with 2 hidden states (s_p, s_q) and a size M observation vocabulary is defined by a 2×2 transition matrix T for the 2 hidden states, and two distributions of observations, p and q , corresponding to the 2 states.*

A sequence of N observations is sampled as follows: First, select an initial state according to the stationary distribution of the underlying Markov chain $[\pi_p, \pi_q]$; Then evolve the Markov chain according to the transition matrix T for N steps; For each $n \in \{1, \dots, N\}$, the n -th observation in the sequence is generated by making an independent draw from either distribution p or q according to whether the Markov chain is in state s_p or s_q at the n -th timestep.

The probability that seeing a bigram (i, j) for the n and the $(n + 1)$ -th observation is given by $\pi_p p_i (T_{p,p} p_j + T_{p,q} q_j) + \pi_q q_i (T_{q,p} p_j + T_{q,q} q_j)$, and hence the expected bigram matrix can be written as $\mathbb{B} = DW D^\top$ with $D = [p, q]$, and $W = \begin{bmatrix} \pi_p & 0 \\ 0 & \pi_q \end{bmatrix} \begin{bmatrix} T_{p,p} & 1 - T_{p,p} \\ 1 - T_{q,q} & T_{q,q} \end{bmatrix}$.

The following corollaries (straightforward by Theorem 2.1) shows that parameter estimation is possible with sample size *linear* in M :

Corollary 2.1. (*Learning 2-topic models*) Suppose we are in the 2-topic model setting. Assume that $\pi_p(1 - \pi_p)\|p - q\|_1 = \Omega(1)$. There exists an algorithm which, given $N = \Omega(M/\epsilon^2)$ bigrams, runs in time $\text{poly}(M)$ and with large probability returns estimates $\hat{\pi}_p, \hat{p}, \hat{q}$ such that

$$|\hat{\pi}_p - \pi_p| < \epsilon, \|\hat{p} - p\|_1 \leq \epsilon, \|\hat{q} - q\|_1 \leq \epsilon.$$

Corollary 2.2. (*Learning 2-state HMMs*) Suppose we are in the 2-state HMM setting. Assume that $\|p - q\|_1 \geq C_1$ and that $\pi_p, T_{p,p}, T_{q,q}$ are lower bounded by C_2 and upper bounded by $1 - C_2$, where both C_1 and C_2 are $\Omega(1)$. There exists an algorithm which, given a sampled chain of length $N = \Omega(M/\epsilon^2)$, runs in time $\text{poly}(M)$ and returns estimates $\hat{\pi}_p, \hat{T}, \hat{p}, \hat{q}$ such that, with high probability, we have (that there is exists a permutation of the model such that)

$$|\hat{\pi}_p - \pi_p| < \epsilon, |\hat{T}_{p,p} - T_{p,p}| < \epsilon, |\hat{T}_{q,q} - T_{q,q}| < \epsilon, \|\hat{p} - p\|_1 \leq \epsilon, \|\hat{q} - q\|_1 \leq \epsilon.$$

Furthermore, it is sufficient for this algorithm to only utilize $\Omega(M/\epsilon^2)$ random bigrams and only $\Omega(1/\epsilon^2)$ random trigrams from this chain.

For topic models with $R > 2$ topics and HMMs with $R > 2$ hidden states, the matrix of bigram probabilities does not uniquely determine the underlying HMM. One can recover the model parameters using sampled trigram sequences (see [7] for the moment structure in the trigrams). However, the core step remains to first obtain an accurate estimate of \mathbb{B} given by Theorem 2.2 and 2.3⁴. We do not go into details here.

2.2.3 Testing vs. Learning

The above theorem and corollaries are tight in an extremely strong sense: for both the topic model and HMM settings, it is information theoretically impossible to perform even the most basic property tests using fewer than $\Theta(M)$ samples. For topic models, the community detection lower bounds [88][73][127] imply that $\Theta(M)$ bigrams are necessary to even distinguish between the case that the underlying model is simply the uniform distribution over bigrams versus the case of a R -topic model in which each topic corresponds to a uniform distribution over disjoint subsets of M/R words. For 2-state HMMs, even if we permit an estimator to have more information than merely bigram counts, namely the *full sequence* of observations, we prove the following linear lower bound.

Theorem 2.4. *There exists a constant $c > 0$ such that for sufficiently large M , given a sequence of observations from a HMM with two states and emission distributions p, q supported on M elements, even if the underlying Markov process is symmetric, with transition probability $1/4$, it is information theoretically impossible to distinguish the case that the two emission distributions, $p = q = \text{Unif}[M]$ from the case that $\|p - q\|_1 = 1$ with probability greater than $2/3$ using a sequence of fewer than cM observations.*

This immediately implies the following corollary for estimating the *entropy rate* of an HMM.

⁴E.g. see [7] for how the bigram matrix can be used in the estimation problem in a “whitening” step to reduce the problem from one of M dimensions to one with effectively R dimensions.

Corollary 2.3. *There exists an absolute constant $c > 0$ such that given a sequence of observations from a HMM with two hidden states and emission distributions supported on M elements, a sequence of cM observations is information theoretically necessary to estimate the entropy rate to within an additive 0.5 with probability of success greater than $2/3$.*

These strong lower bounds for property testing and estimation are striking for several reasons. First, the core of our learning algorithm is a matrix reconstruction step that uses only the set of bigram counts. Conceivably, one could significantly benefit from considering longer sequences of observations — even for HMMs that mix in constant time, there are detectable correlations between observations separated by $O(\log M)$ steps. Regardless, our lower bound shows that actually no additional information from such longer k -grams can be leveraged to yield sublinear sample property testing or estimation.

A second notable point is the apparent brittleness of sublinear property testing and estimation as we deviate from the standard (unstructured) i.i.d sampling setting. Indeed for nearly all distributional property estimation or testing tasks, including testing uniformity and estimating the entropy, sublinear-sample testing and estimation is possible in the i.i.d. sampling setting (e.g. [50, 118, 117]). In contrast to the i.i.d. setting in which estimation and testing require asymptotically fewer samples than *learning*, as the above results illustrate, even in the setting of an HMM with just two hidden states, learning and testing require comparable numbers of observations.

2.3 Outline of our estimation algorithm

Given N samples drawn according to the probability matrix \mathbb{B} . Let B denote the matrix of average empirical counts. By the Poisson assumption on sample size, we have that $[B]_{i,j} \sim \frac{1}{N} \text{Poi}(N\mathbb{B}_{i,j})$.

Before introducing our algorithm, let us consider the naive approach of estimating \mathbb{B} by taking the rank R truncated SVD of the empirical matrix B , which concentrates to \mathbb{B} in spectral distance asymptotically. Unfortunately, this approach leads to a

sample complexity as large as $\Theta(M^2 \log M)$, and in the linear sample size regime, the empirical counts matrix is a poor representation of the underlying distribution. Intuitively, due to the sampling noise, the rows and columns of B corresponding to words with larger marginal probabilities have higher row and column sums in expectation, as well as higher variances that undermine the spectral concentration of the matrix as a whole. This observation leads to the idea of pre-scaling the matrix so that every word (i.e. row/column) is roughly of unit variance. Indeed, with a slight modification of the truncated SVD, we can improve the sample complexity of this approach to $\Theta(M \log M)$, which is nearly linear. Interestingly, if we get to observe a matrix $(\mathbb{B} + E)$ where the noise matrix E are i.i.d. sub-Gaussian variables of unit variance, then truncated SVD indeed gives us the optimal estimator for \mathbb{B} . Our algorithm shows that we can actually shave off the log factor for a broad class of noise (sub-exponential), which require more careful steps than truncated SVD to denoise the empirical matrix.

Next, we sketch the outline of our algorithms (Algorithm 3 for rank 2 case and Algorithm 4 for general rank R case). We only highlight the intuition behind the key ideas, and defer the detailed analysis of the algorithms to Section 2.4 and 2.5.

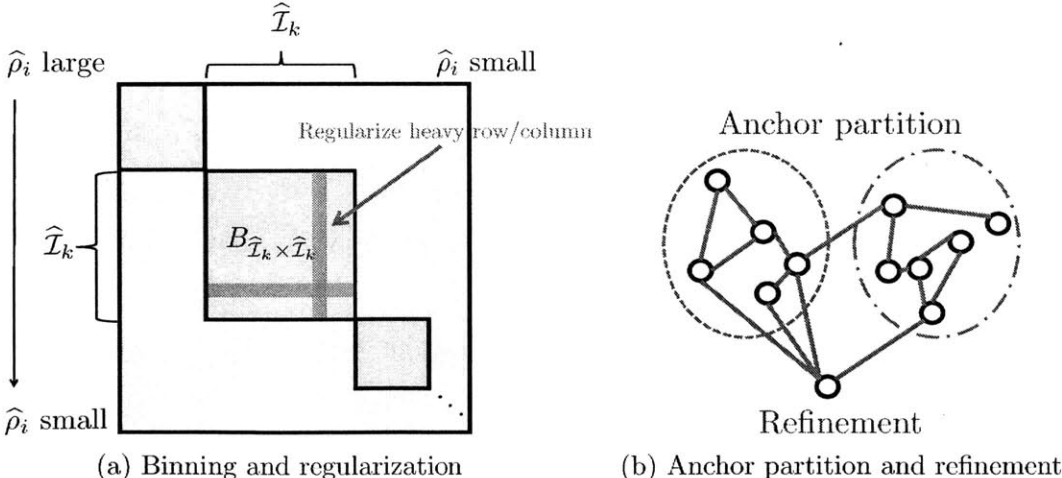


Figure 2-1: The key algorithmic ideas of our algorithm.

2.3.1 Rank 2 algorithm

First, note that it is straightforward to obtain an estimate $\hat{\rho}$ close to the true marginal ρ with linear sample complexity. Also, recall that $\mathbb{B} - \rho\rho^\top = \Delta\Delta^\top$ as per (2.3), hence after subtracting off the relatively accurate rank 1 matrix of $\hat{\rho}\hat{\rho}^\top$, we are essentially left with a rank 1 matrix recovery problem. Our Algorithm 3 consists of two phases:

Phase I: “binning” and “regularization” In Section 2.1, we drew the connection between our problem and the community detection problem in sparse random graphs. Recall that when the word marginals are roughly uniform, namely all in the order of $O(\frac{1}{M})$, the linear sample regime corresponds to the stochastic block model setup where the expected row sums are all in the order of $d_0 = \frac{N}{M} = \Omega(1)$. It is well-known that in this sparse regime, the adjacency matrix, or the empirical count matrix B_N in our problem, does not concentrate to the expectation matrix in the spectral distance. Due to some heavy rows with row sum in the order of $\Omega(\frac{\log M}{\log \log M})$, the leading eigenvectors are polluted by the local properties of these heavy nodes and do not reveal the global structure of the graph, which are precisely the desired information in expectation.

In order to enforce spectral concentration in the linear sample size regime, one of the many techniques is to tame the heavy rows and columns by setting them to 0. This simple idea was first introduced by [47], and followed by analysis works in [44] and many others. Recently in [75] and [76] the authors provided clean and clever proofs to show that *any* such “regularization” essentially leads to better spectral concentration for the adjacency matrix of random graphs whose row/column sums are roughly uniform in expectation.

Phase I of Algorithm 3 leverage such “regularization” ideas in our problem where the marginal probabilities are not uniform with the idea of “binning”. A natural candidate solution would be to partition the vocabulary \mathcal{M} into bins of words according to the word marginals, so that the words in the same bin have roughly uniform marginals. Restricting our attention to the diagonal blocks of \mathbb{B} whose indices are in the same bin, the expected row and column sums are indeed roughly uniform. Then

we can regularize (by removing abnormally heavy rows and columns) each diagonal block separately to restore spectral concentration, to which truncated SVD should then apply. Figure 2-1a visualizes the two operations of “binning” and “regularization” in Phase I of Algorithm 3.

Phase I returns estimates $\hat{\rho}$ and $\hat{\Delta}$ both up to a small constant accuracy in ℓ_1 norm with $\Theta(M)$ samples. There are 3 concerns we rigorously address in order to prove the correctness of the algorithm:

1. We do not have access to the exact marginal ρ . With linear sample size, we only can estimate ρ up to constant accuracy in ℓ_1 norm. If we implement binning according to the empirical marginals, there is considerable probability with which words with large marginals are placed in a bin intended for words with small marginals — which we call “spillover effect”. When directly applied to the empirical bins with such spillover, the existing results of “regularization” in [76] do not lead to the desired concentration result.
2. When restricting to each diagonal block corresponding to a bin, we throw away all the sample counts outside the block. This greatly reduces the effective sample size, and it is not obvious that we retain enough samples in each diagonal block to guarantee meaningful estimation.
3. Even if the “regularization” trick works for each diagonal block, we need to extract the useful information and “stitch” together this information from each block to provide an estimator for the entire matrix, including the off-diagonal blocks.

Phase II: “Anchor partition” Under the separation Assumption 1, Phase II of Algorithm 3 refine the estimates of Phase I to achieve the desired sample complexity bound.

The key to this refining process is to construct an “anchor partition”, which is a bi-partition of the vocabulary \mathcal{M} based on the signs of the estimate of separation vector $\hat{\Delta}$ given by Phase I. We collapse the $M \times M$ matrix B into a 2×2 matrix

corresponding to the bi-partition, and accurately estimate the 2×2 matrix with the N samples. Given this extremely accurate estimate of this 2×2 anchor matrix, we can now iteratively refine our estimates of ρ_i and Δ_i for each word i by solving a simple least square fitting problem.

Similar ideas — estimation refinement based on some crude global information — has appeared in many works for different problems. For example, in a recent paper [36] on community detection, after obtaining a crude classification of nodes using spectral algorithm, one round of a “correction” routine is applied to each node based on its connections to the graph partition given by the first round. This refinement immediately leads to an optimal rate of recovery. Figure 2-1b visualize the example of community detection. In our problem, the nodes are the M words, the edges are the sample counts, and instead of re-assigning the label to each node in the refinement routine, and we refine the estimation of ρ_i and Δ_i for each word.

2.3.2 Rank R algorithm

We summarize the basic ideas of Algorithm below. In Step 1, we again group words according to the empirical marginal probabilities, so that in each bin words are of similar marginals. Then in Step 2, we consider the diagonal blocks of the empirical average bigram matrix B , which rows and columns correspond to the words in the same bin. In each of such diagonal blocks, the entries have roughly uniform expectations, similar to Phase 1 of Algorithm 3, we regularize each diagonal block in the empirical matrix by removing abnormally heavy rows and columns, and then apply truncated SVD to obtain a sharper concentration bound.

After estimate the span of the dictionary restricted to words in each bin by looking at the leading rank R subspace of each diagonal block, in Step 3, we aim to estimate $\text{Diag}(\rho)^{1/2} \mathbb{B} \text{Diag}(\rho)^{1/2}$ accurately in spectral norm. With the marginal probability scaling, such error bound naturally translates into error bound for estimating \mathbb{B} in ℓ_1 norm. To achieve this, we regularize and approximately scale the empirical matrix B with the empirical marginal probability, and then project the entire matrix to a $R \log M$ -dimensional subspace as a union of the spans for each bin found in

Step 2. Since such spans are estimated accurately enough, projecting the $M \times M$ dimensional matrix to the $R \log M$ -dimensional subspace preserves the signal that is correlated with the expectation while significantly reducing the statistical noise from sampling. This guarantees a sharp spectral concentration to the expectation $\text{Diag}(\rho)^{1/2} \mathbb{B} \text{Diag}(\rho)^{1/2}$.

In the last step, similar to the Phase II of Algorithm 3, if the underlying true probability is “well-conditioned” we can further improve the sample complexity by refine the estimation.

Algorithm 3: Rank 2 algorithm

Input: $2N$ sample counts.

Output: Estimates $\hat{\rho}$, $\hat{\Delta}$, \hat{B} .

Divide the sample counts into two independent batches of equal size N , and construct two average empirical matrices. Each of the following two steps uses an independent copy of B .

Phase I**1. Binning according to the empirical marginal probabilities**

Set $\hat{\rho}_i = \sum_{j \in [M]} [B]_{i,j}$. Partition the vocabulary \mathcal{M} into:

$$\hat{\mathcal{I}}_0 = \{i : \hat{\rho}_i < \frac{\epsilon_0}{M}\}, \hat{\mathcal{I}}_{\log} = \left\{i : \hat{\rho}_i > \frac{\log(M)}{M}\right\}, \hat{\mathcal{I}}_k = \left\{i : \frac{\epsilon^k}{M} \leq \hat{\rho}_i \leq \frac{\epsilon^{k+1}}{M}\right\}, k = 1 : \log \log(M).$$

2. Estimate separation vector in each bin (up to sign flip). Set $\hat{\Delta}_{\hat{\mathcal{I}}_0} = 0$.

If $\sum \hat{\rho}_{\hat{\mathcal{I}}_{\log}} < \epsilon_0$, set $\hat{\Delta}_{\hat{\mathcal{I}}_{\log}} = 0$, else

(Rescaling): Set $E = \text{diag}(\hat{\rho}_{\hat{\mathcal{I}}_{\log}})^{-1/2} [B - \hat{\rho}\hat{\rho}^\top]_{\hat{\mathcal{I}}_{\log} \times \hat{\mathcal{I}}_{\log}} \text{diag}(\hat{\rho}_{\hat{\mathcal{I}}_{\log}})^{-1/2}$.

(SVD): Let $u_{\log} u_{\log}^\top$ be the rank-1 truncated SVD of E . Set $v_{\log} = \text{diag}(\hat{\rho}_{\hat{\mathcal{I}}_{\log}})^{1/2} u_{\log}$.

If $\sum \hat{\rho}_{\hat{\mathcal{I}}_k} < \epsilon_0 e^{-k}$, set $\hat{\Delta}_{\hat{\mathcal{I}}_k} = 0$, else

(Regularization): Set $d_k^{\max} = (\sum \hat{\rho}_{\hat{\mathcal{I}}_k}) \frac{\epsilon^{k+\tau}}{M}$, if a row/column of $[B]_{\hat{\mathcal{I}}_k \times \hat{\mathcal{I}}_k}$ has sum larger than $2d_k^{\max}$, set the entire row/column to 0. Let \tilde{B} denote the regularized block.

(SVD): Let $v_k v_k^\top$ be the rank-1 truncated SVD of $(\tilde{B} - \hat{\rho}_{\hat{\mathcal{I}}_k} \hat{\rho}_{\hat{\mathcal{I}}_k}^\top)$.

3. Stitching the segments. Fix $k^* = \arg \max_k \|v_k\|$, set $\hat{\Delta}_{\hat{\mathcal{I}}_{k^*}} = v_{k^*}$.

For all k , define $\mathcal{I}_k^+ = \{i : i \in \hat{\mathcal{I}}_k : \hat{\Delta}_i > 0\}$ and $\mathcal{I}_k^- = \hat{\mathcal{I}}_k \setminus \mathcal{I}_k^+$.

Set $\hat{\Delta}_{\hat{\mathcal{I}}_k} = v_k$ if $\frac{\sum_{i \in \mathcal{I}_k^+, j \in \mathcal{I}_k^+} [B]_{i,j}}{\sum_{i \in \mathcal{M}, j \in \mathcal{I}_k^+} [B]_{i,j}} > \frac{\sum_{i \in \mathcal{I}_k^+, j \in \mathcal{I}_k^-} [B]_{i,j}}{\sum_{i \in \mathcal{M}, j \in \mathcal{I}_k^-} [B]_{i,j}}$, and $\hat{\Delta}_{\hat{\mathcal{I}}_k} = -v_k$ otherwise.

Phase II**1. (Construct anchor partition)** Set $\mathcal{A} = \emptyset$. For all empirical bins, if $\|\hat{\Delta}_{\hat{\mathcal{I}}_k}\|_2 \leq (\sqrt{d_k^{\max}/N})^{1/2}$, skip the bin; otherwise set $\mathcal{A} = \mathcal{A} \cup \{i \in \hat{\mathcal{I}}_k : \hat{\Delta}_i > 0\}$.**2. (Estimate anchor matrix)**

Set $B_{\mathcal{A}} = \begin{bmatrix} \sum_{i \in \mathcal{A}, j \in \mathcal{A}} [B_N]_{i,j} & \sum_{i \in \mathcal{A}, j \in \mathcal{A}^c} [B_N]_{i,j} \\ \sum_{i \in \mathcal{A}^c, j \in \mathcal{A}} [B_N]_{i,j} & \sum_{i \in \mathcal{A}^c, j \in \mathcal{A}^c} [B_N]_{i,j} \end{bmatrix}$. Set vector $b = \begin{bmatrix} \sum_{i \in \mathcal{A}, j \in \mathcal{M}} [B_N]_{i,j} \\ \sum_{i \in \mathcal{A}^c, j \in \mathcal{M}} [B_N]_{i,j} \end{bmatrix}$.

Set aa^\top to be rank-1 truncated SVD of the 2×2 matrix $(B_{\mathcal{A}} - bb^\top)$.

3. (Refine the estimation):

Set $\begin{bmatrix} \hat{\rho}^\top \\ \hat{\Delta}^\top \end{bmatrix} = [a, b]^{-1} \begin{bmatrix} \sum_{i \in \mathcal{A}} [B_N]_{i, \mathcal{M}} \\ \sum_{i \in \mathcal{A}^c} [B_N]_{i, \mathcal{M}} \end{bmatrix}$

Return $\hat{\rho}$, $\hat{\Delta}$, and $\hat{B} = \hat{\rho}\hat{\rho}^\top + \hat{\Delta}\hat{\Delta}^\top$.

Algorithm 4: Rank R algorithm

Input: $4N$ i.i.d. samples from the distribution \mathbb{B} of dimension $M \times M$.

(In each of the 4 steps, B refers to an independent copy of the bigram matrix with N samples.)

Output: Rank R estimator \widehat{B} for \mathbb{B} , and \widehat{V} for the rank R subspace of scaled matrix $D_S \mathbb{B}^{sqr}$.

1. **(Binning according to the empirical marginal probabilities)**

Set $\widehat{\rho}_i = \sum_{j \in [M]} [B]_{i,j}$. Define $\bar{\rho}_k = \frac{1}{M} e^k$. Partition the vocabulary \mathcal{M} into:

$$\widehat{\mathcal{I}}_0 = \{i : \widehat{\rho}_i < \bar{\rho}_1\}, \text{ and } \widehat{\mathcal{I}}_k = \{i : \bar{\rho}_k \leq \widehat{\rho}_i \leq \bar{\rho}_{k+1}\}, \text{ for } k = 1 : \log M.$$

Sort the M words according to $\widehat{\rho}_i$ in ascending order.

Set $\widehat{W}_k = \sum_{i \in \widehat{\mathcal{I}}_k} \widehat{\rho}_i$ and $\widehat{M}_k = |\widehat{\mathcal{I}}_k|$. Set the block diagonal matrix

$$D_S = \begin{bmatrix} \bar{\rho}_1^{-1/2} I_{\widehat{M}_1} & & & \\ & \ddots & & \\ & & \bar{\rho}_{\log M}^{-1/2} I_{\widehat{M}_{\log M}} & \\ & & & \ddots \end{bmatrix}. \quad (2.7)$$

2. **(Estimate dictionary span in each bin)**

For each bin $\widehat{\mathcal{I}}_k$, if $\widehat{W}_k < \epsilon_0 e^{-k}$, set $\widehat{V}_k = 0$; else consider diagonal block $B_k = [B]_{\widehat{\mathcal{I}}_k \times \widehat{\mathcal{I}}_k}$:

(a) **(Regularization):** Set $d_k^{\max} = \widehat{W}_k \bar{\rho}_k$. If a row/column of B_k has sum larger than $2d_k^{\max}$, set the entire row/column to 0. Denote the regularized block by \widetilde{B}_k .

(b) **(R -SVD):** Let the columns of \widehat{V}_k denote the leading R singular vectors of \widetilde{B}_k .

3. **(Estimate dictionary span and an ℓ_1 estimator \widehat{B}_2)** Set the projection matrix

$$\text{Proj}_{\widehat{V}} = \begin{bmatrix} \text{Proj}_{\widehat{V}_1} & & & \\ & \ddots & & \\ & & \text{Proj}_{\widehat{V}_{\log M}} & \\ & & & \ddots \end{bmatrix}. \quad (2.8)$$

(a) **(Regularization):** For each word i in bin $\widehat{\mathcal{I}}_k$, if the corresponding row in B has sum larger than $2\bar{\rho}_k$, set the entire row and column to zero. Denote the regularized average bigram matrix by \widetilde{B} .

(b) **(R -SVD):** Set \widehat{B}_0 to be the rank- R truncated SVD of matrix $\text{Proj}_{\widehat{V}} D_S \widetilde{B} D_S \text{Proj}_{\widehat{V}}$.

Let the columns of \widehat{V} denote the leading R singular vectors of \widehat{B}_0 .

4. **(Refinement to get ℓ_1 estimator)**

Repeat the regularization in Step 3 on B , let \widetilde{B} denote regularized average bigram matrix.

Set $Y = (\widehat{V}^\top D_S \widetilde{B} D_S \widehat{V})^{-1/2} (\widehat{V}^\top D_S \widetilde{B} D_S)$, Set $\widehat{B} = D_S^{-1} Y Y^\top D_S^{-1}$.

Return \widehat{B} and \widehat{V} .

2.4 Details of Rank 2 Algorithm

Given N samples, the goal is to estimate the word marginal vector ρ as well as the dictionary separation vector Δ up to constant accuracy in ℓ_1 norm. We denote the estimates by $\widehat{\rho}$ and $\widehat{\Delta}$. Also, we estimate the underlying probability matrix \mathbb{B} with $\widehat{B} = \widehat{\rho}\widehat{\rho}^\top + \widehat{\Delta}\widehat{\Delta}^\top$. Note that since $\|\Delta\|_1 \leq \|\rho\|_1 = 1$, constant ℓ_1 norm accuracy in $\widehat{\rho}$ and $\widehat{\Delta}$ immediately lead to constant accuracy of \widehat{B} also in ℓ_1 norm.

In this section, we prove Theorem 2.5 and Theorem 2.6 about the correctness of the 2 Phases of Algorithm 3, the detailed proofs are provided in Section 2.7.1 in the appendix.

Throughout the section, we denote the ratio between sample size and the vocabulary size by

$$d_0 = N/M, \tag{2.9}$$

and we assume that d_0 is lower bounded by a large constant such that

$$d_0/\log d_0 > \epsilon_0^{-4}.$$

Theorem 2.5 (Linear sample complexity of Rank 2 algorithm). *Fix ϵ_0 to be a small constant. Given $N = \Theta(M)$ samples, with large probability, Phase I of Algorithm 3 estimates ρ and Δ with accuracy:*

$$\|\widehat{\rho} - \rho\|_1 < \epsilon_0, \quad \|\widehat{\Delta} - \Delta\|_1 < \epsilon_0, \quad \|\widehat{B} - \mathbb{B}\|_1 < \epsilon_0.$$

Under the separation assumptions of Δ , we can refine the estimation to achieve arbitrary ϵ accuracy.

Theorem 2.6 (Refinement of Rank 2 algorithm). *Assume that \mathbb{B} satisfies the $\Omega(1)$ separation assumption. Given N samples, with probability at least $(1 - \delta)$, Phase II*

of our Algorithm 3 estimates ρ and Δ up to accuracy in ℓ_1 norm:

$$\|\widehat{\rho} - \rho\|_1 < \sqrt{M/\delta N}, \quad \|\widehat{\Delta} - \Delta\|_1 < \sqrt{M/\delta N}, \quad \|\widehat{B} - \mathbb{B}\|_1 = O(\sqrt{M/\delta N}).$$

First, we show that it is easy to estimate the marginal probability vector ρ up to constant accuracy.

Lemma 2.1 (Estimate the word marginal probability ρ). *Given the average empirical count matrix B , we estimate the marginal probabilities by:*

$$\widehat{\rho}_i = \sum_{j \in \mathcal{M}} B_{i,j}. \quad (2.10)$$

With probability at least $(1 - \delta)$, we can bound the estimation accuracy by:

$$\|\widehat{\rho} - \rho\|_1 \leq \frac{1}{\sqrt{d_0 \delta}}. \quad (2.11)$$

The hard part is to estimate the separation vector Δ with linear number of sample counts, namely when $d_0 = \Theta(1)$. Recall that in the linear sample size regime, naively taking the rank-1 truncated SVD of $(B - \widehat{\rho}\widehat{\rho}^\top)$ fails to reveal any information about $\Delta\Delta^\top$, since the leading eigenvectors of B are dominated by the statistical noise of the sampling words with large marginal. Algorithm 3 achieves this with delicate steps. The organization of this section is as follows:

1. Section 2.4.1 introduces the binning argument and the necessary notations for the rest of the section. We group the M words into bins according to the empirical marginal probabilities, i.e. $\widehat{\rho}_i$'s. We call a bin “heavy” or “light” according to the marginal probability of a typical word in that bin.
2. Section 2.4.2 analyzes how to estimate the entries of Δ restricted to different empirical bins (up to some common sign flip). To achieve this, for the heaviest bin where words’ marginals are in the order of $\Omega(\log M/M)$, we can simply apply truncated SVD to the properly scaled diagonal block of the empirical average matrix B . For all other empirical bins, we examine the corresponding

diagonal blocks in B . The main challenge here is to deal with the spillover effect due to inexact binning, and Lemma 2.12 shows that with high probability, such spillover effect is very small *for all bins* with high probability. Then we leverage the clever proof techniques from [76] to show that given small spillover effect, we can first regularize each diagonal block and then apply truncated SVD to estimate the segments of separation vector.

3. Section 2.4.3 shows how to stitch the segments of estimates for Δ across different bins.
4. Section 2.4.4 shows that built upon the initialization, if the dictionary further satisfies certain separation condition, we can refine the estimation to improve its dependence on target accuracy ϵ to meet the information theoretic lower bound.

2.4.1 Binning

In order to estimate the separation vector Δ , instead of tackling the empirical count matrix B as a whole, we focus on its diagonal blocks and analyze the spectral concentration restricted to each block separately, using the fact that the entries $\mathbb{B}_{i,j}$ restricted to each diagonal block are roughly uniform.

For any set of words \mathcal{I} , we use $B_{\mathcal{I},\mathcal{I}}$ to denote the diagonal block of B whose row and column indices are in the set \mathcal{I} . When restricting to the diagonal block, the rank 2 decomposition of the expected matrix is given by $\mathbb{B}_{\mathcal{I},\mathcal{I}} = \rho_{\mathcal{I}}\rho_{\mathcal{I}}^{\top} + \Delta_{\mathcal{I}}\Delta_{\mathcal{I}}^{\top}$.

Empirical binning We partition the vocabulary \mathcal{M} according to the empirical marginal $\hat{\rho}$ in (2.10):

$$\hat{\mathcal{I}}_0 = \left\{ i : \frac{\epsilon_0}{M} \leq \hat{\rho}_i < \frac{1}{M} \right\}, \quad \hat{\mathcal{I}}_k = \left\{ i : \frac{e^{k-1}}{M} \leq \hat{\rho}_i < \frac{e^k}{M} \right\}, \quad \hat{\mathcal{I}}_{\log} = \left\{ i : \frac{\log M}{M} \leq \hat{\rho}_i \right\}. \quad (2.12)$$

We call this *empirical binning* to emphasize the dependence on the empirical estimator $\widehat{\rho}$, which is a random variable built from the first batch of N sample counts. We call $\widehat{\mathcal{I}}_0$ the *lightest empirical bin*, and $\widehat{\mathcal{I}}_{\log}$ the *heaviest empirical bin*, and $\widehat{\mathcal{I}}_k$ for $1 \leq k \leq \log \log M$ the *moderate empirical bins*.

For the analysis, we further define the *exact bins* according to the exact marginal probabilities:

$$\mathcal{I}_0 = \left\{ i : \frac{\epsilon_0}{M} \leq \rho_i < \frac{1}{M} \right\}, \quad \mathcal{I}_k = \left\{ i : \frac{e^{k-1}}{M} \leq \rho_i < \frac{e^k}{M} \right\}, \quad \mathcal{I}_{\log} = \left\{ i : \frac{\log M}{M} \leq \rho_i \right\}. \quad (2.13)$$

Note that since the target accuracy of Phase I is a small constant ϵ_0 , we can safely discard all the words with marginals less than ϵ_0/M as that incurs an ℓ_1 error only in the order of $O(\epsilon_0)$.

Spillover effect As N increases asymptotically, we will have $\widehat{\mathcal{I}}_k$ coincides with \mathcal{I}_k for every bin. However, in the linear regime where $N = \Theta(M)$, binning is inexact and we have the following two *spillover effects*:

1. Words from a heavy bin $\mathcal{I}_{k'}$, for k' much larger than k , are placed in a empirical bin $\widehat{\mathcal{I}}_k$;
2. Words from bin \mathcal{I}_k escape from the corresponding empirical bin $\widehat{\mathcal{I}}_k$.

The hope, that we can have good spectral concentration in each diagonal block $B_{\widehat{\mathcal{I}}_k, \widehat{\mathcal{I}}_k}$, crucially relies on the fact that the entries $\mathbb{B}_{i,j}$ restricted to this block are roughly uniform. However, the hope may be ruined by the spillover effects. Next, we show that with high probability the spillover effects are small for all bins with large probability mass:

1. In each empirical bin $\widehat{\mathcal{I}}_k$, the total probability mass of heavy words from the union of bins $\cup_{\{k': k' > k+1\}} \mathcal{I}_{k'}$ is only in the order of $O(e^{-e^k d_0/2})$ (see Lemma 2.12).
2. Most words of \mathcal{I}_k stays within the nearest empirical bins, namely the union of bins $\cup_{\{k': k-1 \leq k' \leq k+1\}} \widehat{\mathcal{I}}_{k'}$, (see Lemma 2.9).

Notations To analyze the spillover effects, we define some additional quantities.

We define the total marginal probability mass in the empirical bins to be:

$$W_k = \sum_{i \in \widehat{\mathcal{I}}_k} \rho_i, \quad (2.14)$$

and let $M_k = |\widehat{\mathcal{I}}_k|$ denote the total number of words in the empirical bin. We also define $\widehat{W}_k = \sum_{i \in \widehat{\mathcal{I}}_k} \widehat{\rho}_i$.

We use $\widehat{\mathcal{J}}_k$ to denote the set of spillover words into the empirical bin $\widehat{\mathcal{I}}_k$:

$$\widehat{\mathcal{J}}_k = \widehat{\mathcal{I}}_k \cap (\cup_{\{k': k' > k+1\}} \mathcal{I}_{k'}), \quad (2.15)$$

and let $\widehat{\mathcal{L}}_k$ denote the “good words” in the empirical bin $\widehat{\mathcal{I}}_k$:

$$\widehat{\mathcal{L}}_k = \widehat{\mathcal{I}}_k \setminus \widehat{\mathcal{J}}_k. \quad (2.16)$$

We also denote the total marginal probability mass of the heavy spillover words $\widehat{\mathcal{J}}_k$ by:

$$\overline{W}_k = \sum_{i \in \widehat{\mathcal{J}}_k} \rho_i. \quad (2.17)$$

Note that these quantities are random variables determined by the randomness of the first batch of N samples, in the binning step. We fix the binning when considering the empirical count matrix B (with independent batches of samples) in the other steps of the algorithm.

Define the upper bound of the “typical” word marginal in the k -th empirical bin to be:

$$\overline{\rho}_k = e^{\tau+1}/M,$$

Recall that we have $\sum_k \mathbb{B}_{i,k} = w_p p + w_q q$ and we assume $w_p, w_q \geq C_w = \Omega(1)$. We can bound each entry in \mathbb{B} by the product of marginal probabilities as

$$\mathbb{B}_{i,j} \leq \frac{2}{C_w^2} \rho_i \rho_j, \quad \forall i, j.$$

Let d_k^{\max} denote the expected max row/column sum of the diagonal block $\mathbb{B}_{\hat{\mathcal{I}}_k, \hat{\mathcal{I}}_k}$:

$$d_k^{\max} =: M_k \max_{i,j \in \mathcal{I}_k} \mathbb{B}_{i,j} = 2M_k \bar{\rho}_k^2 / C_w^2. \quad (2.18)$$

2.4.2 Estimate segments of Δ

Heaviest empirical bin First, we show that the empirical marginal probabilities of words in the heaviest bin concentrate much better than what Lemma 2.1 implies.

Lemma 2.2 (Concentration of marginal probabilities in the heaviest bin). *With high probability, for all the words with marginal probability $\rho_i \geq \epsilon_0 \log(M)/M$, for some universal constant C_1, C_2 ,*

$$C_1 \leq \hat{\rho}_i / \rho_i \leq C_2. \quad (2.19)$$

Lemma 2.2 says that we can estimate the marginal probabilities for every words in the heaviest bin with constant multiplicative accuracy. It also suggests that we do not need to worry about the words from \mathcal{I}_{\log} get spilled over into much lighter bins.

The next lemmas shows that with proper scaling, we can apply truncated SVD to the diagonal block to estimate the entries of separation vector Δ restricted to the empirical heaviest bin.

Lemma 2.3 (Estimate Δ restricted to the heaviest empirical bin). *Suppose that $\widehat{W}_{\log} = \sum \hat{\rho}_{\hat{\mathcal{I}}_{\log}} > \epsilon_0$. Define $\widehat{D}_{\hat{\mathcal{I}}_{\log}} = \text{Diag}(\hat{\rho}_{\hat{\mathcal{I}}_{\log}})$. Consider $B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}}$, the diagonal block corresponding to $\hat{\mathcal{I}}_{\log}$. Let E be the rank 1 truncated SVD of $\widehat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} (B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} - \hat{\rho}_{\hat{\mathcal{I}}_{\log}} \hat{\rho}_{\hat{\mathcal{I}}_{\log}}^\top) \widehat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2}$. Set $v_{\log} = \widehat{D}_{\hat{\mathcal{I}}_{\log}}^{1/2} E^{1/2}$. With large probability, we can estimate the dictionary separation vector restricted to the heaviest empirical bin up to sign flip with accuracy:*

$$\min\{\|\Delta_{\hat{\mathcal{I}}_{\log}} - v_{\log}\|_1, \|\Delta_{\hat{\mathcal{I}}_{\log}} + v_{\log}\|_1\} = O\left(\min\left\{\frac{1/d_0^{1/2}}{\|\Delta_{\hat{\mathcal{I}}_{\log}}\|_1}, 1/d_0^{1/4}\right\}\right). \quad (2.20)$$

The two cases in the above bound correspond to whether the separation is large or small, compared to the statistical noise from sampling, which is in the order $1/d_0^{1/4}$.

If the bin contains a large separation, then the bound follows the standard Wedin's perturbation bound; if the separation is small, i.e. $\|\Delta_{\widehat{\mathcal{I}}_{\log}}\|_1 \ll 1/d_0^{1/4}$, then the bound $1/d_0^{1/4}$ just corresponds to the magnitude of the statistical noise.

Moderate empirical bins In Lemma 2.12, we upper bound the spillover probability \overline{W}_k to show that the spillover effects are small for all the moderate bins. Given that, Lemma 2.5 and Lemma 2.6 show that we can first regularize each diagonal block and then apply truncated SVD to estimate the entries of the separation vector Δ restricted to each bin.

Lemma 2.4 (Bound spillover probabilities). *With high probability, for all empirical bins, we can bound \overline{W}_k defined in (2.17), the spillover probability from much heavier bins, by:*

$$\overline{W}_k \leq 2e^{-e^k d_0/2}. \quad (2.21)$$

Now consider $B_{\widehat{\mathcal{I}}_k, \widehat{\mathcal{I}}_k}$, the diagonal block corresponding to bin $\widehat{\mathcal{I}}_k$. We restrict attention to its spectral concentration on indices of $\widehat{\mathcal{L}}_k$, the set of “good words” defined in (2.16). To ensure the spectral concentration, we “regularize” it by removing the rows and columns with abnormally large sum. Recall that the expected row sum of the diagonal block without spillover is bounded by d_k^{\max} defined in (2.18). Let $\widehat{\mathcal{R}}_k$ denote the indices of the rows and columns in $B_{\widehat{\mathcal{I}}_k, \widehat{\mathcal{I}}_k}$ whose row sum or column sum are larger than $2d_k^{\max}$, namely

$$\widehat{\mathcal{R}}_k = \left\{ i \in \widehat{\mathcal{I}}_k : \sum_{j \in \widehat{\mathcal{I}}_k} B_{i,j} > 2d_k^{\max} \text{ or } \sum_{j \in \widehat{\mathcal{I}}_k} B_{j,i} > 2d_k^{\max} \right\}. \quad (2.22)$$

Starting with $\widetilde{B}_k = B_{\widehat{\mathcal{I}}_k \times \widehat{\mathcal{I}}_k}$, we set all the rows and columns of \widetilde{B}_k indexed by $\widehat{\mathcal{R}}_k$ to 0.

To make the operation of “regularization” more precise, we introduce some additional notations. Define $\widetilde{\rho}_k$ to be a vector with the same length as $\rho_{\widehat{\mathcal{I}}_k}$, with the

entries spillover words $\widehat{\mathcal{J}}_k$ set to 0,

$$(\widetilde{\rho}_k)_i = \rho_i \mathbf{1}_{i \in \widehat{\mathcal{L}}_k}. \quad (2.23)$$

Similarly define vector $\widetilde{\Delta}_k$ to be the separation vector restricted to the good words:

$$(\widetilde{\Delta}_k)_i = \Delta_i \mathbf{1}_{i \in \widehat{\mathcal{L}}_k}. \quad (2.24)$$

We define the matrix $\widetilde{\mathbb{B}}_k$ (of the same size as $B_{\widehat{\mathcal{L}}_k, \widehat{\mathcal{L}}_k}$):

$$\widetilde{\mathbb{B}}_k = \widetilde{\rho}_k \widetilde{\rho}_k^\top + \widetilde{\Delta}_k \widetilde{\Delta}_k^\top. \quad (2.25)$$

Note that by definition the rows and columns in \widetilde{B}_k and $\widetilde{\mathbb{B}}_k$ that are zero-ed out do not necessarily coincide. However, the next lemma shows that \widetilde{B}_k concentrates to $\widetilde{\mathbb{B}}_k$ in the spectral distance.

Lemma 2.5 (Spectral concentration of diagonal blocks.). *Suppose that the marginal of the bin $\widehat{\mathcal{L}}_k$ is large enough $W_k = \sum \rho_{\widehat{\mathcal{L}}_k} > \epsilon_0 e^{-k}$. With probability at least $(1 - M_k^{-r})$, for some universal constant r , we have*

$$\left\| \widetilde{B}_k - \widetilde{\mathbb{B}}_k \right\|_2 \leq Cr^{1.5} \frac{\sqrt{Nd_k^{\max} \log(Nd_k^{\max})}}{N}. \quad (2.26)$$

Proof. Here we highlight the key steps of the proof, and defer the detailed proof to Section 2.7.1.

In Figure 2-2, the rows and the columns of $B_{\widehat{\mathcal{L}}_k, \widehat{\mathcal{L}}_k}$ are sorted according to the exact marginal probabilities of the words in ascending order, with the rows and columns set to 0 by regularization shaded. Consider the block decomposition according to the good words $\widehat{\mathcal{L}}_k$ and the spillover words $\widehat{\mathcal{J}}_k$. We bound the spectral distance of the 4 blocks (A_1, A_2, A_3, A_4) separately. The bound for the entire matrix \widetilde{B}_k is then an immediate result of triangle inequality.

For block A_1 whose rows and columns all correspond to the “good words” with roughly uniform marginals, we show its concentration by applying the result in [76].

For block A_2 and A_3 , we show that after regularization the spectral norm of these two blocks are small. Intuitively, the expected row sums of block A_2 are bounded by $2d_k^{\max}$ and the expected column sums are bounded by $2d_k^{\max} \frac{\bar{W}_k}{W_k} = O(1/N)$, as a result of the bound on \bar{W}_k in Lemma 2.12. Thus the spectral norm of the block A_2 is likely to be bounded by $O(\sqrt{d_k^{\max}/N})$. We show this rigorously with high probability arguments. Lastly for block A_4 , which rows and columns all correspond to the spillover words. We show that the spectral norm of this block is very small, as a result of the small spillover marginal \bar{W}_k . \square

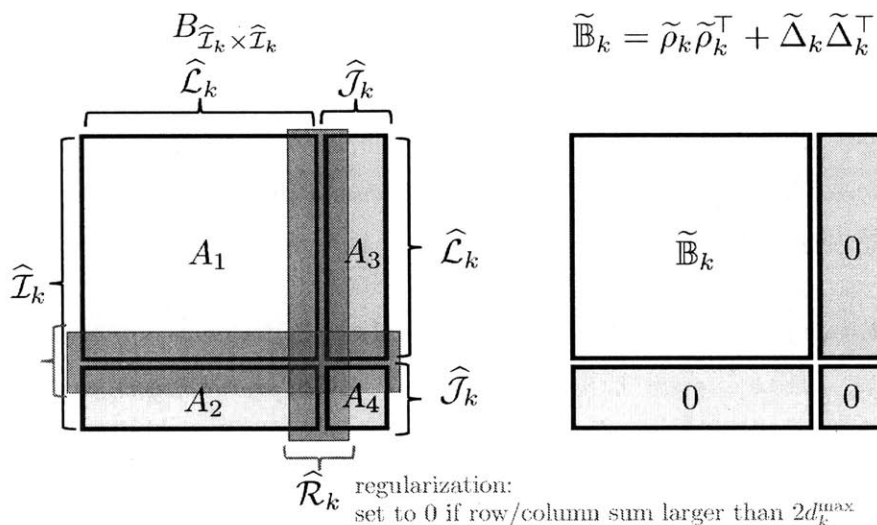


Figure 2-2: block decomposition of the diagonal block of $B_{\hat{\mathcal{I}}_k, \hat{\mathcal{I}}_k}$ corresponding to $\hat{\mathcal{I}}_k$.

Lemma 2.6 (Estimate the separation vector restricted to bins). *Suppose that $W_k = \sum_{i \in \hat{\mathcal{I}}_k} \rho_i > C_1 e^{-k}$ for some fixed constant $C_1 = \Omega(1)$. Let $v_k v_k^\top$ be the rank-1 truncated SVD of the matrix $(\tilde{B}_k - \hat{\rho}_{\hat{\mathcal{I}}_k} \hat{\rho}_{\hat{\mathcal{I}}_k}^\top)$. With high probability, we have*

$$\begin{aligned}
& \min\{\|\tilde{\Delta}_k - v_k\|_2, \|\tilde{\Delta}_k + v_k\|_2\} \\
& = O\left(\min\left\{\frac{\sqrt{Nd_k^{\max} \log(Nd_k^{\max})}}{N} \frac{1}{\|\tilde{\Delta}_{\hat{\mathcal{I}}_k}\|_2}, \left(\frac{\sqrt{Nd_k^{\max} \log(Nd_k^{\max})}}{N}\right)^{1/2}\right\}\right).
\end{aligned} \tag{2.27}$$

Claim 2.1 (Estimate the separation vector restricted to the lightest bin). *Setting $\widehat{\Delta}_{\widehat{\mathcal{I}}_0} = 0$ only incurs a small constant error:*

$$\|\Delta_{\widehat{\mathcal{I}}_0}\|_1 \leq \|\rho_{\widehat{\mathcal{I}}_0}\|_1 \leq \|\rho_{\mathcal{I}_0}\|_1 + \overline{W}_0 \leq \frac{\epsilon_0}{M}M + 2e^{-d_0/2} = O(\epsilon_0),$$

where we used the assumption that $d_0/\log d_0 \geq \epsilon_0^{-4}$.

2.4.3 Stitch the segments of $\widehat{\Delta}$

Given v_k for all k as estimation for $\Delta_{\widehat{\mathcal{I}}_k}$'s up to sign flips. Fix k^* to be one good bin (with large bin marginal and large separation). Partition the words into two groups $\mathcal{I}_{k^*}^+ = \{i : i \in \mathcal{I}_{k^*} : \widehat{\Delta}_i > 0\}$ and $\mathcal{I}_{k^*}^- = \mathcal{I}_{k^*} \setminus \mathcal{I}_{k^*}^+$. Without loss of generality assume that $\sum_{i \in \mathcal{I}_{k^*}^+} \widehat{\Delta}_i \geq \sum_{i \in \mathcal{I}_{k^*}^-} \widehat{\Delta}_i$. We set $\widehat{\Delta}_{\widehat{\mathcal{I}}_{k^*}} = v_{k^*}$. For all other good bins k , we similarly define \mathcal{I}_k^+ and \mathcal{I}_k^- . The next claim shows how to determine the relative sign flip of v_{k^*} and v_k .

Claim 2.2 (Pairwise comparison of bins to fix sign flips). *For all good bins $k \in \mathcal{G}$, we can fix the sign flip to be $\widehat{\Delta}_{\widehat{\mathcal{I}}_k} = v_k$ if:*

$$\frac{\sum_{i \in \mathcal{I}_{k^*}^+, j \in \mathcal{I}_k^+} [B]_{i,j}}{\sum_{i \in \mathcal{M}, j \in \mathcal{I}_k^+} [B]_{i,j}} > \frac{\sum_{i \in \mathcal{I}_{k^*}^+, j \in \mathcal{I}_k^-} [B]_{i,j}}{\sum_{i \in \mathcal{M}, j \in \mathcal{I}_k^-} [B]_{i,j}},$$

and $\widehat{\Delta}_{\widehat{\mathcal{I}}_k} = -v_k$ otherwise.

Proof. This claim is straightforward. When restricted to the good bins, the estimates v_k are accurate enough. We can determine that the sign flips of k^* and k are consistent if and only if the conditional distribution of the two word tuple $(x, y) \in \mathcal{M}^2$ satisfies $\Pr(x \in \mathcal{I}_{k^*}^+ | x \in \mathcal{I}_k^+) > \Pr(x \in \mathcal{I}_{k^*}^+ | x \in \mathcal{I}_k^-)$, and we should revert v_k otherwise. \square

Concatenate the segments of $\widehat{\Delta}$, we can bound the overall estimation error of the separation vector.

Lemma 2.7 (Estimate separation vector in Phase I). *For a fixed small constant $\epsilon_0 = O(1)$, if $d_0/\log(d_0) \geq \epsilon_0^{-4}$, with large probability, Phase I of Algorithm 3 estimates*

the separation vector Δ with constant accuracy in ℓ_1 norm:

$$\|\widehat{\Delta} - \Delta\| = O(\epsilon_0).$$

This concludes the proof for Theorem 2.5.

2.4.4 Refinement

Construct an anchor partition Imagine that we have a way to group the M words in the vocabulary into a new vocabulary with a *constant* number of superwords. The new probability matrix is obtained by summing over the rows and columns of the matrix \mathbb{B} according to the grouping. We similarly define marginal vector ρ_A and separation vector Δ_A over the superwords. If we group the words in a way such that the dictionary over the superwords is still well separated, then with $N = \Omega(M)$ samples we can estimate the constant dimensional ρ_A and Δ_A to arbitrary accuracy. Such estimates provide us some crude and global information about the true original dictionary. Now sum the probability matrix only over the rows accordingly, the expectation can be factorized as $\rho_A \rho^\top + \Delta_A \Delta^\top$. Therefore, given accurate estimates of ρ_A and Δ_A , obtaining refined estimation $\widehat{\rho}$ and $\widehat{\Delta}$ is as simple as solving a least square problem.

Definition 2.3 (Anchor partition). *Consider a partition of the vocabulary $[M]$ into $(\mathcal{A}, \mathcal{A}^c)$. denote $\rho_{\mathcal{A}} = \sum_{i \in \mathcal{A}} \rho_i$ and $\Delta_{\mathcal{A}} = \sum_{i \in \mathcal{A}} \Delta_i$. We call it an anchor partition if for some constant $C_A = \Omega(1)$,*

$$\text{cond} \left(\begin{bmatrix} \rho_{\mathcal{A}} & \Delta_{\mathcal{A}} \\ 1 - \rho_{\mathcal{A}} & -\Delta_{\mathcal{A}} \end{bmatrix} \right) \leq C_A. \quad (2.28)$$

If the dictionary is well separated $\|\Delta\|_1 = \Omega(1)$, it is feasible to find an anchor partition. Moreover, we will show that we can use the estimator $\widehat{\Delta}$ obtained in Phase I to construct such an anchor partition easily. The next lemma states a sufficient condition for constructing an anchor partition.

Lemma 2.8 (Sufficient condition for constructing an anchor partition). *Let $\Delta_{\mathcal{I}}$ be the vector of Δ restricted to a set of words \mathcal{I} . Suppose that $\|\Delta_{\mathcal{I}}\|_1 \geq C\|\Delta\|_1$ for some constant $C = \Omega(1)$, and that for some constant $C' \leq \frac{1}{3}C$, we can estimate $\Delta_{\mathcal{I}}$ up to precision:*

$$\|\widehat{\Delta}_{\mathcal{I}} - \Delta_{\mathcal{I}}\|_1 \leq C'\|\Delta_{\mathcal{I}}\|_1. \quad (2.29)$$

Denote $\widehat{\mathcal{A}} = \{i \in \mathcal{I} : \widehat{\Delta}_i > 0\}$. We have that $(\widehat{\mathcal{A}}, \mathcal{M} \setminus \widehat{\mathcal{A}})$ forms an anchor partition defined in 2.3.

Definition 2.4 (Good bins). *Denote the dictionary separation restricted to the “good words” in each empirical bin $\widehat{\mathcal{I}}_k$ by:*

$$S_k =: \sum_{i \in \widehat{\mathcal{I}}_k} |\Delta_i| = \|\widetilde{\Delta}_k\|_1. \quad (2.30)$$

Fix constants $C_1 = C_2 = \frac{1}{24}\|\Delta\|_1 = \Omega(1)$. We call bin $\widehat{\mathcal{I}}_k$ a “good bin” if it satisfies that:

1. the marginal probability of the bin $W_k \geq C_1 e^{-k}$.
2. the ratio between the separation and the marginal probability of the bin satisfies $\frac{S_k}{2W_k} \geq C_2$.

Let \mathcal{G} denote the set of all the good bins. Next lemma shows that a constant fraction of total probability mass is contained in good bins.

Lemma 2.9 (Total mass in good bins). *With high probability, we can bound the total marginal probability mass in the “good bins” by:*

$$\sum_{k \in \mathcal{G}} W_k \geq \|\Delta\|_1 / 12. \quad (2.31)$$

This implies a bound of total separation contained in all the good words of the good bins:

$$\sum_{i \in \widehat{\mathcal{I}}_k, k \in \mathcal{G}} |\Delta_i| = \sum_{k \in \mathcal{G}} S_k \geq 2C_2 \sum_{k \in \mathcal{G}} W_k \geq \frac{1}{24}(\|\Delta\|_1)^2 = \Omega(1). \quad (2.32)$$

Lemma 2.10 (Estimate the separation vector restricted to good bins). *If the empirical bin $\widehat{\mathcal{I}}_k$ is a good bin, with high probability, the estimate $\widehat{\Delta}_{\widehat{\mathcal{I}}_k}$ from Phase I (Lemma 2.6), for the separation vector restricted to the bin satisfies:*

$$\|\widehat{\Delta}_{\widehat{\mathcal{I}}_k} - \widetilde{\Delta}_k\|_1 \leq \frac{1}{\sqrt{d_0}} \|\Delta_{\widehat{\mathcal{I}}_k}\|_1. \quad (2.33)$$

The above two lemmas suggest that we can focus on the “good words” in the “good bins”, namely $\mathcal{I} = \cup_{k \in \mathcal{G}} \widehat{\mathcal{L}}_k$. Lemma 2.9 showed the separation contained in \mathcal{I} is at least $\sum_{k \in \mathcal{G}} S_k = C \|\Delta\|_1$ for some $C = \Omega(1)$; Lemma 2.10 showed that with linear number of samples we can estimate Δ restricted to \mathcal{I} up to constant accuracy. Therefore by Lemma 2.8 we can construct a valid anchor partition $(\mathcal{A}, \mathcal{M} \setminus \mathcal{A})$ by setting: $\mathcal{A} = \{i : \widehat{\Delta}_i > 0, \text{ for } i \in \widehat{\mathcal{I}}_k, k \in \mathcal{G}\}$.

Ideally, we want to restrict to the “good words” and set the anchor partition to be $\{i : \widehat{\Delta}_i > 0, \text{ for } i \in \widetilde{\mathcal{L}}_k, k \in \mathcal{G}\}$, but we cannot distinguish the “good words” from spillover words. However, the bound on the total marginal of spillover $\sum_k \overline{W}_k = O(e^{-d_0/2})$ guarantees that even if we mis-classify all the spillover words, the construction is still a valid anchor partition.

Estimate the anchor matrix Given the two superwords $(\mathcal{A}, \mathcal{M} \setminus \mathcal{A})$ from the anchor partition, define the 2×2 matrix $D_{\mathcal{A}} = \begin{bmatrix} \rho_{\mathcal{A}}, & \Delta_{\mathcal{A}} \\ 1 - \rho_{\mathcal{A}}, & -\Delta_{\mathcal{A}} \end{bmatrix}$ to be the anchor matrix. To estimate the two scalars $\rho_{\mathcal{A}}$ and $\Delta_{\mathcal{A}}$, we apply the standard concentration bound and argue that with high probability,

$$\left\| \begin{bmatrix} \sum_{i \in \mathcal{A}, j \in \mathcal{A}} [B_N]_{i,j} & \sum_{i \in \mathcal{A}, j \in \mathcal{A}^c} [B_N]_{i,j} \\ \sum_{i \in \mathcal{A}^c, j \in \mathcal{A}} [B_N]_{i,j} & \sum_{i \in \mathcal{A}^c, j \in \mathcal{A}^c} [B_N]_{i,j} \end{bmatrix} - D_{\mathcal{A}} D_{\mathcal{A}}^{\top} \right\| = O\left(\frac{1}{\sqrt{N}}\right)$$

Recall that by anchor partition, we have $|\Delta_{\mathcal{A}}| = \Omega(1)$. Thus we can estimate $\rho_{\mathcal{A}}$ and $\Delta_{\mathcal{A}}$ to accuracy $\frac{1}{\sqrt{N}}$. Since $N = \Omega(M)$ is asymptotically large, we essentially obtain precisely the anchor matrix $D_{\mathcal{A}}$.

Use anchor matrix to refine estimation Now given an anchor partition of the vocabulary $(\mathcal{A}, \mathcal{A}^c)$, and given the exact anchor matrix $D_{\mathcal{A}}$ which has $\Omega(1)$ condition number, refining the estimation of ρ_i and Δ_i for each i is very easy and achieves optimal rate.

Lemma 2.11 (Refine estimation). *With probability at least $1 - \delta$, Phase II of Algorithm 3 outputs estimates $\hat{\rho}$ and $\hat{\Delta}$ such that*

$$\|\hat{\rho} - \rho\| < \sqrt{1/\delta N}, \quad \|\hat{\Delta} - \Delta\| < \sqrt{1/\delta N}.$$

The above lemma implies the ℓ_1 norm accuracy for Theorem 2.6:

$$\|\hat{\rho} - \rho\|_1 < \sqrt{M/\delta N}, \quad \|\hat{\Delta} - \Delta\|_1 < \sqrt{M/\delta N}.$$

2.5 Details of Rank R Algorithm

In this section, we examine each step of Algorithm 4 to prove Theorem 2.2 and Theorem 2.3. Recall that we are given 4 independent batches of N samples, with which we construct 4 independent empirical bigram matrix $B = \mathbb{B} + E$ where the noise matrix E are independent and identical copies of sampling noise E . In each of the 4 steps of the algorithm, an independent and identical copy of the bigram matrix is used. We omit the index $i = 1, \dots, 4$ for notation brevity.

2.5.1 Binning

We focus on the symmetric case where the rank R probability matrix is parameterized as $\mathbb{B} = PWP^T$, and W is a PSD matrix. The algorithm and analysis can be easily extended to deal with more general case. We define the weight $w_r = \sum_i W_{r,i}$. We assume that the weights are lower bounded by

$$w_{min} = \min_r w_r \geq C_w = \Omega(1).$$

The marginal probability is given by

$$\rho_i = \sum_k \mathbb{B}_{i,k} = \sum_r w_r p_i^{(r)}.$$

Note that each entry of the probability matrix \mathbb{B} can be bounded by the product of the corresponding marginal probability as below:

$$\mathbb{B}_{i,j} = \sum_{s,t} W_{s,t} p_i^{(s)} p_j^{(t)} \leq \sum_{s,t} W_{s,t} p_i^{(s)} \sum_{t'} p_j^{(t')} = \rho_i \sum_{t'} p_j^{(t')} \leq \frac{1}{w_{\min}} \rho_i \rho_j. \quad (2.34)$$

Again, binning according to the empirical marginal is given by:

$$\widehat{\mathcal{I}}_0 = \left\{ i : \frac{\epsilon_0}{M} \leq \widehat{\rho}_i < \frac{1}{M} \right\}, \quad \widehat{\mathcal{I}}_k = \left\{ i : \frac{e^{k-1}}{M} \leq \widehat{\rho}_i < \frac{e^k}{M} \right\}, \text{ for } k = 1 : \log M.$$

Let $M_k = |\widehat{\mathcal{I}}_k|$ denote the number of words in bin $\widehat{\mathcal{I}}_k$.

The grouping of words according to the exact marginal probabilities is defined as:

$$\mathcal{I}_0 = \left\{ i : \frac{\epsilon_0}{M} \leq \rho_i < \frac{1}{M} \right\}, \quad \mathcal{I}_k = \left\{ i : \frac{e^{k-1}}{M} \leq \rho_i < \frac{e^k}{M} \right\}, \text{ for } k = 1 : \log M.$$

Define $\bar{\rho}_k$ to be the typical marginal of a word in bin \mathcal{I}_k :

$$\bar{\rho}_k = e^k / M.$$

For $i, j \in \mathcal{I}_k$, we have $\mathbb{B}_{i,j} \leq \bar{\rho}_k^2 / w_{\min}$.

Due to the statistical noise of sampling, $\widehat{\mathcal{I}}_k$ may contain words whose exact marginal is much larger than $\bar{\rho}_k$. The next lemma argues that such spillover effect is small.

Lemma 2.12 (Spillover from heavier bins is small). *With high probability, for all empirical bins $\widehat{\mathcal{I}}_k$, we can bound the spillover probability from much heavier bins by:*

$$\overline{W}_k := \sum_{i \in \mathcal{I}_{k'} : k' > k + \tau} \rho_i \leq 2e^{-e^{\tau+k} d_0 / 2}. \quad (2.35)$$

Definition 2.5 (Big bin). *An empirical bin $\widehat{\mathcal{I}}_k$ is a big bin if*

$$W_k = \sum_{j \in \widehat{\mathcal{I}}_k} \rho_j > e^{-k}. \quad (2.36)$$

We know that a constant fraction of all the probability mass lies in such big bins. Moreover for $d_0 \gg 1$, we have $W_k > e^{-k} \gg 2e^{-(k+\tau)d_0/2} \geq \overline{W}_k$.

Lemma 2.13 (Escaped probability mass). *With high probability, for all big bins, the mass that escapes from the bin is bounded by*

$$W_k^s =: \sum_{i \in \mathcal{I}_k, i \notin \widehat{\mathcal{I}}_{k'} \text{ for } |k-k'| < \tau} \rho_i \leq 4W_k e^{-e^{k+\tau} d_0/2}.$$

2.5.2 Spectral concentration in diagonal blocks

Define the regularized probability matrix $\widetilde{\mathbb{B}}$ by setting the rows/columns corresponding to spillover words from much heavier bins to 0:

$$\widetilde{\mathbb{B}} = \text{Diag}(\mathbf{1}[\rho_i < 2\bar{\rho}_k]) \mathbb{B} \text{Diag}(\mathbf{1}[\rho_i < 2\bar{\rho}_k]) \quad (2.37)$$

Under the assumption that W is a PSD matrix, we define the $M \times R$ matrix \mathbb{B}^{sqr} and $\widetilde{\mathbb{B}}^{sqr}$ to be:

$$\mathbb{B}^{sqr} = P W^{1/2}, \quad \text{and} \quad \widetilde{\mathbb{B}}^{sqr} = \text{Diag}(\mathbf{1}[\rho_i < 2\bar{\rho}_k]) P W^{1/2}. \quad (2.38)$$

Consider the diagonal block corresponding to the k -th empirical bin

$$B_k = [B]_{\widehat{\mathcal{I}}_k \times \widehat{\mathcal{I}}_k}.$$

Similarly, we define the $M_k \times R$ matrix restricting bin $\widehat{\mathcal{I}}_k$ as:

$$\mathbb{B}_{\widehat{\mathcal{I}}_k}^{sqr} = P_{\widehat{\mathcal{I}}_k} W^{1/2}, \quad \text{and} \quad \widetilde{\mathbb{B}}_{\widehat{\mathcal{I}}_k}^{sqr} = \text{Diag}(\mathbf{1}[\rho_i < 2\bar{\rho}_k]) P_{\widehat{\mathcal{I}}_k} W^{1/2},$$

We argue that, given $N = \Omega(MR^2)$, for each diagonal block B_k , Step 2 of Algorithm 4

finds a subspace \widehat{V}_k correlated with $\widetilde{\mathbb{B}}_k^{sgrt}$. We can bound $\|\text{Proj}_{\widehat{V}_k} \widetilde{\mathbb{B}}_{\mathcal{I}_k}^{sgrt} - \widetilde{\mathbb{B}}_{\mathcal{I}_k}^{sgrt}\|$ up to constant accuracy.

Definition 2.6 (Expected row sum in diagonal blocks). *Recall that for $i, j \in \mathcal{I}_k$, we have $\mathbb{B}_{i,j} \leq \bar{\rho}_k^2/w_{min}$. Define the maximal expected row sum of the diagonal block B_k to be:*

$$d_k^{max} = M_k \bar{\rho}_k^2 / w_{min}. \quad (2.39)$$

Note that in the particular parameterization for the problem community detection with uniform marginal, we can simply define d_k^{max} to be $1/M$ and thus get rid of the w_{min}^{-1} dependence, and the rest of the analysis follows to recover the sample complexity result of $N = \Theta(MR^2)$ in the community detection problem with R communities.

Lemma 2.14 (Spectral concentration in each diagonal block). *Regularize the k -th diagonal block B_k by removing the rows/columns with sum larger than $2d_k^{max}$. Run rank R truncated SVD on the regularized block \widetilde{B}_k . Let the columns of the $M_k \times R$ matrix \widehat{V}_k be the leading R singular vectors. Define $\text{Proj}_{\widehat{V}_k} = \widehat{V}_k \widehat{V}_k^\top$. We have*

$$\|\text{Proj}_{\widehat{V}_k} \widetilde{\mathbb{B}}_{\mathcal{I}_k}^{sgrt} - \widetilde{\mathbb{B}}_{\mathcal{I}_k}^{sgrt}\| = O\left(\frac{\sqrt{N d_k^{max} \log N d_k^{max}}}{N}\right)^{1/2}. \quad (2.40)$$

2.5.3 Low rank projection

In Step 3 of Algorithm 4, we “stitch” the subspaces \widehat{V}_k for each bin $\widehat{\mathcal{I}}_k$ learned in Step 2 to get an estimate for the column span of the entire matrix.

Define the diagonal matrix D_S of dimension $M \times M$ to be:

$$D_S = \begin{bmatrix} \bar{\rho}_1^{-1/2} I_{\widehat{M}_1} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \bar{\rho}_{\log M}^{-1/2} I_{\widehat{M}_{\log M}} \end{bmatrix}. \quad (2.41)$$

Define $\text{Proj}_{\widehat{V}}$ to be the block diagonal projection matrix which projects an $M \times M$

matrix to a subspace \widehat{V} of dimension at most $R \log M$:

$$\text{Proj}_{\widehat{V}} = \begin{bmatrix} \text{Proj}_{\widehat{V}_1} & & \\ & \ddots & \\ & & \text{Proj}_{\widehat{V}_{\log M}} \end{bmatrix}. \quad (2.42)$$

Now consider the empirical average bigram B with the 3rd batch of samples. We regularize the entire matrix in the following way. For each row in B , if the word i is in bin $\widehat{\mathcal{L}}_k$ as defined in Step 1, and if the corresponding row has sum larger than $2\bar{\rho}_k$, we set the entire row and column to zero. Let \widetilde{B} denote the regularized matrix.

Lemma 2.15. *Let \widehat{B}_1 denote the rank R truncated SVD of $\text{Proj}_{\widehat{V}} D_S \widetilde{B} D_S \text{Proj}_{\widehat{V}}$. With large probability, we can bound the spectral distance between \widehat{B}_1 and $D_S \widetilde{B} D_S$ by:*

$$\|\widehat{B}_1 - D_S \widetilde{B} D_S\| = O\left(\frac{\log(Nw_{\min}^2/M)}{Nw_{\min}^2/M}\right)^{1/4}. \quad (2.43)$$

Lemma 2.16. *Let $\widehat{B}_2 = D_S^{-1} \widehat{B}_1 D_S^{-1}$, we can get the ℓ_1 error bound as:*

$$\|\widehat{B}_2 - \mathbb{B}\|_1 = O\left(\frac{\log(Nw_{\min}^2/M)}{Nw_{\min}^2/MR^2}\right)^{1/4}.$$

So far we have proved Theorem 2.2 for Algorithm 4.

2.5.4 Refinement

For a given PSD matrix $X = UU^\top$, whose SVD is given by $X = V\Sigma V^\top$, we define $X^{1/2}$ to be $X^{1/2} = V\Sigma^{1/2}$. Note that $V = UH$ for some unknown rotation matrix H .

Lemma 2.17 (Refinement with separation). *Recall that initialization \widehat{B}_1 obtained from Step 3 such that $\|\widehat{B}_1 - D_S \widetilde{B} D_S\| \leq (w_{\max} MR^2 / w_{\min} N)^{1/4}$. Assume that $\sigma_{\min}(D_S \widetilde{B} D_S) > (w_{\max}^2 MR^2 / w_{\min}^2 N)^{1/4}$.*

Let \widehat{V} denote the R leading left singular vectors of \widehat{B}_1 . Regularize B from the 4-th

batch of samples in the same way as in Step 3. Set

$$Y = (\widehat{V}^\top D_S \widetilde{B} D_S \widehat{V})^{-1/2} (\widehat{V}^\top D_S \widetilde{B} D_S).$$

Let $\widehat{B}_3 = Y^\top Y$ and $\widehat{B} = D_S^{-1} \widehat{B}_3 D_S^{-1}$. We can bound the spectral distance by

$$\|\widehat{B}_3 - D_S \widetilde{B} D_S\|_F = O\left(\sqrt{\frac{MR}{N}}\right), \quad (2.44)$$

$$\|\widehat{B}_4 - \mathbb{B}\|_1 = O\left(\sqrt{\frac{MR}{N}}\right). \quad (2.45)$$

2.6 Sample complexity lower bounds

Lower bound for estimating probabilities

We reduce the estimation problem to the community detection for a specific set of model parameters.

Consider the following topic model with equal mixing weights, i.e. $w = w^c = 1/2$. For some constant $C_\Delta = \Omega(1)$, the two word distributions are given by:

$$p = \left[\frac{1 + C_\Delta}{M}, \dots, \frac{1 + C_\Delta}{M}, \frac{1 - C_\Delta}{M}, \dots, \frac{1 - C_\Delta}{M} \right],$$

$$q = \left[\frac{1 - C_\Delta}{M}, \dots, \frac{1 - C_\Delta}{M}, \frac{1 + C_\Delta}{M}, \dots, \frac{1 + C_\Delta}{M} \right].$$

The expectation of the sum of samples is given by

$$\mathbb{E}[B_N] = N \frac{1}{2} (pp^\top + qq^\top) = \frac{N}{M^2} \begin{bmatrix} 1 + C_\Delta^2 & 1 - C_\Delta^2 \\ 1 - C_\Delta^2 & 1 + C_\Delta^2 \end{bmatrix}.$$

Note that the expected row sum is in the order of $\Omega(\frac{N}{M})$. When N is small, with high probability the entries of the empirical sum B_N only take value either 0 or 1, and B_N approximately corresponds to a SBM ($G(M, a/M, b/M)$) with parameter $a = \frac{N}{M}(1 + C_\Delta^2)$ and $b = \frac{N}{M}(1 - C_\Delta^2)$.

If the number of sample document is large enough for any algorithm to estimating

the dictionary vector p and q up to ℓ_1 accuracy ϵ for a small constant ϵ , it can then be used to achieve partial recovery in the corresponding SBM, namely correctly classify a γ proportion of all the nodes for some constant $\gamma = \frac{\epsilon}{C_\Delta}$.

According to Zhang & Zhou [127], there is a universal constant $C > 0$ such that if $(a - b)^2/(a + b) < c \log(1/\gamma)$, then there is no algorithm that can recover a γ -correct partition in expectation. This suggests that a necessary condition for us to learn the distributions is that

$$\frac{(2(N/M)C_\Delta^2)^2}{2(N/M)} \geq c \log(C_\Delta/\epsilon),$$

namely $(N/M) \geq c \log(C_\Delta/\epsilon)/2C_\Delta^4$. In the well separated regime, this means that the sample complexity is at least linear in the vocabulary size M .

Note that this lower bound is in a sense a worst case constructed with a particular distribution of p and q , and for other choices of p and q it is possible that the sample complexity can be much lower than that $\Omega(M)$.

Lower bound for testing property of HMMs

In this section, we prove an information theoretic lower bound for testing whether a sequence of observations consists of independent draws from $Unif[M]$, versus being a sequence of observations generated by a 2-state HMM with observation distributions supported on $\{1, \dots, M\}$. Such a lower bound will immediately yield a lower bound for estimating various properties of HMMs, including estimating the entropy rate, as a sequence of independent draws from $Unif[M]$ has entropy rate $\log(M)$, whereas the 2-state HMMs we consider have an entropy rate that is an additive constant lower. We note that this HMM lower bound is significantly stronger than the analogous task of testing whether a matrix of probabilities has rank 1 versus rank 2. Such a task corresponds to only using the bi-gram counts extracted from the sequence of observations. It is conceivable that by leveraging longer sequences (i.e. k -grams for $k > 2$), more information can be extracted about the instance. While this is the case, as our lower bound shows, even with such information, $\Theta(M)$ observations are

required to perform this test and distinguish these two cases.

Theorem 2.7 (Theorem 2.4 restated). *Consider a sequence observations from a HMM with two hidden states $\{s_p, s_q\}$, emission distributions p, q supported on M elements, and probability $t = \Omega(1)$ of transitioning from s_p to s_q and from s_q to s_p . For sufficiently large M , given a sequence of N observations for $N = o(M)$, it is information theoretically impossible to distinguish the case that the two emission distributions are well separated, i.e. $\|p - q\|_1 \geq 1/2$, from the case that both p and q are uniform distribution over $[M]$, namely the HMM is degenerate of rank 1.*

In order to derive a lower bound for the sample complexity, it suffices to show that given a sequence of $N = o(M)$ consecutive observations, one can not distinguish whether it is generated by a random instance from a class of 2-state HMMs (Definition 2.2) with well-separated emission distribution p and q , or the sequence is simply N i.i.d. samples from the uniform distribution over \mathcal{M} , namely a degenerate HMM with $p = q$.

We shall focus on a class of well-separated HMMs parameterized as below: a symmetric transition matrix $T = \begin{bmatrix} 1-t & t \\ t & 1-t \end{bmatrix}$, where we set the transition probability to $t = 1/4$; the initial state distribution is $\pi_p = \pi_q = 1/2$ over the two states s_p and s_q ; the corresponding emission distribution p and q are uniform over two disjoint subsets of the vocabulary, \mathcal{A} and $\mathcal{M} \setminus \mathcal{A}$, separately. Moreover, we treat the set \mathcal{A} as a random variable, which can be any of the $\binom{M}{M/2}$ subsets of the vocabulary of size $M/2$, chosen with equal probability $1/\binom{M}{M/2}$. Note that there is a one to one mapping between the set \mathcal{A} and an instance in the class of well-separated HMM.

Now consider a random sequence of N words $G_1^N = [g_1, \dots, g_N] \in \mathcal{M}^N$. If this sequence is generated by an instance of the 2-state HMM denoted by \mathcal{A} , the joint probability of (G_1^N, \mathcal{A}) is given by:

$$\Pr_2(G_1^N, \mathcal{A}) = \Pr_2(G_1^N | \mathcal{A}) \Pr_2(\mathcal{A}) = \Pr_2(G_1^N | \mathcal{A}) \frac{1}{\binom{M}{M/2}} \quad (2.46)$$

Moreover, given \mathcal{A} , since the support of p and q are disjoint over \mathcal{A} and $\mathcal{M} \setminus \mathcal{A}$ by

our assumption, we can perfectly infer the sequence of hidden states $S_1^N(G_1^N, \mathcal{A}) = [s_1, \dots, s_N] \in \{s_p, s_q\}^N$ simply by the rule $s_i = s_p$ if $g_i \in \mathcal{A}$ and $s_i = s_q$ otherwise. Thus we have:

$$\Pr_2(G_1^N | \mathcal{A}) = \Pr_2(G_1^N, S_1^N | \mathcal{A}) = \frac{1/2}{M/2} \prod_{i=2}^N \frac{(1-t)\mathbf{1}[s_i = s_{i-1}] + t\mathbf{1}[s_i \neq s_{i-1}]}{M/2}. \quad (2.47)$$

On the other hand, if the sequence G_1^N is simply i.i.d. samples from the uniform distribution over \mathcal{M} , its probability is given by

$$\Pr_1(G_1^N) = \frac{1}{M^N}. \quad (2.48)$$

We further define a joint distribution rule $\Pr_1(G_1^N, \mathcal{A})$ such that the marginal probability agrees with $\Pr_1(G_1^N)$. In particular, we define:

$$\Pr_1(G_1^N, \mathcal{A}) = \Pr_1(\mathcal{A} | G_1^N) \Pr_1(G_1^N) \equiv \frac{\Pr_2(G_1^N | \mathcal{A})}{\sum_{\mathcal{B} \in \binom{\mathcal{M}}{M/2}} \Pr_2(G_1^N | \mathcal{B})} \Pr_1(G_1^N), \quad (2.49)$$

where we define the conditional probability $\Pr_1(\mathcal{A} | G_1^N)$ using the properties of the 2-state HMM class.

The main idea of the proof of Theorem 2.7 is to show that if $N = o(M)$, the total variation distance between \Pr_1 and \Pr_2 vanishes to zero. It follows immediately from the connection between the error bound of hypothesis testing and total variation distance between two probability rules, that if $TV(\Pr_1(G_1^N), \Pr_2(G_1^N))$ is too small we are not able to test which probability rule the random sequence G_1^N is generated according to.

The detailed proofs are provided in Appendix 2.7.4.

As an immediate corollary of this theorem, it follows that many natural properties of HMMs cannot be estimated using a sublinear length sequence of observations:

Corollary 2.4. *For HMMs with 2 states and emission distributions supported on a domain of size at most M , to estimate the entropy rate up to an additive constant $c \leq 1$ requires a sequence of $\Omega(M)$ observations.*

2.7 Proofs for Chapter 2

2.7.1 Proofs for Rank 2 Algorithm, Phase I

Proof. (to Lemma 2.1 (Estimate the word marginal probability ρ))

We analyze how accurate the empirical average $\hat{\rho}$ is. Note that under the assumption of Poisson number of samples, we have $\hat{\rho}_i \sim \frac{1}{N}\text{Poi}(N\rho_i)$, and $\text{Var}(\hat{\rho}_i) = \frac{1}{N}\rho_i$. Apply Markov inequality:

$$\Pr\left(\sum_{i=1}^M \left|\frac{\hat{\rho}_i - \rho_i}{\sqrt{\rho_i}}\right|^2 > t\right) \leq \frac{M}{tN},$$

thus probability at least $1 - \delta$, we can bound

$$\sum_{i=1}^M \left|\frac{\hat{\rho}_i - \rho_i}{\sqrt{\rho_i}}\right|^2 \leq \frac{M}{N\delta}. \quad (2.50)$$

Then apply Cauchy-Schwartz, we have

$$\sum_{i=1}^M |\hat{\rho}_i - \rho_i| \leq \left(\sum_{i=1}^M \sqrt{\rho_i}^2 \sum_{i=1}^M \left|\frac{\hat{\rho}_i - \rho_i}{\sqrt{\rho_i}}\right|^2\right)^{1/2} \leq \frac{1}{\sqrt{d_0\delta}}.$$

□

Proof. (to Lemma 2.2 (Concentration of marginal probabilities in the heaviest bin))

Fix constants $C_1 = \frac{1}{2}$ and $C_2 = 2$, apply Corollary 2.5 of Poisson tail (note that for word in the heaviest bin, we have $N\rho_i > d_0 \log M$ to be a super constant), we show that $\hat{\rho}_i$ concentrates well:

$$\Pr(C_1 N\rho_i < \text{Poi}(N\rho_i) < C_2 N\rho_i) \geq 1 - 4e^{-N\rho_i/2} \geq 1 - 4e^{-N \log M/(2M)}.$$

Note that the number of words in the heaviest bin is upper bounded by $M_{\log} \leq \frac{1}{\min_{i \in \mathcal{I}_{\log}} \rho_i} \leq \frac{M}{\log(M)}$. Take a union bound, we have that with high probability, all the

estimates $\hat{\rho}_i$'s in the heaviest bin concentrate well:

$$\begin{aligned}
\Pr(\forall i \in \mathcal{I}_{\log} : C_1 \rho_i < \hat{\rho}_i < C_2 \rho_i) &\geq 1 - \frac{M}{\log M} e^{-N \log M / (2M)} \\
&\geq 1 - 4e^{-N \log M / (2M) + \log M - \log \log M} \\
&\geq 1 - M^{-(d_0/2-1)} \\
&\geq 1 - M^{-1},
\end{aligned}$$

where recall that $d_0 = N/M$ is a large constant. \square

Proof. (to Lemma 2.3 (Estimate the dictionary separation restricted to the empirical heaviest bin))

(1) First, we claim that with high probability, no word from \mathcal{I}_k for $k \leq \log(M) - e^2$ is placed in $\hat{\mathcal{I}}_{\log M}$. Namely all the words in $\hat{\mathcal{I}}_{\log M}$ have true marginals at least $\Omega(\frac{\log M}{M})$. This is easy to show, by the Corollary 2.5 of Poisson tail bound, each of the word from the much lighter bins is placed in $\hat{\mathcal{I}}_{\log M}$ with probability less than $2e^{-N \log M / M}$. Take a union bound over all words with marginal at least $1/M$, we can bound the probability that any of the words being placed in $\hat{\mathcal{I}}_{\log M}$ by $2Me^{-d_0 \log M} = O(M^{-d_0+1})$.

(2) The appropriate scaling with the diagonal matrix $\text{diag}(\hat{\rho}_{\hat{\mathcal{I}}_{\log}})^{-1/2}$ on both sides of the diagonal block is very important, which allows us to apply matrix Bernstein inequality at a sharper rate.

Note that with the two independent batches of samples, the empirical count matrix B considered here is independent from the empirical marginal vector $\hat{\rho}$. Thus for every fixed realization of $\hat{\rho}$, we have that with probability at least $1 - M^{-1}$,

$$\begin{aligned}
\|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} (B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} - \mathbb{B}_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}}) \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2}\|_2 &\leq \sqrt{\frac{\log(M_{\log}/\delta)}{N} + \frac{\frac{M}{\log(M)} \log(M_{\log}/\delta)}{N}} \\
&= O\left(\frac{M}{N} \left(1 + \frac{\log(1/\delta)}{\log(M)}\right)\right) \\
&= O\left(\frac{M}{N}\right),
\end{aligned}$$

where we used the fact that the all the marginals in the heaviest bin can be estimated

with constant multiplicative accuracy given by Lemma 2.2; also, note that compared to the Bernstein matrix inequality directly applied to the entire matrix as in (2.73), here with the proper scaling we have $Var \leq 1$ and $B \leq \frac{M}{\log(M)}$, since $\hat{\rho}_i > \log(M)/M$ for all $i \in \hat{\mathcal{I}}_{\log}$.

We will show that $B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}}$, the diagonal block of the empirical count matrix, concentrates well enough to ensure that we can estimate the separation restricted to the heaviest bin by the leading eigenvector of $(B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} - \hat{\rho}_{\hat{\mathcal{I}}_{\log}} \hat{\rho}_{\hat{\mathcal{I}}_{\log}}^\top)$. Note that

$$\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \mathbb{B}_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} = \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \rho_{\hat{\mathcal{I}}_{\log}} (\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \rho_{\hat{\mathcal{I}}_{\log}})^\top + \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta'_{\hat{\mathcal{I}}_{\log}} (\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta_{\hat{\mathcal{I}}_{\log}})^\top.$$

Apply triangle inequality we have

$$\begin{aligned} & \left\| \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} (B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} - \hat{\rho}_{\hat{\mathcal{I}}_{\log}} \hat{\rho}_{\hat{\mathcal{I}}_{\log}}^\top) \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} - \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta'_{\hat{\mathcal{I}}_{\log}} (\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta_{\hat{\mathcal{I}}_{\log}})^\top \right\|_2 \\ & \leq \left\| \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} (B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} - \mathbb{B}_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}}) \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \right\|_2 + \left\| \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} (\hat{\rho}_{\hat{\mathcal{I}}_{\log}} \hat{\rho}_{\hat{\mathcal{I}}_{\log}}^\top - \rho_{\hat{\mathcal{I}}_{\log}} \rho_{\hat{\mathcal{I}}_{\log}}^\top) \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \right\|_2 \\ & = O\left(\frac{M}{N}\right) + \sqrt{\frac{M}{N}} \\ & = O\left(\sqrt{\frac{M}{N}}\right). \end{aligned}$$

(3) Let uu^\top be the rank-1 truncated SVD of $\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} (B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} - \hat{\rho}_{\hat{\mathcal{I}}_{\log}} \hat{\rho}_{\hat{\mathcal{I}}_{\log}}^\top) \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2}$. Let $v = \hat{D}_{\hat{\mathcal{I}}_{\log}}^{1/2} u$ be our estimate for $\Delta_{\hat{\mathcal{I}}_{\log}}$. Apply Wedin's theorem to rank-1 matrix (Lemma 2.22), we can bound the distance between vector u and $\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta_{\hat{\mathcal{I}}_{\log}}$ by:

$$\min\{\|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta_{\hat{\mathcal{I}}_{\log}} - u\|_2, \|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta_{\hat{\mathcal{I}}_{\log}} + u\|_2\} = O\left(\min\left\{\frac{(M/N)^{1/2}}{\|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta_{\hat{\mathcal{I}}_{\log}}\|_2}, (M/N)^{1/4}\right\}\right).$$

Note that $\|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{1/2} \mathbf{1}\|_2 = \|\hat{\rho}_{\hat{\mathcal{I}}_{\log}}^{1/2}\|_2 = 1$. Apply Cauchy-Schwartz, for any vector x , we have

$$\|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} x\|_2 \geq \|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} x\|_2 \|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{1/2} \mathbf{1}\|_2 \geq \left| \langle \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} x, \hat{D}_{\hat{\mathcal{I}}_{\log}}^{1/2} \mathbf{1} \rangle \right| = \|x\|_1,$$

therefore, we can bound

$$\min\{\|\Delta_{\widehat{\mathcal{I}}_{\log}} - v\|_1, \|\Delta_{\widehat{\mathcal{I}}_{\log}} + v\|_1\} \leq O\left(\min\left\{\frac{(M/N)^{1/2}}{\|\Delta_{\widehat{\mathcal{I}}_{\log}}\|_1}, (M/N)^{1/4}\right\}\right).$$

In the above inequalities we absorb all the universal constants and focus on the scaling factors.

□

Proof. (to Lemma 2.4 (Spillover from much heavier bins is small in all bins))

Define $d_k = e^k d_0$, which is not to be confused with d_k^{\max} defined in (2.18).

(1) Consider $k' = k + \tau$. The probability that a word i from $\mathcal{I}_{k'}$ falls into $\widehat{\mathcal{I}}_k$ is bounded by:

$$\Pr\left(N\frac{e^{k-1}}{M} < \text{Poi}(N\rho_i) < N\frac{e^k}{M}\right) < \Pr\left(\text{Poi}\left(N\frac{e^{(\tau+k)}}{M}\right) < N\frac{e^k}{M}\right) \leq 2e^{-e^{\tau+k}d_0/2} \quad (2.51)$$

where we apply the Poisson tail bound (1) in Corollary 2.5, and set $c = e^{-\tau} < 1/2$ for $\tau \geq 1$. Note that this bound is doubly exponentially decreasing in τ and exponentially decreasing in d_0 .

In expectation, we can bound \overline{W}_k by:

$$\begin{aligned} \mathbb{E}\overline{W}_k &= \mathbb{E} \sum_{i \in \mathcal{M}} \rho_i \sum_{k': k' \geq k + \tau + 1} \mathbf{1}[i \in \mathcal{I}_{k'}] \Pr\left(N\frac{e^{k-1}}{M} < \text{Poi}(N\rho_{k'}) < N\frac{e^k}{M}\right) \\ &\leq \sum_{i \in \mathcal{M}} \rho_i \Pr\left(\text{Poi}\left(N\frac{e^{(\tau+k)}}{M}\right) < N\frac{e^k}{M}\right) \\ &\leq 2e^{-e^{\tau+k}d_0/2}. \end{aligned}$$

Similarly, apply the Poisson tail bound (2) in Corollary 2.5, and set $c' = e^\tau \geq e$ for $\tau \geq 1$, we can bound the probability with which a word i from much lighter bins,

namely $\cup_{\{k':k' < k-\tau\}} \mathcal{I}_{k'}$, is placed in the empirical bin $\widehat{\mathcal{I}}_k$ by:

$$\Pr\left(N\frac{e^k}{M} < \text{Poi}(N\rho_i) \leq N\frac{e^{k+1}}{M}\right) \leq \Pr\left(\text{Poi}\left(N\frac{e^{(k-\tau)}}{M}\right) > N\frac{e^k}{M}\right) \leq 2e^{-e^k d_0}, \quad (2.52)$$

and bound the total marginal probability by:

$$\mathbb{E}W_k \leq 2e^{-e^k d_0}.$$

(2) Next, we apply Bernstein's bound to get a high probability argument. We show that with high probability, for all the $\log \log(M)$ bins, we can bound the spillover probability mass by $\overline{W}_k \leq \mathbb{E}[\overline{W}_k] + O(\frac{1}{\text{poly}(M)})$, which implies that asymptotically as the vocabulary size $M \rightarrow \infty$, we have $\overline{W}_k \leq 2e^{-e^{\tau+k} d_0/2}$ for all k .

Consider the word i from the exact bin $\mathcal{I}_{k'}$, for some $k' \geq k + \tau$. Let

$$\lambda_i = 2e^{-e^{k'} d_0/2}$$

denote the upper bound (as shown in (2.51)) of the independent probability with which word i is placed in the empirical bin $\widehat{\mathcal{I}}_k$ (recall the Poisson number of samples assumption). The spillover probability mass is a random variable and can be written as

$$\overline{W}_k = \sum_{i \in \mathcal{I}_{k'}: (k+\tau) < k' \leq \log \log(M)} \rho_i \text{Ber}(\lambda_i),$$

Note that the summation of word i is over all the bin $\mathcal{I}_{k'}$ for $(k+\tau) \leq k' \leq \log \log(M)$, where recall that in Lemma 2.2 we showed that with high probability the heaviest words are retained in the empirical bin $\widehat{\mathcal{I}}_{\log}$. Apply Bernstein's inequality to bound \overline{W}_k :

$$\Pr(\overline{W}_k - \mathbb{E}\overline{W}_k > t) \leq e^{-\frac{t^2/2}{\sum_i \rho_i^2 \lambda_i + \max_i \rho_i t/3}}.$$

To ensure that the right hand side is bounded by $e^{-\log M}$ (this is to create space for

the union bound over the $\log \log M$ bins), we can fix some large universal constant C and set t to be

$$t = 2 \left(\left(\sum_i \rho_i^2 \lambda_i \right)^{1/2} + \max_i \rho_i \right) \log(M).$$

which right hand side can be bounded by:

$$\begin{aligned} \left(\sum_i \rho_i^2 \lambda_i \right)^{1/2} + \max_i \rho_i &\leq \left(\max_{i \in \mathcal{I}_{k'}: (k+\tau) \leq k' \leq \log \log(M)} \left(\frac{1}{\rho_i} \right) (\rho_i^2) (2e^{-e^{k'} d_0/2}) \right)^{1/2} + \frac{\log M}{M} \\ &\leq \left(2 \max_{(k+\tau) \leq k' \leq \log \log(M)} \frac{e^{k'}}{M} e^{-e^{k'} d_0/2} \right)^{1/2} + \frac{\log M}{M} \\ &\leq \frac{2e^{-e^{k+\tau} d_0/4}}{\sqrt{M}} + \frac{\log M}{M}, \end{aligned}$$

where the first inequality uses the worst case to bound the summation, and the last inequality uses the fact that $d_0 = \Omega(1)$ is a large constant. Therefore, we can set $t = 2 \left(\frac{e^{-e^{k+\tau} d_0/4} \log M}{\sqrt{M}} + \frac{(\log M)^2}{M} \right)$. Finally, take a union bound over at most $\log \log(M)$ moderate bins, we argue that with high probability (at least $1 - O(1/M)$), for all the empirical moderate bins, we can bound the spillover marginal by:

$$\overline{W}_k \leq \mathbb{E} \overline{W}_k + O\left(\frac{1}{\text{poly}(M)}\right).$$

(3) Moreover, assume that $W_k \geq e^{-k}$, we can bound the number of the heavy spillover words \overline{M}_k compared to number of words in the exact bin M_k .

First note that $\overline{M}_k \leq \frac{\overline{W}_k}{e^{\tau+k}/M}$. Recall that $d_k^{\max} = N W_k (e^{\tau+k}/M)$ was defined in (2.18). Also, since $W_k \geq e^{-k} \gg \overline{W}_k \approx e^{-e^{k+\tau} d_0}$, we can lower bound the number of words in the empirical bin $\widehat{\mathcal{L}}_1$ by:

$$M_k \geq \frac{W_k}{e^{k+\tau}/M}.$$

Thus we can bound

$$\begin{aligned}
\overline{M}_k \frac{d_k^{\max}}{M_k} &\leq \left(\frac{\overline{W}_k}{e^{k+\tau}/M} \right) \frac{(NW_k(e^{k+\tau}/M))}{\left(\frac{W_k}{e^{k+\tau}/M} \right)} \\
&\leq e^{k+\tau} d_0 \overline{W}_k \\
&\leq \frac{2e^{k+\tau} d_0}{e^{e^{k+\tau} d_0}} \\
&\leq 1,
\end{aligned}$$

where the second last inequality we used the high probability upper bound for \overline{W}_k , and in the last inequality we use the fact that $e^x > 2x$ for all x . \square

Proof. (of Lemma 2.5 (Concentration of the regularized diagonal block \tilde{B}_k .)

In Figure 2-2, the rows and the columns of $B_{\hat{\mathcal{I}}_k \times \hat{\mathcal{I}}_k}$ are sorted according to the exact marginal probabilities of the words in ascending order. The rows and columns that are set to 0 by regularization are shaded.

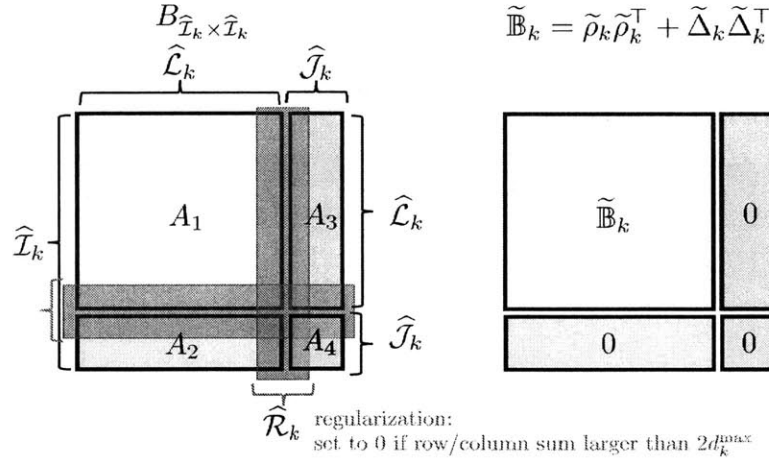


Figure 2-3: block decomposition of the diagonal block of B_{N^2} corresponding to $\hat{\mathcal{I}}_k$.

On the left hand side, it is the empirical matrix without regularization. We denote the removed elements by matrix $E \in \mathbb{R}_+^{M_k \times M_k}$, whose only nonzero entries are those that are removed from in the regularization step (in the strips with orange color), namely $E = [B_{i,j} \mathbf{1}[i \text{ or } j \in \hat{\mathcal{R}}_k]]$. We denote the retained elements by matrix $\tilde{B}_k = B_{\hat{\mathcal{I}}_k \times \hat{\mathcal{I}}_k} \setminus E = B_{\hat{\mathcal{I}}_k \times \hat{\mathcal{I}}_k} - E$.

On the right hand side it is the same block decomposition applied to the matrix which we want the regularized empirical count matrix converges to. Recall that we defined $\tilde{\mathbb{B}}_k = \tilde{\rho}_k \tilde{\rho}_k^\top + \tilde{\Delta}_k \tilde{\Delta}_k^\top$ in (2.25), where we set entries corresponding to the words in the spillover set $\hat{\mathcal{J}}_k$ to 0.

We bound the spectral distance of the 4 blocks (A_1, A_2, A_3, A_4) separately. The bound for the entire matrix \hat{B}_k is then an immediate result of triangle inequality:

$$\begin{aligned} \|\tilde{B}_k - \tilde{\mathbb{B}}_k\| &= \|[B_{N2}]_{\hat{\mathcal{L}}_k \times \hat{\mathcal{L}}_k} - E - \tilde{\mathbb{B}}_k\| \\ &= \|A_1 \setminus E + A_2 \setminus E + A_3 \setminus E + A_4 \setminus E - N\tilde{\mathbb{B}}_k\| \\ &\leq \|A_1 \setminus E - \mathbb{B}_{\hat{\mathcal{L}}_k \times \hat{\mathcal{L}}_k}\| + \|A_2 \setminus E\| + \|A_3 \setminus E\| + \|A_4 \setminus E\|. \end{aligned}$$

We bound the 4 parts separately below in **(a)-(c)**.

(a) To bound $\|A_1 \setminus E - \mathbb{B}_{\hat{\mathcal{L}}_k \times \hat{\mathcal{L}}_k}\|$, we first make a few observations:

1. By definition of $\hat{\mathcal{J}}_k$ and $\hat{\mathcal{L}}_k$, every entry of the random matrix A_1 is distributed as an independent Poisson variable $\frac{1}{N} \text{Poi}(\lambda_k)$, where $\lambda_k \leq N \left(\frac{e^{k+\tau}}{M}\right)^2 \leq d_0 \frac{\log(M)}{M} = o(1)$.
2. The expected row sum of A_1 is bounded by of d_k^{\max} .
3. With the regularization of removing the heavy rows and columns in E , every column sum and the row sum of A_1 is bounded by $2d_k^{\max}$.

Therefore, by applying the Lemma 2.25 (an immediate extension of the main theorem in [76]), we can argue that with probability at least $1 - M_k^{-r}$ for some constant $r = \Theta(1)$,

$$\|A_1 \setminus E - \mathbb{B}_{\hat{\mathcal{L}}_k \times \hat{\mathcal{L}}_k}\|_2 = O(\sqrt{d_k^{\max}/N}).$$

(b) To bound $\|A_2 \setminus E\|$ and $\|A_3 \setminus E\|$, the key observations are:

1. Every row sum of $A_2 \setminus E$ and every column sum of $A_3 \setminus E$ is bounded by $2d_k^{\max}$.

2. For every non-zero row of A_2 , its distribution is entry-wise dominated by a multinomial distribution $\frac{1}{N} \text{Mul} \left(\frac{\rho_{\hat{L}_k}}{\sum_{i \in \hat{L}_k} \rho_i}; (2N d_k^{\max}) \right)$, while the entries in E are set to 0, and note that in A_2 the columns are restricted to the good words $\hat{\mathcal{L}}_k$. Moreover, by the Poisson assumption on N (recall that $d_k^{\max} = W_k \bar{\rho}_k$), we have that the distributions of the entries in the row are independently dominated by $\frac{1}{N} \text{Poi} \left(\frac{2N d_k^{\max}}{M_k} \right)$.

Lemma 2.18 (row/column-wise ℓ_1 norm to ℓ_2 norm bound (Lemma 2.5 in [76])). *Consider a matrix B in which each row has \mathcal{L}_1 norm at most a and each column has \mathcal{L}_1 norm at most b , then $\|B\|_2 \leq \sqrt{ab}$.*

Claim 2.3 (Sparse decomposition of $(A_2 \setminus E)$). *With high probability, the index subset $\hat{\mathcal{J}}_k \times \hat{\mathcal{L}}_k$ of $(A_2 \setminus E)$ can be decomposed into two disjoint subsets \mathcal{R} and \mathcal{C} such that: each row of \mathcal{R} and each column of \mathcal{C} has row/column sum at most $(\frac{r}{N} \log(N d_k^{\max}))$, for some constant r .*

Recall that from regularization we know that each column of \mathcal{R} and each row of \mathcal{C} in $A_2 \setminus E$ has column/row sum at most $2d_k^{\max}$. Therefore we can apply Lemma 2.18 and conclude that with high probability

$$\|A_2 \setminus E\|_2 \leq 2 \sqrt{\frac{r d_k^{\max} \log(N d_k^{\max})}{N}}.$$

Proof. (to Claim 2.3)

We sketch the proof of Claim 2.3, which mostly follows the sparse decomposition argument in Theorem 6.3 in [76]. We adapt their argument in our setup where the entries are distributed according to independent Poisson distributions. We first show (in (1)) that, with high probability, any square submatrix in $(A_2 \setminus E)$ actually contains a sparse column with almost only a constant column sum; then, with this property we can (in (2)) iteratively take out sparse columns and rows from $(A_2 \setminus E)$ to construct the \mathcal{R} and \mathcal{C} .

(1) With high probability, in any square submatrix of size $m \times m$ in $(A_2 \setminus E)$, there exists a sparse column whose sum is at most $(\frac{r}{N} \log(N d_k^{\max}))$.

To show this, consider an arbitrary column in an arbitrary submatrix of size $m \times m$ in $(A_2 \setminus E)$. Recall our observation (b).2, that the column sum is dominated by $\frac{1}{N} \text{Poi}(\lambda)$ with rate

$$\lambda = 2Nd_k^{\max} \frac{m}{M_k}.$$

Therefore, we can bound the column sum by applying the Chernoff bound for Poisson distribution (Lemma 2.23):

$$\begin{aligned} \Pr \left(\text{a column sum} > \left(\frac{r}{N} \log Nd_k^{\max} \right) \right) &\leq \Pr \left(\text{Poi}(\lambda) > (r \log Nd_k^{\max}) \right) \\ &\leq e^{-\lambda} \left(\frac{r \log Nd_k^{\max}}{e\lambda} \right)^{-r \log Nd_k^{\max}} \\ &\leq \left(\frac{rM_k}{2Nd_k^{\max}m} \right)^{-r \log Nd_k^{\max}} \\ &\leq \left(\frac{rM_k}{2m} \right)^{-r}, \end{aligned}$$

where in the last inequality we used the fact that for Nd_k^{\max} and r to be large constant, the following simple inequality holds:

$$\log(Nd_k^{\max}) \log \left(\frac{rM_k}{2Nd_k^{\max}m} \right) \geq \log \left(\frac{rM_k}{m} \right).$$

Then consider all the m columns in the submatrix of size $m \times m$, which column sums are independently dominated by Poisson distributions, we have

$$\Pr \left(\text{every column sum} > \left(\frac{r}{N} \log Nd_k^{\max} \right) \right) \leq \left(\frac{rM_k}{2m} \right)^{-rm}.$$

Next, take a union bound over all the $m \times m$ submatrices of $(A_2 \setminus E)$ for m ranging between 1 and \overline{M}_k , and recall that block $(A_2 \setminus E)$ is of size $\overline{M}_k \times (M_k - \overline{M}_k)$. We can

bound for all the submatrices:

$$\begin{aligned}
& \Pr \left(\text{for every submatrix in } (A_2 \setminus E), \text{ there exist a column whose sum } \leq \left(\frac{r}{N} \log N d_k^{\max} \right) \right) \\
& \geq 1 - \sum_{m=1}^{\overline{M}_k} \binom{M_k}{m} \binom{\overline{M}_k}{m} \left(\frac{r M_k}{2m} \right)^{-rm} \\
& \geq 1 - \sum_{m=1}^{\overline{M}_k} \left(\frac{M_k}{m} \right)^{2m} \left(\frac{r M_k}{2m} \right)^{-rm} \\
& \geq 1 - M_k^{-(r-2)}. \tag{2.53}
\end{aligned}$$

Note that this is indeed a high probability event, since for $W_k \geq \epsilon_0 e^{-k}$, we have shown that $M_k \geq M e^{-2k+\tau}$.

(2) Perform iterative row and column deletion to construct \mathcal{R} and \mathcal{C} .

Given $(A_2 \setminus E)$ of size $\overline{M}_k \times M_k$, we apply the argument above in (1) iteratively. First select a sparse column and remove it to \mathcal{C} , and apply it to remove columns until the remaining number of columns and rows are equal, then apply it alternatively to the rows (move to \mathcal{R}) and columns (move to \mathcal{C}) until empty. By construction, there are at most M_k such sparse columns in \mathcal{C} , each column of \mathcal{C} has sum bounded by $(\frac{r}{N} \log N d_k^{\max})$, and each row of \mathcal{C} bounded by $2d_k^{\max}$ because it is in the regularized $(A_2 \setminus E)$; similarly \mathcal{R} has at most \overline{M}_k rows and each row of \mathcal{R} has sum at most $(\frac{r}{N} \log N d_k^{\max})$ and each column has sum at most $2d_k^{\max}$. □

The proof for the other narrow strip $(A_3 \setminus E)$ is in parallel with the above analysis for $(A_2 \setminus E)$.

(c) To bound $\|A_4 \setminus E\|$, the two key observation are:

1. The total marginal probability mass of spillover heavy words $\overline{W}_k = \sum_{i \in \hat{\mathcal{J}}_k} \rho_i \leq 2e^{-e^{k+\tau} d_0/2}$. (shown in Lemma 2.12).
2. Similar to the observation in (b).2 above, the distributions of the entries in each row of $(A_4 \setminus E)$ are independently dominated by $\frac{1}{N} \text{Poi} \left(2N d_k^{\max} \frac{\overline{W}_k}{W_k} \frac{1}{\overline{M}_k} \right)$.

In parallel with Claim 2.3, we make a claim about the spectral norm of the block $(A_4 \setminus E)$:

Claim 2.4 (Sparse decomposition of $(A_4 \setminus E)$). *With high probability, the index subset $\widehat{\mathcal{I}}_k \times \widehat{\mathcal{J}}_k$ of A_2 can be decomposed into two disjoint subsets \mathcal{R} and \mathcal{C} such that: each row of \mathcal{R} and each column of \mathcal{C} has sum at most $\frac{r}{N}$; each column of \mathcal{R} and each row of \mathcal{C} has sum at most d_k^{\max} .*

Proof. (to Claim 2.4)

To show this, we construct sparse decomposition similar to that of $(A_2 \setminus E)$.

The only difference is that, when considering all the $m \times m$ submatrices, we only need to consider all the submatrices contained in the small square $(A_4 \setminus E)$ of size $\widehat{\mathcal{I}}_k \times \widehat{\mathcal{J}}_k$, instead of all submatrices in the wide strip $(A_2 \setminus E)$ of size $\widehat{\mathcal{L}}_k \times \widehat{\mathcal{J}}_k$. In this case, taking the union bound leads to factors of \overline{M}_k , compared to that of M_k in (2.53).

Here we only highlight the difference in the inequalities. Consider an arbitrary column in an arbitrary submatrix of size $m \times m$ in $(A_4 \setminus E)$. Recall that this column sum is dominated by $\frac{1}{N}\text{Poi}(\lambda)$ with rate

$$\lambda = 2Nd_k^{\max} \frac{\overline{W}_k}{W_k} \frac{m}{M_k}.$$

Thus we can bound the probability of having a dense column by:

$$\Pr(\text{a column sum} > \frac{r}{N}) \leq \Pr(\text{Poi}(\lambda) > r) \leq e^{-\lambda} \left(\frac{r}{e\lambda}\right)^{-r} \leq \left(\frac{r}{e\lambda}\right)^{-r}.$$

Take a union over all the square matrices of size $m \times m$ in block $(A_4 \setminus E)$, we can

bound:

$$\begin{aligned}
& \Pr(\text{for every submatrix in } (A_4 \setminus E), \text{ there exist a column whose sum } \leq \frac{r}{N}) \\
& \geq 1 - \sum_{m=1}^{\bar{M}_k} \binom{\bar{M}_k}{m}^2 \left(\frac{r}{e\lambda}\right)^{-rm} \\
& \geq 1 - \sum_{m=1}^{\bar{M}_k} \binom{\bar{M}_k}{m}^{2m} \left(\frac{\bar{M}_k}{m} \frac{rW_k}{e2d_k^{\max}\bar{W}_k}\right)^{-rm} \\
& \geq 1 - M_k^{-(r-2)},
\end{aligned}$$

where in the last inequality we used the fact that $d_k^{\max} = NW_k \frac{e^{k+\tau}}{M}$, and plug in the high probability upper bound of $\bar{W}_k \leq 2e^{-e^{k+\tau}d_0}$ as in (2.21), we have:

$$\frac{rW_k}{e2d_k^{\max}\bar{W}_k} = \frac{rW_k M}{2eNW_k e^{k+\tau} e^{-e^{k+\tau}d_0}} = \frac{re^{(e^{k+\tau}d_0)}}{2e(e^{k+\tau}d_0)} \gg 1.$$

Again note that given that the bin has significant total marginal probability, thus $M_k \geq Me^{-2k}$, the above probability bound is indeed a high probability statement. \square

\square

Proof. (to Lemma 2.6 (Given spectral concentration of block \tilde{B}_k , estimate the separation $\tilde{\Delta}_k$))

Recall the result of Lemma 2.5 about the concentration of the diagonal block with regularization. For empirical bin with large enough marginal W_k , we have with high probability,

$$\left\| \tilde{B}_k - \tilde{\mathbb{B}}_k \right\|_2 \leq C \sqrt{\frac{d_k^{\max} \log^2 d_k^{\max}}{N}}.$$

Also recall that $\hat{\rho}_{\tilde{\mathcal{I}}_k}$ is defined to be the exact marginal vector restricted to the empirical bin $\tilde{\mathcal{I}}_k$.

We can also bound

$$\begin{aligned}
\left\| (\tilde{B}_k - \hat{\rho}_{\hat{\mathcal{I}}_k} \hat{\rho}_{\hat{\mathcal{I}}_k}^\top) - \tilde{\Delta}_k \tilde{\Delta}_k^\top \right\|_2 &\leq \left\| (\tilde{B}_k - \hat{\rho}_{\hat{\mathcal{I}}_k} \hat{\rho}_{\hat{\mathcal{I}}_k}^\top) - (\tilde{\mathbb{B}}_k - \tilde{\rho}_k \tilde{\rho}_k^\top) \right\|_2 \\
&\leq \left\| \tilde{B}_k - \tilde{\mathbb{B}}_k \right\|_2 + \left\| \hat{\rho}_{\hat{\mathcal{I}}_k} \hat{\rho}_{\hat{\mathcal{I}}_k}^\top - \tilde{\rho}_k \tilde{\rho}_k^\top \right\|_2 \\
&\leq C \sqrt{\frac{d_k^{\max} \log^2 d_k^{\max}}{N}}.
\end{aligned}$$

Note that in the last inequality above we ignored the term $\|\hat{\rho}_{\hat{\mathcal{I}}_k} - \tilde{\rho}_k\|_2$ as it is small for all bins (with large probability):

$$\begin{aligned}
\left\| \hat{\rho}_{\hat{\mathcal{I}}_k} \hat{\rho}_{\hat{\mathcal{I}}_k}^\top - \tilde{\rho}_k \tilde{\rho}_k^\top \right\|_2 &\leq 4 \left\| \hat{\rho}_{\hat{\mathcal{I}}_k} - \tilde{\rho}_k \right\|_2 \left\| \hat{\rho}_{\hat{\mathcal{I}}_k} \right\|_2 \leq \sqrt{\underbrace{\bar{\rho}_k (M_k/N)}_{\text{over } \hat{\mathcal{L}}_k} + \underbrace{\bar{\rho}_k^2 (\bar{W}_k / \bar{\rho}_k)}_{\text{over } \hat{\mathcal{J}}_k}} \sqrt{M_k \bar{\rho}_k^2} \\
&\leq \sqrt{\frac{M_k^2 \bar{\rho}_k^3}{N}} = o\left(\sqrt{\frac{d_k^{\max}}{N}}\right),
\end{aligned}$$

where in the second inequality we write $\left\| \hat{\rho}_{\hat{\mathcal{I}}_k} - \tilde{\rho}_k \right\|_2^2$ into two parts over the set of good words $\hat{\mathcal{L}}_k$ and the set of bad words $\hat{\mathcal{J}}_k$. To bound the sum over $\hat{\mathcal{L}}_k$ we used the Markov inequality as in the proof of Lemma 2.1; and to bound the sum over $\hat{\mathcal{J}}_k$ as well as the term $\left\| \hat{\rho}_{\hat{\mathcal{I}}_k} \right\|_2^2$ we used the fact that if a word i appears in $\hat{\mathcal{I}}_k$, we must have $\hat{\rho}_i \leq \bar{\rho}_k$. The last inequality is due to $M_k^2 \bar{\rho}_k^3 \leq W_k^2 \bar{\rho}_k \leq W_k \bar{\rho}_k = d_k^{\max}$.

Let $v_k v_k^\top$ be the rank-1 truncated SVD of the regularized block $(\tilde{B}_k - \hat{\rho}_k \hat{\rho}_k^\top)$. Apply Wedin's theorem to rank-1 matrix (Lemma 2.22), we can bound the distance between vector v_k and $\tilde{\Delta}_k$ by:

$$\begin{aligned}
&\min \left\{ \|\tilde{\Delta}_k - v_k\|, \|\tilde{\Delta}_k + v_k\| \right\} \\
&= O \left(\min \left\{ \sqrt{\frac{d_k^{\max}}{N} \log(N d_k^{\max})} \frac{1}{\|\tilde{\Delta}_k\|_2}, \left(\sqrt{\frac{d_k^{\max}}{N} \log(N d_k^{\max})} \right)^{1/2} \right\} \right).
\end{aligned}$$

□

Proof. (to Lemma 2.7 (Accuracy of $\hat{\Delta}$ in Phase I)) Consider for each empirical bin. If $\widehat{W}_k < \epsilon_0 e^{-k}$ set $\hat{\Delta}_{\hat{\mathcal{I}}_k} = 0$. We can bound the total ℓ_1 norm error incurred in those bins

by ϵ_0 . Also, for the lightest bin, we can bound the total ℓ_1 norm error from setting $\widehat{\Delta}_{\widehat{\mathcal{I}}_0} = 0$ by ϵ_0 small constant. If $\widehat{W}_k > \epsilon_0 e^{-k}$, we can apply the concentration bounds in Lemma 2.2, 2.6, and note that $\|\widehat{\Delta}_{\widehat{\mathcal{I}}_k} - \Delta_{\widehat{\mathcal{I}}_k}\|_1 \leq \sqrt{M_k} \|\widehat{\Delta}_{\widehat{\mathcal{I}}_k} - \Delta_{\widehat{\mathcal{I}}_k}\|_2$.

Note that we need to take a union bound of probability that spectral concentration results holds (Lemma 2.5) for all the bins with large enough marginal. This is true because we have at most $\log \log M$ bins, and each bin's spectral concentration holds with high probability $(1 - 1/\text{poly}(M))$, thus even after taking the union bound the failure probability is still inverse poly in M .

Actually throughout our discussion the small constant failure probability is only incurred when bounding the estimation error of $\widehat{\rho}$, for the same reason of estimating a simple and unstructured distribution.

Overall, we can bound the estimation error in ℓ_1 norm by:

$$\begin{aligned}
\|\widehat{\Delta} - \Delta\|_1 &\leq \underbrace{\epsilon_0}_{\text{lightest bin}} + \underbrace{1/d_0^{1/4}}_{\text{heaviest bin}} + \underbrace{\epsilon_0}_{\text{moderate bins with small marginal}} + \underbrace{\sum_k \sqrt{M_k} \left(\sqrt{\frac{d_k^{\max} \log N d_k^{\max}}{N}} \right)^{1/2}}_{\text{moderate bins with large marginal}} \\
&\leq 2\epsilon_0 + 1/d_0^{1/4} + \sum_k \left(\frac{M_k^2 W_k \bar{\rho}_k \log(N W_k \bar{\rho}_k)}{N} \right)^{1/4} \\
&\leq 2\epsilon_0 + (\log(d_0)/d_0)^{1/4} (1 + e^\tau \sum_k \left(\frac{W_k^2 M_k}{M} \right)^{1/4}) \\
&\leq 2\epsilon + (\log(d_0)/d_0)^{1/4} (1 + e^\tau) \\
&= O(\epsilon_0)
\end{aligned}$$

where in the second last inequality used Cauchy-Schwartz and the fact $W_k \leq 1$, so that $\sum_k (W_k^2 M_k)^{1/4} \leq \sum_k (\sqrt{W_k} \sqrt{M_k})^{1/2} \leq (\sum_k W_k \sum_k M_k)^{1/4} \leq M^{1/4}$, and in the last inequality above we use the assumption that $d_0 =: N/M$ satisfies that $d_0/\log(d_0) \geq 1/\epsilon_0^4$. \square

2.7.2 Proofs for Rank 2 Algorithm Phase II

Proof. (to Lemma 2.8 (Sufficient condition for constructing an anchor partition))

(1) First, we show that if for some constant $c = \Omega(1)$, a set of words \mathcal{A} satisfy

$$\left| \sum_{i \in \mathcal{A}} \Delta_i \right| \geq c \|\Delta\|_1, \quad (2.54)$$

then $(\mathcal{A}, [M] \setminus \mathcal{A})$ is a pair of anchor set defined in 2.3.

By the assumption of constant separation ($C_\Delta = \Omega(1)$), $\sum_{i \in \mathcal{A}} \Delta_i = \Omega(1)$. We can bound the condition number of the anchor partition matrix by:

$$\text{cond} \left(\begin{bmatrix} \rho_{\mathcal{A}} & \Delta_{\mathcal{A}} \\ 1 - \rho_{\mathcal{A}} & -\Delta_{\mathcal{A}} \end{bmatrix} \right) = \frac{\sqrt{T^2 - 4D} + T}{\sqrt{T^2 - 4D} - T} \leq \frac{\sqrt{1 + 4cC_\Delta} + 1}{\sqrt{1 + 4cC_\Delta} - 1} = \Omega(1),$$

where $T = \rho_{\mathcal{A}} - \Delta_{\mathcal{A}} \leq 1$ and $D = -\rho_{\mathcal{A}}\Delta_{\mathcal{A}} - (1 - \rho_{\mathcal{A}})\Delta_{\mathcal{A}} = -\Delta_{\mathcal{A}}$.

(2) Next we show that $\widehat{\mathcal{A}}$ defined in the lemma statement satisfies (2.54).

Denote $\mathcal{A}^* = \{i \in \mathcal{I} : \Delta_i > 0\}$. Note that $\|\Delta_{\mathcal{I}}\|_1 = \sum_{i \in \mathcal{A}^*} \Delta_i - \sum_{i \in \mathcal{I} \setminus \mathcal{A}^*} \Delta_i$. Without loss of generality we assume that $\sum_{i \in \mathcal{A}^*} \Delta_i \geq \frac{1}{2} \|\Delta_{\mathcal{I}}\|_1 \geq \frac{1}{2} C \|\Delta\|_1$, where the last inequality is by the condition $\|\Delta_{\mathcal{I}}\|_1 \geq C \|\Delta\|_1$.

Given $\widehat{\Delta}_{\mathcal{I}}$ that satisfies (2.29). We look at $\widehat{\mathcal{A}} = \{i \in \mathcal{I} : \widehat{\Delta}_i > 0\}$.

$$\begin{aligned} \sum_{i \in \widehat{\mathcal{A}}} \Delta_i &= \sum_{i \in \widehat{\mathcal{A}} \cap \mathcal{A}^*} \Delta_i - \sum_{i \in \widehat{\mathcal{A}} \cap (\mathcal{I} \setminus \mathcal{A}^*)} \Delta_i \\ &= \sum_{i \in \mathcal{A}^*} \Delta_i - \sum_{i \in (\widehat{\mathcal{A}} \cap (\mathcal{I} \setminus \mathcal{A}^*)) \cup (\mathcal{A}^* \cap (\mathcal{I} \setminus \widehat{\mathcal{A}}))} |\Delta_i| \\ &\geq \sum_{i \in \mathcal{A}^*} \Delta_i - \|\widehat{\Delta}_{\mathcal{I}} - \Delta_{\mathcal{I}}\|_1 \\ &\geq \left(\frac{1}{2}C - C'\right) \|\Delta\|_1 \\ &\geq \frac{1}{6} C C_\Delta, \end{aligned}$$

where in the second last inequality we used the fact that, if the sign of $\widehat{\Delta}_i$ and Δ_i are different, it must be that $|\widehat{\Delta}_i - \Delta_i| > |\Delta_i|$. \square

Proof. (to Lemma 2.10 (Estimate the separation restricted to the k -th good bin))

Since it is a good bin, we have the ℓ_2 bound given by Lemma 2.6 as below (as-

suming the possible sign flip has been fixed as in Lemma 2.2):

$$\|\tilde{\Delta}_k - v_k\|_2 \leq \frac{\sqrt{d_k^{\max}} \log^2 d_k^{\max}}{N} \frac{1}{\|\tilde{\Delta}_k\|_2}.$$

Then we can convert the bound to ℓ_1 distance by:

$$\begin{aligned} \frac{\|v_k - \tilde{\Delta}_k\|_1}{\|\tilde{\Delta}_k\|_1} &\leq \frac{\sqrt{M_k} \|v_k - \tilde{\Delta}_k\|_2}{\|\tilde{\Delta}_k\|_1} \\ &\leq \frac{\sqrt{M_k} \|v_k - \tilde{\Delta}_k\|_2 \|\tilde{\Delta}_k\|_2}{\|\tilde{\Delta}_k\|_2 \|\tilde{\Delta}_k\|_1} \\ &\leq \frac{M_k \|v_k v_k^\top - \tilde{\Delta}_k \tilde{\Delta}_k^\top\|}{\|\tilde{\Delta}_k\|_1^2} \\ &\leq C \frac{M_k}{W_k^2} \sqrt{d_k^{\max}} \frac{\log^2(d_0)}{N} \\ &\leq C \frac{M_k}{W_k^2} e^\tau \sqrt{N W_k \frac{W_k}{M_k} \log^2(d_0)} \\ &\leq C e^\tau \sqrt{\frac{M \log^2(d_0)}{N W_k e^k}} \\ &= O\left(\sqrt{\frac{\log(d_0)}{d_0}}\right), \end{aligned}$$

where in the second last inequality, we used the fact that $M_k \frac{e^k}{M} \leq W_k$ again, and in the last inequality we used the assumption $W_k \geq \epsilon_0/e^k$.

□

Proof. (to Lemma 2.9 (With $\Omega(1)$ separation, most words fall in “good bins” with high probability))

This proof is mostly playing around with the probability mass and converting something obviously true in expectation to high probability argument.

(1) Note that by their definition we know that $W_k \geq \frac{1}{2}S_k$, and we have

$$\begin{aligned}
& \sum_k W_k \left(\mathbf{1}_{\left[\frac{S_k}{2W_k} \geq C_2\right]} + \frac{S_k}{2W_k} \mathbf{1}_{\left[\frac{S_k}{2W_k} < C_2\right]} \right) \\
& \geq \sum_k W_k \frac{S_k}{2W_k} \left(\mathbf{1}_{\left[\frac{S_k}{2W_k} \geq C_2\right]} + \mathbf{1}_{\left[\frac{S_k}{2W_k} < C_2\right]} \right) \\
& = \sum_k W_k \frac{S_k}{2W_k} \\
& = \frac{1}{2} \sum_k S_k,
\end{aligned}$$

Moreover, note that by definition of W_k , we have $\sum_k W_k = 1$, therefore

$$\sum_k W_k \frac{S_k}{2W_k} \mathbf{1}_{\left[\frac{S_k}{2W_k} < C_2\right]} \leq \sum_k C_2 W_k = C_2.$$

From the above two inequalities we can bound

$$\sum_k W_k \mathbf{1}_{\left[\frac{S_k}{2W_k} \geq C_2\right]} \geq \frac{1}{2} \sum_k S_k - C_2.$$

Also note that

$$\sum_k W_k \mathbf{1}_{\left[W_k < \frac{C_1}{2^k}\right]} \leq C_1.$$

Therefore according to the definition of “good bins” we have that:

$$\sum_{k \in \mathcal{G}} W_k = \sum_k W_k \mathbf{1}_{\left[\frac{S_k}{2W_k} \geq C_2 \text{ and } W_k \geq \frac{C_1}{2^k}\right]} \geq \frac{1}{2} \sum_k S_k - C_2 - C_1. \quad (2.55)$$

(2) We want to lower bound the quantity $\sum_k S_k$ to be a constant fraction of $\|\Delta\|_1$.

Note that by definition of S_k we can equivalently write the sum as:

$$\sum_k S_k = \sum_k \sum_{i \in \widehat{\mathcal{I}}_k \cap (\cup_{\{k': k' \leq k+\tau\}} \mathcal{I}_{k'})} |\Delta_i| = \sum_i |\Delta_i| \sum_k \mathbf{1}_{[i \in \mathcal{I}_k \cap \cup_{\{k': k' \leq k+\tau\}} \widehat{\mathcal{I}}_{k'}]}.$$

Consider for each word i . Assume that word $i \in \mathcal{I}_k$. Given $N = d_0 M$ for

some large constant d_0 , denote $d_k = e^k d_0$, we can bound the probability $\Pr(i \in \mathcal{I}_k \cap \cup_{\{k':k' \leq k+\tau\}} \widehat{\mathcal{I}}_{k'})$ as follows:

$$\begin{aligned} \Pr(i \in \mathcal{I}_k \cap \cup_{\{k':k' \leq k+\tau\}} \widehat{\mathcal{I}}_{k'}) &\geq 1 - \Pr(\text{Poi}(N\rho_i) > e^\tau \rho_i) - \Pr(\text{Poi}(N\rho_i) < e^{-\tau} N\rho_i/2) \\ &\geq 1 - \frac{e^{-(\tau-1)e^{(\tau+k)d_0}}}{\sqrt{2\pi e^{k+\tau} d_0}} - e^{-d_k} \\ &\geq 1 - 2e^{-d_k}. \end{aligned}$$

Therefore at least in expectation we can lower bound the sum by

$$\mathbb{E}[\sum_k S_k] = \sum_i |\Delta_i| \sum_k \Pr(i \in \mathcal{I}_k \cap \cup_{\{k':k' \leq k+\tau\}} \widehat{\mathcal{I}}_{k'}) \geq (1 - 2e^{-d_0}) \|\Delta\|_1.$$

(3) Restrict to the exact good bins, for which we know that the exact $\|\rho_{\widehat{\mathcal{I}}_k}\|_1 \geq e^{-k}$ and $\|\Delta_{\widehat{\mathcal{I}}_k}\|_1 / \|\rho_{\mathcal{I}_k}\|_1 \geq C$.

Here we know that if $\widehat{\mathcal{I}}_k$ is an good bin, the number of words in this exact good bin is lower bounded by $M_k \geq e^{-k} / \rho_k \geq M/e^{-k}$, and since $k \leq \log \log M$ we have that $M_k \geq \frac{M}{\log(M)}$. This is important for use to apply Bernstein concentration of the words in the bin.

Since $\|\Delta_{\widehat{\mathcal{I}}_k}\|_1 / \|\rho_{\widehat{\mathcal{I}}_k}\|_1 \geq C$, and that $|\Delta_i| \leq \rho_i$, we have that out of the M_k words there are at least a constant fraction of words with $|\Delta_i| \geq \frac{1}{2}C\rho_k$. Recall that we denote $\rho_k = e^k/M$. This is easy to see as $x\rho_k + (M_k - x)\frac{1}{2}C\rho_k \geq \|\Delta_{\widehat{\mathcal{I}}_k}\|_1 \geq CM_k\rho_k$ thus $x \geq C/2 - CM_k$.

Then we bound the probability that out of these cM_k words with good separation, a constant fraction of them do not escape from the closest τ empirical bins. Denote $\lambda_k = 2e^{-d_k}$, which is the upper bound of the escaping probability for each of the word, and is very small. By a simple application of Bernstein bounds of the Bernoulli sum,

for a small constant c_0 , we have

$$\begin{aligned}
\Pr\left(\sum_{i=1,\dots,cM_k} \text{Ber}_i(\lambda_k) \geq c_0 M_k\right) &\leq \Pr\left(\sum_{i=1,\dots,cM_k} \text{Ber}_i(\lambda_k) - \lambda_k M_k \geq (c_0 - \lambda_k) M_k\right) \\
&\leq e^{-\frac{\frac{1}{2}(c_0 - \lambda_k)^2 M_k^2}{M_k \lambda_k + \frac{1}{3}(c_0 - \lambda_k) M_k}} \\
&\approx e^{-c_0 M_k}.
\end{aligned}$$

Then union bound over all the exact good bins. That gives a $\log \log M$ multiply of the probability.

We now know that restricting to the non-escaping good words in the exact good bins, they already contribute a constant fraction (due to constant non-escaping, constant ratio $\|\Delta_{\hat{\mathcal{I}}_k}\|_1 / \|\rho_{\hat{\mathcal{I}}_k}\|_1$, and constant $\sum_{k \in \text{exact good bin}} W_k$) of the total separation $\|\Delta\|_1$. Therefore we can conclude that for some universal constant C we have

$$\sum_k S_k \geq C \|\Delta\|_1.$$

(4) Finally plug the above bound of $\sum_k S_k$ into (2.55), and note the assumption on the constants C_1 and C_2 , we can conclude that the total marginal probability mass contained in “good bins” is large:

$$\sum_{k \in \mathcal{G}} W_k \geq \left(C - \frac{1}{24} - \frac{1}{24}\right) \|\Delta\|_1 = \frac{1}{12} \|\Delta\|_1.$$

□

Proof. (to Lemma 2.11 (Estimate ρ and Δ with accuracy in ℓ_2 distance))

Consider word i we have that

$$\begin{bmatrix} \rho_{\mathcal{A}}, & \Delta_{\mathcal{A}} \\ 1 - \rho_{\mathcal{A}}, & -\Delta_{\mathcal{A}} \end{bmatrix} \begin{bmatrix} \rho_i \\ \Delta_i \end{bmatrix} = \begin{bmatrix} \sum_{j \in \mathcal{A}} \mathbb{B}_{j,i} \\ \sum_{j \in \mathcal{A}^c} \mathbb{B}_{j,i} \end{bmatrix}$$

Set

$$\begin{bmatrix} \widehat{\rho}_i \\ \widehat{\Delta}_i \end{bmatrix} = \begin{bmatrix} \rho_{\mathcal{A}}, & \Delta_{\mathcal{A}} \\ 1 - \rho_{\mathcal{A}}, & -\Delta_{\mathcal{A}} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j \in \mathcal{A}} B_{j,i} \\ \sum_{j \in \mathcal{A}^c} B_{j,i} \end{bmatrix}$$

Since $\sum_{j \in \mathcal{A}} B_{j,i} \sim \frac{1}{N} \text{Poi}(N(\rho_{\mathcal{A}}\rho_i + \Delta_{\mathcal{A}}\Delta_i))$, apply Markov inequality, we have that

$$\Pr\left(\sum_i \left(\sum_{j \in \mathcal{A}} B_{j,i} - \sum_{j \in \mathcal{A}} \mathbb{B}_{j,i}\right)^2 > \epsilon^2\right) \leq \frac{\frac{1}{N} \sum_i (\rho_{\mathcal{A}}\rho_i + \Delta_{\mathcal{A}}\Delta_i)}{\epsilon^2} = \frac{\rho_{\mathcal{A}}}{N\epsilon^2}$$

Note that $\rho_{\mathcal{A}} = \Omega(1)$ and that $\text{cond}(D_{\mathcal{A}}) = \Omega(1)$, we can propagate the concentration to the estimation error of ρ and Δ as, for some constant $C = \Omega(1)$,

$$\Pr(\|\widehat{\rho} - \rho\| > \epsilon) \leq \frac{C}{N\epsilon^2}, \quad \Pr(\|\widehat{\Delta} - \Delta\| > \epsilon) \leq \frac{C}{N\epsilon^2}.$$

□

2.7.3 Proofs for Rank R Algorithm

Proof. (to Lemma 2.12)

$$\mathbb{E}[\overline{W}_k] \leq 2e^{-e^{\tau+k}d_0/2}.$$

First we argue that with high probability, all words in bins $k > \log \log M$ concentrates well. For $\rho_i > \frac{\log M}{M}$, set constant $C = 1/2$, we have

$$\Pr(\text{Poi}(N\rho_i) < \frac{1}{2}N\rho_i) \leq 2e^{-N\rho_i/2} \leq 2e^{N \log M/2M}.$$

There are at most $\frac{M}{\log M}$ such heavy words. Take a union bound over them we have

$$\begin{aligned}
& \Pr(\forall i \text{ s.t. } i \in \mathcal{I}_k, k > \log \log M : i \in \widehat{\mathcal{I}}_{k'}, k' < k - 1) \\
& \geq 1 - \frac{M}{\log M} 2e^{-N \log M/2M} \\
& \geq 1 - 2\exp(-N \log M/2M + \log M - \log \log M) \\
& \geq 1 - 2\exp(-N \log M/4M) \\
& = 1 - 2M^{-d_0/4}.
\end{aligned}$$

Second, define $\overline{\mathcal{I}}_k = \{i : i \in \mathcal{I}_{k'}, k + \tau < k' < \log \log M\}$. For word $i \in \mathcal{I}_{k'}$, let $\lambda_i = 2e^{-e^{k'} d_0/2}$, we have $\overline{W}_k = \sum_{i \in \overline{\mathcal{I}}_k} \rho_i \text{Ber}(\lambda_i)$. By Bernstein inequality:

$$\Pr(\overline{W}_k - \mathbb{E}\overline{W}_k > t) \leq \exp\left(-\frac{t^2}{\sum_{i \in \overline{\mathcal{I}}_k} \rho_i^2 \lambda_i + \max_{i \in \overline{\mathcal{I}}_k} \rho_i t}\right).$$

In order to bound the probability by $\exp(-2 \log \log M)$, so that we can take a union bound over the $\log M$ bins, we set t to be $t = 2 \log \log M (1/\sqrt{M} + \log M/M) = O(1/\text{poly}(M))$, and note that

$$\begin{aligned}
\left(\sum_{i \in \overline{\mathcal{I}}_k} \rho_i^2 \lambda_i\right)^{1/2} + \max_{i \in \overline{\mathcal{I}}_k} \rho_i & \leq \left(\max_{i \in \overline{\mathcal{I}}_k} \frac{1}{\rho_i} \rho_i^2 \lambda_i\right)^{1/2} + \frac{\log M}{M} \\
& \leq \left(\max_{k' > k + \tau} (e^{k'}/M) 2e^{-e^{k'} d_0/2}\right)^{1/2} + \log M/M \\
& \leq 1/\sqrt{M} + \log M/M.
\end{aligned}$$

Therefore, we argue that with high probability, for all empirical bins $\widehat{\mathcal{I}}_k$, we can bound the spillover probability from heavy bins by:

$$\overline{W}_k \leq e^{-e^{\tau+k} d_0/2}.$$

□

Proof. (to Lemma 2.13) Consider for a typical word in the bin \mathcal{I}_k , we can bound the

probability that it is not contained in bin $\widehat{\mathcal{I}}_k$ by:

$$\Pr(\text{Poi}(N\rho_i) < \frac{1}{2}N\rho_i \text{ or } \text{Poi}(N\rho_i) > 2N\rho_i) \leq 4e^{-e^k d_0/2}.$$

Apply Bernstein inequality to all the $W_k/\bar{\rho}_k$ words in bin \mathcal{I}_k , denote $\lambda_k = 4e^{-e^k d_0/2}$, we have

$$\Pr(W_k^s - \mathbb{E}W_k^s > t) < \exp\left(-\frac{t^2}{M_k \bar{\rho}_k^2 \lambda_k + \bar{\rho}_k t}\right)$$

Since the bin is big, we have $W_k > e^{-k}$, we can set

$$t = (M_k \bar{\rho}_k (\lambda_k / M_k)^{1/2} + \bar{\rho}_k) \log \log M \leq 4W_k e^{-e^k d_0/4} \frac{\log M}{\sqrt{M_k}} + \frac{\log M \log \log M}{M} = o(1),$$

where the last inequality is due to $M_k > M e^{-2k}$. Take a union bound over all $\log M$ bins, we can ensure that with high probability, for each bin, the escaped mass is bounded by $4W_k e^{-e^k d_0/2}$.

□

Proof. (to Lemma 2.14) In parallel with the analysis for Rank 2 (see Lemma 2.5), we know that regularization restores spectral concentration in the diagonal blocks. Denote the noise matrix in the regularized diagonal block by $E_k = \widetilde{B}_k - \widetilde{\mathbb{B}}_k$.

$$\|E_k\| = \|\widetilde{B}_k - \widetilde{\mathbb{B}}_k\| = O\left(\frac{\sqrt{N d_k^{\max} \log N d_k^{\max}}}{N}\right)$$

Denote the R -SVD of \widetilde{B}_k by $\widehat{V}_k \widehat{\Lambda}_k \widehat{V}_k^\top$.

$$\begin{aligned} \|\text{Proj}_{\widehat{V}_k} \widetilde{\mathbb{B}}_k \text{Proj}_{\widehat{V}_k} - \widetilde{\mathbb{B}}_k\| &= \|\text{Proj}_{\widehat{V}_k} (\widetilde{B}_k - E_k) \text{Proj}_{\widehat{V}_k} - \widetilde{\mathbb{B}}_k\| \\ &\stackrel{(a)}{\leq} \|\text{Proj}_{\widehat{V}_k} \widetilde{B}_k \text{Proj}_{\widehat{V}_k} - \widetilde{\mathbb{B}}_k\| + \|\text{Proj}_{\widehat{V}_k} E_k\| \\ &\stackrel{(b)}{\leq} \|\text{Proj}_{\widehat{V}_k} \widetilde{B}_k \text{Proj}_{\widehat{V}_k} - \widetilde{B}_k\| + \|\widetilde{B}_k - \widetilde{\mathbb{B}}_k\| + \|\text{Proj}_{\widehat{V}_k} E_k\| \\ &\stackrel{(c)}{\leq} \|\widetilde{\mathbb{B}}_k - \widetilde{B}_k\| + \|E_k\| + \|\text{Proj}_{\widehat{V}_k} E_k\| \\ &\leq 3\|E_k\|, \end{aligned} \tag{2.56}$$

where inequality (a) (b) are simply triangle inequality; and inequality (c) used the fact that $\text{Proj}_{\widehat{V}_k} \widetilde{B}_k \text{Proj}_{\widehat{V}_k}$ from truncated SVD is the best rank R approximation to \widetilde{B}_k that minimizes the spectral norm. Finally, apply Lemma 2.19 we have

$$\|\text{Proj}_{\widehat{V}_k} \widetilde{\mathbb{B}}_{\mathcal{I}_k}^{sqr} - \widetilde{\mathbb{B}}_{\mathcal{I}_k}^{sqr}\| \leq \sqrt{\|\text{Proj}_{\widehat{V}_k} \widetilde{B}_k \text{Proj}_{\widehat{V}_k} - \widetilde{B}_k\|}.$$

□

Lemma 2.19. *Let U be a matrix of dimension $M \times R$. Let P be a projection matrix, we have*

$$\|U - PU\|^2 \leq \|UU^\top - PU(PU)^\top\|.$$

Proof. (to Lemma 2.19) Let $P^\perp = I - P$, so $U - PU = P^\perp U$. We can write

$$\begin{aligned} UU^\top - PU(PU)^\top &= (P + P^\perp)UU^\top(P + P^\perp) - PU(PU)^\top \\ &= P^\perp UU^\top P^\perp + PUU^\top P^\perp + P^\perp UU^\top P. \end{aligned}$$

Let vector v denote the leading left singular vector of $P^\perp U$ and $P^\perp U$, by orthogonal projection it must be that $Pv = 0$. We can bound

$$\begin{aligned} \|UU^\top - PU(PU)^\top\| &\geq |v^\top (P^\perp UU^\top P^\perp + PUU^\top P^\perp + P^\perp UU^\top P)v| \\ &= |v^\top P^\perp UU^\top P^\perp v| \\ &= \|P^\perp U\|^2. \end{aligned}$$

□

Lemma 2.20 (Scaled noise matrix). *Consider a noise matrix E_S with independent entries, and each entry has sub-exponential tail with parameter $(\sigma_{i,j} = \frac{1}{\sqrt{N}}, b_{i,j} = \frac{1}{N\sqrt{\rho_i \rho_j}})$, for $b_{i,j} \leq \frac{M}{N}$.*

Consider a fixed matrix V of dimension $M \times R$ whose columns are orthonormal,

with large probability we can bound the norm of $V^\top E_S V$ and $V^\top E_S$ separately by:

$$\|V^\top E_S V\| = O\left(\sqrt{\frac{R^2}{M}}\right), \quad \text{and} \quad \|\widehat{V}^\top E_S\| = O\left(\sqrt{\frac{RM}{N}}\right).$$

Proof. To bound the norm of the projected matrix, note that we have

$$\|V^\top E_S V\|_2^2 \leq \|V^\top E_S V\|_F^2 = \text{Tr}(V^\top E_S V V^\top \widetilde{E}^\top V).$$

By Markov inequality, we have

$$\begin{aligned} \Pr(\text{Tr}(\widehat{V}^\top E_S V V^\top E_S^\top \widehat{V}) > t) &\leq \frac{1}{t} \mathbb{E} \text{Tr}(\widehat{V}^\top E_S V V^\top E_S^\top \widehat{V}) \\ &= \frac{1}{t} \text{Tr}(\widehat{V}^\top \underbrace{\mathbb{E}[E_S V V^\top E_S^\top]}_X \widehat{V}) \\ &= \frac{1}{t} \frac{R^2}{N}, \end{aligned}$$

where the last equality is because for the i, j -th entry of X (let E_i denote the i -th row of E and V_r denote the r -th column of V)

$$X_{i,j} = \mathbb{E}\left[\sum_r (E_i V_r)(E_j V_r)\right] = \delta_{i,j} \sigma_{i,j}^2 \sum_r \|V_r\|_2^2 = \delta_{i,j} \frac{R}{N}.$$

Therefore, with probability at least $1 - \delta$, we have

$$\|V^\top E_S V\| \leq \sqrt{\frac{R^2}{N\delta}}.$$

Similarly, note that

$$\|\widehat{V}^\top E_S\|_2^2 \leq \|\widehat{V}^\top E_S\|_F^2 = \text{Tr}(\widehat{V}^\top E_S E_S^\top \widehat{V}).$$

By Markov inequality, we have

$$\Pr(\text{Tr}(\widehat{V}^\top E_S E_S^\top \widehat{V}) > t) \leq \frac{1}{t} \mathbb{E} \text{Tr}(\widehat{V}^\top \widetilde{E} E_S^\top \widehat{V}) = \frac{1}{t} \text{Tr}(\widehat{V}^\top \mathbb{E}[E_S E_S^\top] \widehat{V}) = \frac{1}{t} \frac{RM}{N}.$$

□

Proof. (to Lemma 2.15)

We first show the spectral concentration of $\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{B} D_S \text{Proj}_{\hat{\mathcal{V}}}$ to $D_S \tilde{\mathbb{B}} D_S$ can be bounded as:

$$\|\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{B} D_S \text{Proj}_{\hat{\mathcal{V}}} - D_S \tilde{\mathbb{B}} D_S\| = O\left(\frac{\log(Nw_{\min}^2/M)}{Nw_{\min}^2/M}\right)^{1/4}. \quad (2.57)$$

Note that by definition the rows and columns that are set to zero do not necessarily coincide in \tilde{B} and $\tilde{\mathbb{B}}$ (defined in (2.37)), and we do not observe the sparsity pattern in $\tilde{\mathbb{B}}$.

Define $\tilde{E} = \tilde{B} - \tilde{\mathbb{B}}$. Apply triangle inequalities we have

$$\begin{aligned} & \|\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{B} D_S \text{Proj}_{\hat{\mathcal{V}}} - D_S \tilde{\mathbb{B}} D_S\| \\ & \leq \|\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}} D_S \text{Proj}_{\hat{\mathcal{V}}} - D_S \tilde{\mathbb{B}} D_S\| + \|\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{E} D_S \text{Proj}_{\hat{\mathcal{V}}}\| \end{aligned} \quad (2.58)$$

Next, we bound the two terms in (2.58) separately.

(1) First, to bound the term $\|\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}} D_S \text{Proj}_{\hat{\mathcal{V}}} - D_S \tilde{\mathbb{B}} D_S\|$, we note that

$$\|D_S \tilde{\mathbb{B}}^{sqr}\| \leq \|\text{diag}(\rho^{-1/2}) \mathbb{B} \text{diag}(\rho^{-1/2})\| = 1. \quad (2.59)$$

Apply the block concentration result in Lemma 2.14, and recall that $d_k^{max} = M_k \bar{\rho}_k^2 / w_{\min} = W_k \rho_k / w_{\min}$, we have

$$\begin{aligned}
\|\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}}^{sqr} - D_S \tilde{\mathbb{B}}^{sqr}\| &\leq \left(\sum_k \bar{\rho}_k^{-1} \|\text{Proj}_{\hat{\mathcal{V}}_k} \tilde{\mathbb{B}}_{\mathcal{I}_k}^{sqr} - \tilde{\mathbb{B}}_{\mathcal{I}_k}^{sqr}\|^2 \right)^{1/2} \\
&= O\left(\left(\sum_k \frac{\sqrt{N d_k^{max} \log N d_k^{max}}}{N \bar{\rho}_k} \right)^{1/2} \right) \\
&\leq O\left(\frac{1}{\sqrt{N}} \sum_k \sqrt{\log\left(\frac{N W_k \bar{\rho}_k}{w_{min}}\right) \frac{M_k}{w_{min}}} \right)^{1/2} \\
&= O\left(\left(\sqrt{\frac{M}{N w_{min}}} \sum_k \sqrt{\log\left(\frac{N W_k e^k}{M w_{min}}\right) W_k e^{-k}} \right)^{1/2} \right) \\
&= O\left(\left(\sqrt{\frac{M}{N w_{min}}} \sum_k \sqrt{\left(\log\left(\frac{N}{M w_{min}}\right) + \log(W_k e^k)\right) W_k e^{-k}} \right)^{1/2} \right) \\
&\leq O\left(\left(\sqrt{\frac{M}{N w_{min}}} \log \frac{N}{M w_{min}} \sum_k \sqrt{\log(W_k e^k) W_k e^{-k}} \right)^{1/2} \right) \\
&= O\left(\frac{\log(N w_{min}^2 / M) + 3 \log(1/w_{min})}{N w_{min} / M} \right)^{1/4}, \\
&= O\left(\frac{\log(N w_{min}^2 / M)}{N w_{min}^2 / M} \right)^{1/4}, \tag{2.60}
\end{aligned}$$

where the last inequality is because $\log(1/w_{min}) \leq 1/w_{min}$, and the second last inequality is because

$$\sum_k \sqrt{\log(W_k e^k) W_k e^{-k}} \leq \sum_k \sqrt{W_k k e^{-k}} \leq \sqrt{\sum_k W_k \sum_k k e^{-k}} \leq 2.$$

Therefore, with (2.59) and (2.60) we can bound the first term in (2.58) by:

$$\begin{aligned}
&\|\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}} D_S \text{Proj}_{\hat{\mathcal{V}}} - D_S \tilde{\mathbb{B}} D_S\| \\
&\leq \|(\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}}^{sqr} - D_S \tilde{\mathbb{B}}^{sqr})(\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}}^{sqr})^\top\| + \|D_S \tilde{\mathbb{B}}^{sqr}(\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}}^{sqr} - \tilde{\mathbb{B}}^{sqr})^\top\| \\
&\leq 2 \|D_S \tilde{\mathbb{B}}^{sqr}\| \|\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}}^{sqr} - D_S \tilde{\mathbb{B}}^{sqr}\| \\
&= O\left(\frac{\log(N w_{min}^2 / M)}{N w_{min}^2 / M} \right)^{1/4}. \tag{2.61}
\end{aligned}$$

(2) Second, to bound $\|\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{E} D_S \text{Proj}_{\hat{\mathcal{V}}}\|$, we carefully analyze the regularization to take care of the spillover effect. In Figure 2-4 we divide \tilde{E} into different regions according to the sparsity pattern of $\tilde{\mathbb{B}}$ (as defined in (2.37)) and the regularized empirical matrix \tilde{B} in this step. We only highlight the division in one diagonal block, but it applies to the entire matrix across different bins. We bound the spectral norm of the matrix \tilde{E} restricting to different regions separately.

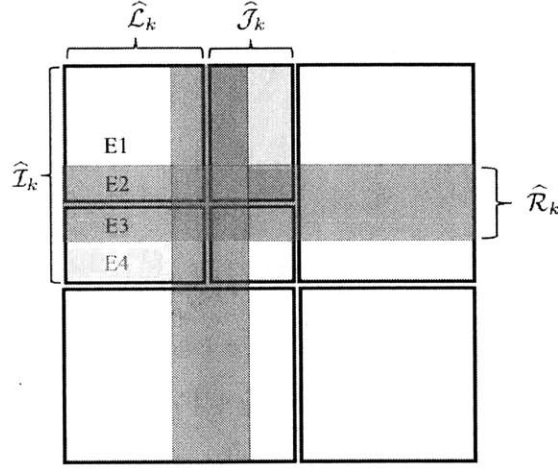


Figure 2-4: decomposition of \tilde{E} corresponding to $\hat{\mathcal{I}}_k, \hat{\mathcal{L}}_k$ and $\hat{\mathcal{R}}_k$.

In particular, region E_1 is where rows/columns are not removed by either $\tilde{\mathbb{B}}$ or \tilde{B} .

The entries are dominated by independent variables $\frac{1}{N\sqrt{\rho_i\rho_j}}(\text{Poi}(N\mathbb{B}_{i,j}) - N\mathbb{B}_{i,j})$ and have sub-exponential tail with parameter $(\sigma_{i,j} = \frac{1}{\sqrt{N}}, b_{i,j} = \frac{1}{N\sqrt{\rho_i\rho_j}} < 1)$.

Also, since independent copies of the empirical bigram matrix are used in each step of the algorithm, the noise is independent with the $(R \log M)$ -dimensional projection matrix $\text{Proj}_{\hat{\mathcal{V}}}$. Apply Lemma 2.20, with probability at least $1 - \delta$ we can bound the norm of projected noise as:

$$\|\text{Proj}_{\hat{\mathcal{V}}} D_S E_1 D_S \text{Proj}_{\hat{\mathcal{V}}}\| \leq 2\|\text{Proj}_{\hat{\mathcal{V}}} E_S \text{Proj}_{\hat{\mathcal{V}}}\| = O\left(\sqrt{\frac{(R \log M)^2}{N\delta}}\right) = o(1)$$

Region E_3 corresponds to the rows/columns that are removed by both $\tilde{\mathbb{B}}$ and \tilde{B} , thus $E_3 = 0$.

Region E_2 is set to 0 in \tilde{B} but not in $\tilde{\mathbb{B}}$, thus the entries of E_2 are equal to $[\tilde{\mathbb{B}}]_{i,j}$.

For the rows of $\widetilde{\mathbb{B}}^{sqr}$ restricted to bin $\widehat{\mathcal{I}}_k$, the row sums are bounded by $2\bar{\rho}_k$, and the column sums are bounded by $W_k^s = O(W_k e^{-e^k d_0/2})$ (Lemma 2.13). Recall the fact that if a matrix X in which each row has \mathcal{L}_1 norm at most a and each column has \mathcal{L}_1 norm at most b , then $\|X\|_2 \leq \sqrt{ab}$. Therefore, we can bound

$$\begin{aligned} \|\text{Proj}_{\widehat{\mathcal{V}}} D_S E_2 D_S \text{Proj}_{\widehat{\mathcal{V}}}\| &\leq \sum_k \left(\frac{1}{\sqrt{\bar{\rho}_k}} \sqrt{W_k^s \bar{\rho}_k} \right)^2 \\ &= \sum_k W_k^s \\ &= O(e^{-N/2M}). \end{aligned} \tag{2.62}$$

Region E_4 is set to 0 in $\widetilde{\mathbb{B}}$ but not in \widetilde{B} , corresponding to a subset of spillover words, and $E_4 = \widetilde{B}_4$. There are at most $\bar{W}_k/\bar{\rho}_k$ rows of region E_4 in each bin $\widehat{\mathcal{I}}_k$. Moreover, the row/column sum in are bounded by $2\bar{\rho}_k$. Conditional on the row sum, the entries in the row are distributed as multinomial $\text{Mul}(\rho; 2\bar{\rho}_k)$, thus the entries of $D_S E_4 D_S$ are dominated by subexponential tail with parameter ($\sigma_{i,j} = \frac{1}{\sqrt{N}}$, $b_{i,j} = \frac{1}{N\sqrt{\rho_i \rho_j}} < 1$). With probability at least δ we can bound

$$\begin{aligned} \|\text{Proj}_{\widehat{\mathcal{V}}} D_S E_4 D_S \text{Proj}_{\widehat{\mathcal{V}}}\| &\leq \|\text{Proj}_{\widehat{\mathcal{V}}} [E_S]_4 \text{Proj}_{\widehat{\mathcal{V}}} + \text{Proj}_{\widehat{\mathcal{V}}} D_S (\bar{\rho}_k \mathbf{1} \rho^\top) D_S \text{Proj}_{\widehat{\mathcal{V}}}\| \\ &\leq \|\text{Proj}_{\widehat{\mathcal{V}}} [E_S]_4 \text{Proj}_{\widehat{\mathcal{V}}}\| + \|\text{Proj}_{\widehat{\mathcal{V}}} D_S (\bar{\rho}_k \mathbf{1} \rho^\top) D_S \text{Proj}_{\widehat{\mathcal{V}}}\| \\ &\leq \sqrt{\frac{\min(\sum_k \frac{\bar{W}_k}{\bar{\rho}_k}, R \log M)(R \log M)}{N\delta}} + \left(\sum_k \left(\frac{1}{\sqrt{\bar{\rho}_k}} \sqrt{W_k \bar{\rho}_k} \right)^2 \right)^{1/2} \\ &= O\left(\sqrt{\frac{(R \log M)^2}{N\delta}} + e^{-N/2M} \right), \end{aligned}$$

where the second last inequality is by the same argument as that in (2.62).

By triangle inequality over the 4 different regions, we can bound:

$$\|\text{Proj}_{\widehat{\mathcal{V}}} D_S \widetilde{E} D_S \text{Proj}_{\widehat{\mathcal{V}}}\| = O\left(\sqrt{\frac{(R \log M)^2}{N\delta}} + e^{-N/2M} \right). \tag{2.63}$$

Therefore, with (2.61) and (2.63) we can bound (2.58) by:

$$\|\text{Proj}_{\widehat{\mathcal{V}}} D_S \widetilde{B} D_S \text{Proj}_{\widehat{\mathcal{V}}} - D_S \widetilde{\mathbb{B}} D_S\| = O\left(\frac{\log(Nw_{\min}^2/M)}{Nw_{\min}^2/M}\right)^{1/4}$$

Finally note that \widehat{B}_1 is the best rank R approximation of $\text{Proj}_{\widehat{\mathcal{V}}} D_S \widetilde{B} D_S \text{Proj}_{\widehat{\mathcal{V}}}$ and that $D_S \widetilde{\mathbb{B}} D_S$ is of rank at most R . We have

$$\begin{aligned} \|\widehat{B}_1 - D_S \widetilde{\mathbb{B}} D_S\| &\leq \|\widehat{B}_1 - \text{Proj}_{\widehat{\mathcal{V}}} D_S \widetilde{B} D_S \text{Proj}_{\widehat{\mathcal{V}}}\| + \|\text{Proj}_{\widehat{\mathcal{V}}} D_S \widetilde{B} D_S \text{Proj}_{\widehat{\mathcal{V}}} - D_S \widetilde{\mathbb{B}} D_S\| \\ &\leq 2\|\text{Proj}_{\widehat{\mathcal{V}}} D_S \widetilde{B} D_S \text{Proj}_{\widehat{\mathcal{V}}} - D_S \widetilde{\mathbb{B}} D_S\|. \end{aligned}$$

□

Proof. (to Lemma 2.16) By triangle inequality we have

$$\|\widehat{B}_2 - \mathbb{B}\|_1 \leq \|\widehat{B}_2 - \widetilde{\mathbb{B}}\|_1 + \|\widetilde{\mathbb{B}} - \mathbb{B}\|_1 \quad (2.64)$$

Note that

$$\|\widetilde{\mathbb{B}} - \mathbb{B}\|_1 \leq 2 \sum_k W_k^s = 2 \sum_k e^{-e^k N/2M} = o\left(\frac{MR^2}{Nw_{\min}^2}\right)^{1/2}.$$

Apply Cauchy-Schwartz to the first term we have:

$$\begin{aligned} \|\widehat{B}_2 - \mathbb{B}\|_1 &= \sum_{i,j} |(\widehat{B}_2)_{i,j} - \mathbb{B}_{i,j}| \frac{1}{\sqrt{\rho_i \rho_j}} \sqrt{\rho_i \rho_j} \\ &\leq \|D_S(\widehat{B}_2 - \mathbb{B})D_S\|_F \sqrt{\sum_{i,j} \rho_i \rho_j} \\ &\leq \sqrt{2R} \|D_S(\widehat{B}_2 - \mathbb{B})D_S\| \\ &= O\left(\frac{\log(Nw_{\min}^2/M)}{Nw_{\min}^2/MR^2}\right)^{1/4}. \end{aligned}$$

where the second inequality used the fact that if a matrix X if of rank R then $\|X\|_F \leq \sqrt{R}\|X\|$.

□

Proof. (to Lemma 2.17) After removing the abnormally heavy rows and columns, we have that each row /column corresponding to a word in $\widehat{\mathcal{L}}_k$ (defined according to Step 1 binning) has row sum and column sum less than $2\bar{\rho}_k$. We have that entries of $E_S =: (D_S \widetilde{\mathbb{B}} D_S - D_S \mathbb{B} D_S)$ are dominated by entry-wise independent zero mean sub-exponential variable with parameter $(\sigma_{i,j} = \frac{1}{\sqrt{N}}, b_{i,j} = \frac{1}{N\sqrt{\rho_i \rho_j}})$, and $b_{i,j} \leq \frac{M}{N} < 1$. Given the initialization \widehat{B}_1 from Step 3 such that $\|\widehat{B}_1 - D_S \widetilde{\mathbb{B}} D_S\| < \frac{1}{4} \sigma_{\min}(D_S \widetilde{\mathbb{B}} D_S)$, the correctness of Step 4 of Algorithm 4 follows Lemma 2.21 below. □

Lemma 2.21 (Refinement with separation condition). *Consider a noisy low rank matrix $X = UU^\top + E_S$. Assume that the noise entries are zero mean, independent and $\mathbb{E}[[E_S]_{i,j}^2] \leq \frac{1}{N}$. Assume that $\sigma_{\min}(U) > (MR^2/N)^{1/8}$. Given initialization \widehat{U} such that $\|\widehat{U}\widehat{U}^\top - UU^\top\| = \epsilon_0 \leq \frac{1}{4} \sigma_{\min}(U)^2$. We can find \widehat{X} such that*

$$\|\widehat{X} - X\|_F = O\left(\sqrt{\frac{MR}{N}}\right).$$

Proof. Let \widehat{V} and V denote the leading left singular vectors of \widehat{U} and U .

$$\|\text{Proj}_{\widehat{V}^\perp} UU^\top \text{Proj}_{\widehat{V}^\perp}\| = \|\text{Proj}_{\widehat{V}^\perp} (\widehat{U}\widehat{U}^\top - UU^\top) \text{Proj}_{\widehat{V}^\perp}\| \leq \epsilon_0.$$

First, consider the $R \times R$ matrix $\widehat{V}^\top X \widehat{V}$, we know that with large probability,

$$\|\widehat{V}^\top X \widehat{V} - \widehat{V}^\top UU^\top \widehat{V}\| = \|\widehat{V}^\top E_S \widehat{V}\| = O\left(\sqrt{\frac{R^2}{N}}\right).$$

Let $Z = (\widehat{V}^\top X \widehat{V})^{1/2}$, we know that there exists some unknown rotation matrix H_Z such that

$$\|Z - \widehat{V}^\top U H_Z\| = o(1).$$

Note that $U = \text{Proj}_{\widehat{V}} U + \text{Proj}_{\widehat{V}^\perp} U$, we have $\sigma_{\min}(U) \leq \sigma_{\min}(\widehat{V}^\top U) + \sigma_{\max}(\text{Proj}_{\widehat{V}^\perp} U)$.

By assumption of ϵ_0 we have

$$\sigma_{\min}(Z) = \sigma_{\min}(\widehat{V}^\top U) \geq \sigma_{\min}(U) - \epsilon_0^{1/2} \geq \frac{1}{2}\sigma_{\min}(U).$$

Next, consider the matrix $\widehat{V}^\top X$ we know that it can be factorized as:

$$\begin{aligned} \widehat{V}^\top X &= \widehat{V}^\top (UU^\top + E_S) \\ &= \widehat{V}^\top UH_Z(UH_Z)^\top + \widehat{V}^\top E_S \\ &= Z(UH_Z)^\top + \widehat{V}^\top E_S + o(1). \end{aligned}$$

Let $\widehat{U} = (Z^{-1}\widehat{V}^\top X)^\top$. Note that $\widehat{U} - UH_Z = (Z^{-1}\widehat{V}^\top E_S)^\top$. Thus we can bound that

$$\begin{aligned} \|\widehat{U}\widehat{U}^\top - UU^\top\|_F &= \|\widehat{U}\widehat{U}^\top - UH_Z(UH_Z)^\top\|_F \\ &\leq \|(\widehat{U} - UH_Z)(UH_Z)^\top\|_F + \|(\widehat{U} - UH_Z)\widehat{U}^\top\|_F \\ &\leq \|UH_Z Z^{-1}\widehat{V}^\top E_S\|_F + \|\widehat{U} Z^{-1}\widehat{V}^\top E_S\|_F \end{aligned}$$

We bound the two terms separately. First

$$\begin{aligned} \|UH_Z Z^{-1}\widehat{V}^\top E_S\|_F &\leq \|\widehat{V}^\top UH_Z Z^{-1}\widehat{V}^\top E_S\|_F + \|(\widehat{V}^\perp)^\top UH_Z Z^{-1}\widehat{V}^\top E_S\|_F \\ &\leq \|ZZ^{-1}\widehat{V}^\top E_S\|_F + \|(\widehat{V}^\perp)^\top U\| \|Z^{-1}\| \|\widehat{V}^\top E_S\|_F \\ &\leq \|\widehat{V}^\top E_S\|_F (1 + \sqrt{\epsilon_0}/\sigma_{\min}) \\ &\leq 2\|\widehat{V}^\top E_S\|_F \end{aligned}$$

We then bound the second term

$$\begin{aligned} \|\widehat{U} Z^{-1}\widehat{V}^\top E_S\|_F &\leq \|UH_Z Z^{-1}\widehat{V}^\top E_S\|_F + \|(Z^{-1}V^\top E_S)^\top Z^{-1}\widehat{V}^\top E_S\|_F \\ &\leq \|\widehat{V}^\top E_S\|_F (2 + \|\widehat{V}^\top E_S\|_F / \sigma_{\min}(Z)^2) \\ &\leq 6\|\widehat{V}^\top E_S\|_F, \end{aligned}$$

where the last inequality is by the assumption $\sigma_{\min}(U)^2 > (MR/N)^{1/4} > (MR/N)^{1/2} =$

$$\|\widehat{V}^\top E_S\|_F.$$

Finally by Lemma 2.20 we have that with large probability

$$\|\widehat{U}\widehat{U}^\top - UU^\top\|_F \leq 8\|\widehat{V}^\top E_S\|_F = O\left(\sqrt{\frac{MR}{N}}\right).$$

□

2.7.4 Proofs for HMM testing lower bound

Proof. (to Theorem 2.7)

$$TV(\Pr_1(G_1^N), \Pr_2(G_1^N)) \tag{2.65}$$

$$\leq TV(\Pr_1(G_1^N, \mathcal{A}), \Pr_2(G_1^N, \mathcal{A})) \tag{2.66}$$

$$\begin{aligned} &= \frac{1}{2} \sum_{G_1^N \in [M]^N, \mathcal{A} \in \binom{[M]}{M/2}} |\Pr_2(G_1^N, \mathcal{A}) - \Pr_1(G_1^N, \mathcal{A})| \\ &= \frac{1}{2} \sum_{G_1^N \in [M]^N, \mathcal{A} \in \binom{[M]}{M/2}} \Pr_1(G_1^N, \mathcal{A}) \left| \frac{\Pr_2(G_1^N, \mathcal{A})}{\Pr_1(G_1^N, \mathcal{A})} - 1 \right| \\ &\stackrel{(a)}{\leq} \frac{1}{2} \left(\sum_{G_1^N \in [M]^N, \mathcal{A} \in \binom{[M]}{M/2}} \Pr_1(G_1^N, \mathcal{A}) \left(\frac{\Pr_2(G_1^N, \mathcal{A})}{\Pr_1(G_1^N, \mathcal{A})} - 1 \right)^2 \right)^{1/2} \\ &= \frac{1}{2} \left(\sum_{G_1^N \in [M]^N, \mathcal{A} \in \binom{[M]}{M/2}} \Pr_1(G_1^N, \mathcal{A}) \left(\frac{\Pr_2(G_1^N, \mathcal{A})}{\Pr_1(G_1^N, \mathcal{A})} \right)^2 - 1 \right)^{1/2} \\ &\stackrel{(b)}{=} \frac{1}{2} \left(\left(\frac{M^N}{\binom{[M]}{M/2}} \right)^2 \sum_{G_1^N \in [M]^N, \mathcal{A} \in \binom{[M]}{M/2}} \Pr_1(G_1^N, \mathcal{A}) \left(\sum_{\mathcal{B} \in \binom{[M]}{M/2}} \Pr_2(G_1^N | \mathcal{B}) \right)^2 - 1 \right)^{1/2} \\ &\stackrel{(c)}{=} \frac{1}{2} \underbrace{\left(\frac{M^N}{\binom{[M]}{M/2}^2} \sum_{G_1^N \in [M]^N} \left(\sum_{\mathcal{B} \in \binom{[M]}{M/2}} \Pr_2(G_1^N | \mathcal{B}) \right)^2 - 1 \right)^{1/2}}_Y, \tag{2.67} \end{aligned}$$

where inequality (a) used the fact that $\mathbb{E}[X] \leq (\mathbb{E}[X^2])^{1/2}$; equality (b) used the joint distributions in (2.46) and (2.49); and equality (c) takes sum over the summand \mathcal{A}

first and makes use of the marginal probability $\Pr_1(G_1^N)$ as in (2.48).

In order to bound the term $Y = \frac{M^N}{\binom{M}{M/2}^2} \sum_{G_1^N \in [M]^N} \left(\sum_{\mathcal{B} \in \binom{M}{M/2}} \Pr_2(G_1^N | \mathcal{B}) \right)^2$ in equation (2.67), we break the square and write:

$$Y = \frac{M^N}{\binom{M}{M/2}^2} \sum_{\mathcal{B}, \mathcal{B}' \in \binom{M}{M/2}} \sum_{G_1^N \in [M]^N} \Pr_2(G_1^N | \mathcal{B}) \Pr_2(G_1^N | \mathcal{B}'). \quad (2.68)$$

In Claim 2.5 below we explicitly compute the term $\sum_{G_1^N \in [M]^N} \Pr_2(G_1^N | \mathcal{B}) \Pr_2(G_1^N | \mathcal{B}')$ for any two subsets $\mathcal{B}, \mathcal{B}' \in \binom{M}{M/2}$. Then in Claim 2.6 we compute the sum over $\mathcal{B}, \mathcal{B}'$ and bound $Y \leq \frac{1}{\sqrt{1 - \frac{2(1-2t)N}{M}}}$. To conclude, we can bound the total variation distance as:

$$TV(\Pr_1(G_1^N), \Pr_2(G_1^N)) \leq \frac{1}{2} \sqrt{\frac{1}{\sqrt{1 - \frac{2(1-2t)N}{M}}} - 1}.$$

In the case that $N \leq M$, this is bounded as $TV(\Pr_1(G_1^N), \Pr_2(G_1^N)) \leq \frac{\sqrt{1-2t}}{2} \sqrt{\frac{N}{M}}$, which vanishes as $N = o(M)$ for any constant transition probability t .

□

Claim 2.5. *In the same setup of Theorem 2.7, given two subsets $\mathcal{B}, \mathcal{B}' \in \mathcal{M}$ and $|\mathcal{B}| = |\mathcal{B}'| = M/2$, let $\bar{\mathcal{B}} = \mathcal{M} \setminus \mathcal{B}, \bar{\mathcal{B}}' = \mathcal{M} \setminus \mathcal{B}'$ denote the corresponding complement. Define $x \in [0, 1]$ to be:*

$$x = |\mathcal{B} \cap \mathcal{B}'| / (M/2). \quad (2.69)$$

Let $\gamma_1(x) = \frac{1}{2} \left(1 + (1-2t)^2 + \sqrt{(1 - (1-2t)^2)^2 + (2(1-2t))^2 x^2} \right)$ and let $\gamma_2(x) = \frac{1}{2} \left(1 + (1-2t)^2 - \sqrt{(1 - (1-2t)^2)^2 + (2(1-2t))^2 x^2} \right)$ be functions of $|\mathcal{B} \cap \mathcal{B}'|$ and t .

We have:

$$\sum_{G_1^N \in [M]^N} \Pr_2(G_1^N | \mathcal{B}) \Pr_2(G_1^N | \mathcal{B}') = \frac{1}{(M/2)^N} \left(\frac{\gamma_1(x)^N (1 - 2\gamma_2(x)) - \gamma_2(x)^N (1 - 2\gamma_1(x))}{2(\gamma_1(x) - \gamma_2(x))} \right),$$

Proof. (to Claim 2.5)

(1) For an instance of 2-state HMM which support for q is specified by set $\mathcal{B} \in \binom{M}{M/2}$, consider two consecutive outputs (g_{n-1}, g_n) . We first show how to compute the probability $\Pr_2(g_n|g_{n-1}, \mathcal{B})$.

Given \mathcal{B} and another set \mathcal{B}' , we can partition the vocabulary \mathcal{M} into four subsets as:

$$\mathcal{M}_1 = \mathcal{B} \cap \mathcal{B}', \quad \mathcal{M}_2 = \mathcal{B} \cap \overline{\mathcal{B}'}, \quad \mathcal{M}_3 = \overline{\mathcal{B}} \cap \mathcal{B}', \quad \mathcal{M}_4 = \overline{\mathcal{B}} \cap \overline{\mathcal{B}'}$$

Note that we have $|\mathcal{M}_1| = |\mathcal{M}_4| = xM/2$ and $|\mathcal{M}_2| = |\mathcal{M}_3| = (1-x)M/2$.

Define a subset of tuples $\mathcal{J}_B \subset [4]^2$ to be

$$\mathcal{J}_B = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 3), (3, 4), (4, 3), (4, 4)\}, \quad \mathcal{J}_B^c = [4]^2 \setminus \mathcal{J}_B.$$

If $g_{n-1} \in \mathcal{M}_{j'}, g_n \in \mathcal{M}_j$ and $(j', j) \in \mathcal{J}_B$, we know that the hidden state for the HMM associated with set \mathcal{B} does not change between time slot $n-1$ and n , namely $s_{n-1} = s_n$. Thus $\Pr_2(g_n|g_{n-1}, \mathcal{B}) = \Pr_2(s_n|s_{n-1}, \mathcal{B})\Pr_2(g_n|s_n, \mathcal{B}) = \frac{1-t}{M/2}$. Also, if $(j', j) \in \mathcal{J}_B^c$, we know that there is the state transition and we have $\Pr_2(g_n|g_{n-1}, \mathcal{B}) = \frac{t}{M/2}$.

Similarly, for the 2-state HMM associated with set \mathcal{B}' , we can define the set of tuples

$$\mathcal{J}_{B'} = \{(1, 1), (1, 3), (3, 1), (3, 3), (2, 2), (2, 4), (4, 2), (4, 4)\}, \quad \mathcal{J}_{B'}^c = [4]^2 \setminus \mathcal{J}_{B'}.$$

Here $\Pr_2(g_n|g_{n-1}, \mathcal{B}') = \frac{1-t}{M/2}$ if $(j', j) \in \mathcal{J}_{B'}$ and equals $\frac{t}{M/2}$ if $(j', j) \in \mathcal{J}_{B'}^c$.

(2) Next, we show how to compute the target sum of the claim statement in a recursive way.

For fixed sets \mathcal{B} and \mathcal{B}' , define $F_{n,j}$ for $n \leq N$ and $j = 1, 2, 3, 4$ as below

$$F_{n,j} = \sum_{G_1^n \in [M]^n} \Pr_2(G_1^n | \mathcal{B}) \Pr_2(G_1^n | \mathcal{B}') \mathbf{1}[g_n \in \mathcal{M}_j],$$

and the target sum is $\sum_{G_1^N \in [M]^N} \Pr_2(G_1^N | \mathcal{B}) \Pr_2(G_1^N | \mathcal{B}') = \sum_{j=1:4} F_{N,j}$. Also, we have

that

$$F_{1,j} = |\mathcal{M}_j|/M^2.$$

Making use of the recursive property of the probability rule of the 2-state HMM as in (2.47), we can write the following recursion in terms of $F_{n,j}$ for $n \geq 2$:

$$\begin{aligned} F_{n,j} &= \sum_{G_1^n \in [M]^n} \Pr_2(G_1^n | \mathcal{B}) \Pr_2(G_1^n | \mathcal{B}') \mathbf{1}[g_n \in \mathcal{M}_j] \\ &= \sum_{G_1^n \in [M]^n} \Pr_2(G_1^{n-1} | \mathcal{B}) \Pr_2(G_1^{n-1} | \mathcal{B}') \Pr_2(g_n | G_1^{n-1}, \mathcal{B}) \Pr_2(g_n | G_1^{n-1}, \mathcal{B}') \\ &\quad \sum_{j'=1:4} \mathbf{1}[g_{n-1} \in \mathcal{M}_{j'}, g_n \in \mathcal{M}_j] \\ &= \sum_{G_1^{n-1} \in [M]^{n-1}} \Pr_2(G_1^{n-1} | \mathcal{B}) \Pr_2(G_1^{n-1} | \mathcal{B}') \sum_{g_n \in [M]} \\ &\quad \sum_{j'=1:4} \mathbf{1}[g_{n-1} \in \mathcal{M}_{j'}, g_n \in \mathcal{M}_j] \Pr_2(g_n | g_{n-1}, \mathcal{B}) \Pr_2(g_n | g_{n-1}, \mathcal{B}') \\ &= |\mathcal{M}_j| \sum_{j'=1:4} F_{n-1,j'} \left(\frac{1-t}{M/2} \mathbf{1}[(j', j) \in \mathcal{J}_B] + \frac{t}{M/2} \mathbf{1}[(j, j') \in \mathcal{J}_B^c] \right) \\ &\quad \left(\frac{1-t}{M/2} \mathbf{1}[(j', j) \in \mathcal{J}_{B'}] + \frac{t}{M/2} \mathbf{1}[(j, j') \in \mathcal{J}_{B'}^c] \right) \end{aligned}$$

where we used the probability $\Pr_2(g_n | g_{n-1}, \mathcal{B})$ derived in (1).

Equivalently we can write the recursion as:

$$\begin{pmatrix} F_{n,1} \\ F_{n,2} \\ F_{n,3} \\ F_{n,4} \end{pmatrix} = \frac{1}{(M/2)} D_x T \begin{pmatrix} F_{n-1,1} \\ F_{n-1,2} \\ F_{n-1,3} \\ F_{n-1,4} \end{pmatrix},$$

for diagonal matrix $D_x = \begin{pmatrix} x & & & \\ & 1-x & & \\ & & 1-x & \\ & & & x \end{pmatrix}$ and the symmetric stochastic matrix

T given by

$$T = \begin{pmatrix} (1-t)^2 & (1-t)t & (1-t)t & t^2 \\ (1-t)t & (1-t)^2 & t^2 & (1-t)t \\ (1-t)t & t^2 & (1-t)^2 & (1-t)t \\ t^2 & (1-t)t & (1-t)t & (1-t)^2 \end{pmatrix} = \sum_{i=1}^4 \lambda_i v_i v_i^\top,$$

where the singular values and singular vectors of T are specified as follows: $\lambda_1 = 1$, $\lambda_4 = (1-2t)^2$, and $v_1 = \frac{1}{2}[1, 1, 1, 1]^\top$, $v_4 = \frac{1}{2}[1, -1, -1, 1]^\top$. And $\lambda_2 = \lambda_3 = 1-2t$ with $v_2 = \frac{1}{\sqrt{2}}[0, 1, -1, 0]^\top$ and $v_3 = \frac{1}{\sqrt{2}}[1, 0, 0, -1]^\top$.

Note that we can write $(F_{1,1}, F_{1,2}, F_{1,3}, F_{1,4})^\top = \frac{M/2}{M^2} D_x(1, 1, 1, 1)^\top$.

(3) Finally we can compute the target sum as:

$$\begin{aligned} & \sum_{G_1^N \in [M]^N} \Pr_2(G_1^N | \mathcal{B}) \Pr_2(G_1^N | \mathcal{B}') \\ &= \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} F_{N,1} & F_{N,2} & F_{N,3} & F_{N,4} \end{pmatrix}^\top \\ &= \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \frac{1}{(M/2)^{N-1}} (D_x T)^{N-1} \frac{M/2}{M^2} D_x \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}^\top \\ &\stackrel{(a)}{=} \frac{1}{M^N} v_1^\top (2D_x T)^N v_1 \\ &= \frac{1}{M^N} (1 \ 0) \underbrace{\begin{pmatrix} 1 & (2x-1) \\ (1-2t)^2(2x-1) & (1-2t)^2 \end{pmatrix}^N}_{H(x)^N} (1 \ 0)^\top \\ &\stackrel{(b)}{=} \frac{1}{M^N} \left(\frac{\gamma_1(x)\gamma_2(x)^N - \gamma_2(x)\gamma_1(x)^N}{\gamma_1(x) - \gamma_2(x)} + \frac{\gamma_1(x)^N - \gamma_2(x)^N}{\gamma_1(x) - \gamma_2(x)} \right) \\ &= \frac{1}{M^N} \frac{\gamma_1^N(1-\gamma_2) - \gamma_2^N(1-\gamma_1)}{\gamma_1 - \gamma_2}, \end{aligned}$$

where in (a) we used the fact that

$$2D_x T v_1 = v_1 + (2x-1)v_4, \text{ and } 2D_x T v_4 = (1-2t)^2((2x-1)v_1 + v_4).$$

In (b) we used the Cayley-Hamilton theorem to obtain that for 2×2 matrix $H(x)$ parameterized by x and with 2 distinct eigenvalue $\gamma_1(x)$ and $\gamma_2(x)$, its power can be written as $H(x)^N = \frac{\gamma_1(x)\gamma_2(x)^N - \gamma_2(x)\gamma_1(x)^N}{\gamma_1(x) - \gamma_2(x)} I_{2 \times 2} + \frac{\gamma_1(x)^N - \gamma_2(x)^N}{\gamma_1(x) - \gamma_2(x)} H(x)$. Moreover, the distinct eigenval-

ues of the 2×2 matrix $H(x)$ can be written explicitly as follows:

$$\gamma_1(x) = \frac{1}{2} \left(1 + (1 - 2t)^2 + \sqrt{(1 - (1 - 2t)^2)^2 + (2(1 - 2t))^2 x^2} \right), \quad (2.70)$$

$$\gamma_2(x) = \frac{1}{2} \left(1 + (1 - 2t)^2 - \sqrt{(1 - (1 - 2t)^2)^2 + (2(1 - 2t))^2 x^2} \right). \quad (2.71)$$

where recall that we defined $x = \frac{|\mathcal{B} \cap \mathcal{B}'|}{M/2}$ so $0 \leq x \leq 1$, also we have the transition probability $0 < t < 1/2$ to be a constant, therefore we have $\gamma_1 > \gamma_2$ to be two distinct real roots. \square

The next claim makes use of the above claim and bounds the right hand side of (2.68).

Claim 2.6. *In the same setup of Theorem 2.7, we have*

$$Y = \frac{M^N}{\binom{M}{M/2}^2} \sum_{\mathcal{B}, \mathcal{B}' \in \binom{M}{M/2}} \sum_{G_1^N \in [M]^N} \Pr_2(G_1^N | \mathcal{B}) \Pr_2(G_1^N | \mathcal{B}') \leq \frac{1}{\sqrt{1 - \frac{2(1-2t)^N}{M}}}.$$

Proof. (to Claim 2.6)

Define $f(x) = \frac{\gamma_1(x)^N(1-\gamma_2(x))-\gamma_2(x)^N(1-\gamma_1(x))}{\gamma_1(x)-\gamma_2(x)}$ with $\gamma_1(x)$ and $\gamma_2(x)$ defined in (2.70) and (2.71) as functions of x . Recall that $x = |\mathcal{B} \cap \mathcal{B}'|/(xM/2) \in [0, 1]$.

Use the result of Claim 2.5 we have:

$$\begin{aligned} Y &= \frac{M^N}{\binom{M}{M/2}^2} \sum_{\mathcal{B}, \mathcal{B}' \in \binom{M}{M/2}} \frac{1}{M^N} f(x) \\ &\stackrel{(a)}{=} \frac{1}{\binom{M}{M/2}^2} \binom{M}{M/2} \sum_{i=1}^{M/2} \binom{M/2}{i}^2 f\left(\frac{i}{M/2}\right) \\ &= \frac{1}{\binom{M}{M/2}} \sum_{i=1}^{M/2} \binom{M/2}{i}^2 f\left(\frac{i}{M/2}\right), \end{aligned} \quad (2.72)$$

where equality (a) is obtained by counting the number of subsets $\mathcal{B}, \mathcal{B}' \in \binom{M}{M/2}$: for each fixed \mathcal{B} , there are $\binom{M/2}{i}$ choices of \mathcal{B}' such that $|\mathcal{B} \cap \mathcal{B}'| = i$.

Next we approximately bound Y when M is asymptotically large. First, note that

$\gamma_1(0) = 1$ and $\gamma_1(1) = 1 + (1 - 2t)^2$, we can bound $\gamma_1(x)$ as by exponential function:

$$\gamma_1(x) = \frac{1}{2} \left(1 + (1 - 2t)^2 + \sqrt{(1 - (1 - 2t)^2)^2 + (2(1 - 2t))^2 x^2} \right) \leq e^{(1-2t)(1-2x)^2},$$

Then note that for N increasing with M and thus asymptotically large, we have $\gamma_2^N(1 - \gamma_1) = o(1)$, so we bound $f(x)$ by:

$$\lim_{M \rightarrow \infty} f(x) \approx \gamma_1(x)^N \frac{1 - \gamma_2(x)}{\gamma_1(x) - \gamma_2(x)} \leq e^{(1-2t)(1-2x)^2 N},$$

where we used the fact that $\frac{1 - \gamma_2(x)}{\gamma_1(x) - \gamma_2(x)} = \frac{1}{2} \left(1 + 1/\sqrt{1 + x^2 \left(\frac{2(1-2t)}{1 - (1-2t)^2} \right)^2} \right) \leq 1$. and that $1/2 \leq \gamma_1 \leq 1$.

Second, we use Stirling's approximation for the combinatorial coefficients $\binom{M/2}{i}^2$ and $\binom{M}{M/2}$,

$$\begin{aligned} \binom{M}{M/2} &\approx \frac{4^{M/2}}{\sqrt{\pi M/2}}, \\ \binom{M/2}{i}^2 &\approx \left(\binom{M/2}{M/4} e^{-(M/2 - 2i)^2 / 2(M/2)} \right)^2 \\ &\approx \frac{4^{M/2}}{\pi M/4} e^{-2M(i/(M/2) - 1/2)^2}, \quad \text{for } \log M \leq i \leq (M/2) - \log M. \end{aligned}$$

Finally we can approximately bound Y in (2.72) as follows:

$$\begin{aligned} Y &\approx \frac{1}{\sqrt{\pi M}} \frac{4}{\sqrt{2}} \sum_{i=1}^{M/2} e^{-2M(\frac{i}{M/2} - \frac{1}{2})^2} f\left(\frac{i}{M/2}\right) + \frac{2}{\binom{M}{M/2}} \sum_{i=1}^{\log M} \binom{M/2}{i} f\left(\frac{i}{M/2}\right) \\ &\leq \frac{1}{\sqrt{\pi M}} \frac{4}{\sqrt{2}} \left(\frac{M}{2} \int_{y=-1/2}^{1/2} e^{-2My^2 + 4(1-2t)y^2 N} \right) + o(1) \\ &= \sqrt{\frac{2M}{\pi}} \int_{y=-1/2}^{1/2} e^{-y^2 2M(1-2(1-2t)N/M)} \\ &= \frac{1}{\sqrt{1 - \frac{2(1-2t)N}{M}}}, \end{aligned}$$

where the second inequality is because for M asymptotically large and $N = O(M)$,

we have

$$\begin{aligned} & \frac{2}{\binom{M}{M/2}} \sum_{i=1}^{\log M} \binom{M/2}{i} f\left(\frac{i}{M/2}\right) \\ & \leq 2\sqrt{\pi(M/2)} 4^{-M/2} (\log M) (M/2)^{\log M} e^{4(1-2t)(2\log M/M)^2 N} \\ & = o(1). \end{aligned}$$

□

2.7.5 Analyze truncated SVD

The reason that truncated SVD does not concentrate at the optimal rate is as follows. What truncated SVD actually optimizes is the spectral distance from the estimator to the empirical average (minimizing $\|\widehat{B} - \frac{1}{N}B_N\|_2$), yet not to the expected matrix \mathbb{B} . It is only “optimal” in some very special setup, for example when $(\frac{1}{N}B_N - \mathbb{B})$ are entry-wise i.i.d. Gaussian. In the asymptotic regime when $N \rightarrow \infty$ it is indeed true that under mild condition any sampling noise converges to i.i.d Gaussian. However in the sparse regime where $N = \Omega(M)$, the sampling noise from the probability matrix is very different from additive Gaussian noise.

Claim 2.7 (Truncated SVD has sample complexity super linear). *In order to achieve ϵ accuracy, the sample complexity of rank-2 truncated SVD estimator is in given by $N = O(M^2 \log M)$.*

Example 1: $a = b = w = 1/2$, dictionary given by

$$\begin{aligned} p &= \left[\frac{1+C_\Delta}{M}, \dots, \frac{1+C_\Delta}{M}, \frac{1-C_\Delta}{M}, \dots, \frac{1-C_\Delta}{M} \right], \\ q &= \left[\frac{1-C_\Delta}{M}, \dots, \frac{1-C_\Delta}{M}, \frac{1+C_\Delta}{M}, \dots, \frac{1+C_\Delta}{M} \right]. \end{aligned}$$

Sample complexity is $O(M \log M)$.

Example 2: modify Example 1 so that a constant fraction of the probability mass lies in a common word, namely $p_1 = q_1 = 1/2\rho_1 = 0.1$, while the marginal probability as well as the separation in all the other words are roughly uniform. Sample complexity

is $O(M^2 \log N)$.

Proof. (to Claim 2.7 (Truncated SVD has sample complexity super linear))

(1) We formalize this and examine the sample complexity of t-SVD by applying Bernstein matrix inequality. The concentration of the empirical average matrix at the following rate:

$$\Pr(\|\frac{1}{N}B_N - \mathbb{B}\| \geq t) \leq e^{-\frac{(Nt)^2}{NVar + BNt/3} + \log(M)},$$

where $Var = \|\mathbb{E}[e_i e_i^\top]\|_2 = \|\text{diag}(\rho)\|_2 = \max_i \rho_i$, and $B = \max_{i,j} \|e_i e_j\|_2 = 1$.

Therefore, with probability at least $1 - \delta$, we have that

$$\|\frac{1}{N}B_N - \mathbb{B}\| \leq \sqrt{\frac{\max_i \rho_i \log(M/\delta)}{N}} + \frac{1}{3} \frac{1}{N} \log(M/\delta). \quad (2.73)$$

Since $\|x\|_1 \leq \sqrt{M}\|x\|_2$, in order to guarantee that $\|\hat{\Delta} - \Delta\|_1 \leq \epsilon$, it suffices to ensure that $\|\hat{\Delta} - \Delta\|_2 \leq \epsilon/\sqrt{M}$. Note that the leading two eigenvectors are given by $\sigma_1(\mathbb{B}) \geq \|\rho\|_2 = 1/\sqrt{M}$ and $\sigma_2(\mathbb{B}) = \|\Delta\|_2 = C_\Delta/\sqrt{M}$. Assume that we have the exact marginal probability ρ , by Davis-Kahan, it suffices to ensure that

$$\|\frac{1}{N}B_N - \mathbb{B}\|_2 \leq \epsilon \frac{\|\Delta\|_2}{\sqrt{M}}.$$

Example 1. Consider the example of (p, q) in community detection problem, where the marginal probability ρ_i is roughly uniform. We have $\|\Delta\|_2 = C_\Delta/\sqrt{M}$ and $\max_i \rho_i = 1/M$, and the concentration bound becomes

$$\|\frac{1}{N}B_N - \mathbb{B}\| \leq \sqrt{\frac{\log(M/\delta)}{MN}}, \quad (2.74)$$

and by requiring

$$\sqrt{\frac{\log(M/\delta)}{MN}} \leq \epsilon \frac{\|\Delta\|_2}{\sqrt{M}} = \epsilon \frac{C_\Delta}{M}$$

we get a sample complexity bound $N = \Omega(M \log(M/\delta))$, which is worse than the

lower bound by a $\log(M)$ factor.

Example 2. Moreover, modify Example 1 so that a constant fraction of the probability mass lies in a common word, namely $p_1 = q_1 = 1/2\rho_1 = 0.1$, while the marginal probability as well as the separation in all the other words are roughly uniform. In this case, $\|\Delta\|_2$ is still roughly C_Δ/\sqrt{M} , however we have $\max_i \rho_i = 0.1$, and the sample complexity becomes $N = \Omega(M^2 \log(M/\delta))$. This is even worse than the first example, as the same separation gets swamped by the heavy common words.

(2) (square root of the empirical marginal scaling (from 1st batch of samples) on both side of the empirical count matrix (from 2nd batch of samples)). \square

Take a closer look at the above proof and we can identify two misfortunes that make the truncated SVD deviate from linear sample complexity:

1. In the worst case, the nonuniform marginal probabilities costs us an M factor in the first component of Bernstein's inequality;
2. We pay another $\log(M)$ factor for the spectral concentration of the $M \times M$ random matrix.

To resolve these two issues, the two corresponding key ideas of Phase I algorithm are “binning” and “regularization”:

1. “Binning” means that we partition the vocabulary according to the marginal probabilities, so that for the words in each bin, their marginal probabilities are roughly uniform. If we are able to apply spectral method in each bin separately, we could possibly get rid of the M factor.
2. Now restrict our attention to the diagonal block of the empirical average matrix $\frac{1}{N}B_N$ whose indices corresponding to the words in a bin. Assume that the bin has sufficiently many words, so that the expected row sum and column sum are at least constant, namely the effective number of samples is at least in the order of the number of words in the bin.

We apply regularized spectral method for the empirical average with indices restricted to the bin. By “regularization” we mean removing the rows and

column, whose row and column sum are much higher than the expected row sum, from the empirical. Then we apply t-SVD to the remaining. This regularization idea is motivated by the community detection literature in the sparse regime, where the total number of edges of the random network is only linear in the number of nodes.

2.7.6 Auxiliary Lemmas

Lemma 2.22 (Wedin's theorem applied to rank-1 matrix). *Denote symmetric matrix $X = vv^\top + E$. Let $\widehat{v}\widehat{v}^\top$ denote the rank-1 truncated SVD of X . There is a positive universal constant C such that*

$$\min\{\|v - \widehat{v}\|, \|v + \widehat{v}\|\} \leq \begin{cases} \frac{C\|E\|}{\|v\|} & \text{if } \|v\|^2 > C\|E\|; \\ C\|E\|^{1/2} & \text{if } \|v\|^2 < C\|E\|. \end{cases}$$

Lemma 2.23 (Chernoff Bound for Poisson variables).

$$\Pr(\text{Poi}(\lambda) \geq x) \leq e^{-\lambda} \left(\frac{x}{e\lambda}\right)^{-x}, \quad \text{for } x > \lambda,$$

$$\Pr(\text{Poi}(\lambda) \leq x) \leq e^{-\lambda} \left(\frac{x}{e\lambda}\right)^{-x}, \quad \text{for } x < \lambda.$$

Lemma 2.24 (Upper bound of Poisson tails (Proposition 1 in [49])). *Assume $\lambda > 0$, consider the Poisson distribution $\text{Poi}(\lambda)$.*

(1) *if $0 \leq n < \lambda$, the left tail can be upper bounded by:*

$$\Pr(\text{Poi}(\lambda) \leq n) \leq \left(1 - \frac{n}{\lambda}\right)^{-1} \Pr(\text{Poi}(\lambda) = n).$$

(2) *if $n > \lambda - 1$, for any $m \geq 1$, the right tail can be upper bounded by:*

$$\Pr(\text{Poi}(\lambda) \geq n) \leq \left(1 - \left(\frac{\lambda}{n+1}\right)^m\right)^{-1} \sum_{i=n}^{n+m-1} \Pr(\text{Poi}(\lambda) = i).$$

Corollary 2.5. *Let $\lambda > C$ for some large universal constant C . For any constant*

$c' > e$, $0 \leq c < 1/2$, we have the following Poisson tail bounds:

$$\begin{aligned}\Pr(\text{Poi}(\lambda) \leq c\lambda) &\leq 2e^{-\lambda/2}, \\ \Pr(\text{Poi}(\lambda) \geq c'\lambda) &\leq 2e^{-c'\lambda}.\end{aligned}$$

Proof. Apply Stirling's bound for λ large, we have $\lambda! \geq (\frac{\lambda}{e})^\lambda$. Then, the bound in Lemma 2.24 (1) can be written as

$$\begin{aligned}\Pr(\text{Poi}(\lambda) \leq c\lambda) &\leq (1-c)^{-1} \Pr(\text{Poi}(\lambda) = c\lambda) \\ &\leq 2e^{-\lambda}(\lambda)^{c\lambda}/(c\lambda)! \\ &\leq 2e^{-\lambda}(\lambda)^{c\lambda}/(c\lambda e^{-1})^{c\lambda} \\ &\leq 2e^{-\lambda+c\lambda \log(e/c)} \\ &\leq 2e^{-\lambda/2},\end{aligned}$$

where in the second inequality we used the assumption that $c < 1/2$, and in the last inequality we used the inequality $1 - c \log(e/c) \geq 1/2$ for all $0 \leq c < 1$.

Similarly, set $m = 1$ in Lemma 2.24 (2), we can write the bound as

$$\begin{aligned}\Pr(\text{Poi}(\lambda) \geq c'\lambda) &\leq \left(1 - \frac{\lambda}{c'\lambda + 1}\right)^{-1} \Pr(\text{Poi}(\lambda) = c'\lambda) \\ &\leq (1 - 1/c')^{-1} e^{-\lambda}(\lambda)^{c'\lambda}/(c'\lambda)! \\ &\leq 2e^{-\lambda}(\lambda)^{c'\lambda}/(c'\lambda e^{-1})^{c'\lambda} \\ &\leq 2e^{-c'\lambda \log(c'/e) - 1} \\ &\leq 2e^{-c'\lambda},\end{aligned}$$

where in both the second and the last inequality we used the assumption that $c' > e$ and λ is a large constant. \square

Lemma 2.25 (Slight variation of Vershynin's theorem (Poisson instead of Bernoulli)).
Consider a random matrix A of size $M \times M$, where each entry follows an independent Poisson distribution $A_{i,j} \sim \text{Poi}(P_{i,j})$. Define $d_{\max} = M \max_{i,j} P_{i,j}$. For any $r \geq 1$, the

following holds with probability at least $1 - M^{-r}$. Consider any subset consisting of at most $10\frac{M}{d_{\max}}$, and decrease the entries in the rows and the columns corresponding to the indices in the subset in an arbitrary way. Then for some universal large constant c the modified matrix A' satisfies:

$$\|A' - \mathbb{E}A\| \leq Cr^{3/2}(\sqrt{d_{\max}} + \sqrt{d'}),$$

where d' denote the maximal row sum in the modified random matrix.

Proof. The original proof in [75] is for independent Bernoulli entries $A_{i,j} \sim \text{Ber}(P_{i,j})$. The specific form of the distribution is only used when bounding the $\ell_\infty \rightarrow \ell_1$ norm of the adjacency matrix by applying Bernstein inequality:

$$\Pr\left(\sum_{i,j=1}^M X_{i,j} > M^2t\right) \leq \exp\left(\frac{M^2t^2/2}{\frac{1}{M^2} \sum_{i,j}^M P_{i,j} + t/3}\right)$$

where $X_{i,j} = (A_{i,j} - \mathbb{E}[A_{i,j}])x_i y_j$ for any fixed $x_i, y_j \in \{+1, -1\}$.

Recall that a random variable X is sub-exponential if there are non-negative parameters (σ, b) such that $\mathbb{E}[e^{t(X-\mathbb{E}[X])}] \leq e^{t^2\sigma^2/2}$ for all $|t| < \frac{1}{b}$. Note that a Poisson variables $X \sim \text{Poi}(\lambda)$ has sub-exponential tail bound with parameters $(\sigma = \sqrt{2\lambda}, b = 1)$, since

$$\log(\mathbb{E}[e^{t(X-\lambda)}]e^{-t^2\sigma^2/2}) = (\lambda(e^t - 1) - \lambda t) - \lambda t^2 \leq 0, \text{ for } |t| < 1.$$

Therefore, when the entries are replaced by independent Poisson entries $A_{i,j} \sim \text{Poi}(P_{i,j})$, we can apply Bernstein inequality for sub-exponential random variables to obtain similar concentration bound:

$$\Pr\left(\sum_{i,j=1}^M X_{i,j} > M^2t\right) \leq \exp\left(\frac{M^2t^2/2}{\frac{1}{M^2} \sum_{i,j}^M \text{Var}(X_{i,j}) + bt}\right) \leq \exp\left(\frac{M^2t^2/2}{2\frac{1}{M^2} \sum_{i,j}^M P_{i,j} + t}\right).$$

The same arguments of the proof in [75] then go through. \square

Chapter 3

Learning Gaussian Mixtures in High Dimensions

3.1 Problem Statement

3.1.1 Formulation

In a Gaussian mixture model, there are k unknown n -dimensional multivariate Gaussian distributions. Samples are generated by first picking one of the k Gaussians, then drawing a sample from that Gaussian distribution. Given samples from the mixture distribution, our goal is to estimate the means and covariance matrices of these underlying Gaussian distributions.

3.1.2 Related Work

Learning mixtures of Gaussians is a fundamental problem in statistics and learning theory, whose study dates back to [94]. Gaussian mixture models arise in numerous areas including physics, biology and the social sciences ([82, 115]), as well as in image processing ([100]) and speech ([95]).

In a Gaussian mixture model, there are k unknown n -dimensional multivariate Gaussian distributions. Samples are generated by first picking one of the k Gaussians, then drawing a sample from that Gaussian distribution. Given samples from the

mixture distribution, our goal is to estimate the means and covariance matrices of these underlying Gaussian distributions¹.

This problem has a long history in theoretical computer science. The seminal work of [37] gave an algorithm for learning spherical Gaussian mixtures when the means are well separated. Subsequent works ([39, 103, 121, 29]) developed better algorithms in the well-separated case, relaxing the spherical assumption and the amount of separation required.

When the means of the Gaussians are not separated, after several works ([22, 64]), [23] and [85] independently gave algorithms that run in polynomial time and with polynomial number of samples for a fixed number of Gaussians. However, both running time and sample complexity depend *super* exponentially on the number of components k^2 . Their algorithm is based on the *method of moments* introduced by [94]: first estimate the $O(k)$ -order moments of the distribution, then try to find the parameters that agree with these moments. [85] also show that the exponential dependency of the sample complexity on the number of components is necessary, by constructing an example of two mixtures of Gaussians with very different parameters, yet with exponentially small statistical distance.

Recently, [57] applied spectral methods to learning mixture of spherical Gaussians. When $n \geq k + 1$ and the means of the Gaussians are linearly independent, their algorithm can learn the model in polynomial time and with polynomial number of samples. This result suggests that the lower bound example in [85] is only a *degenerate* case in high dimensional space. In fact, *most* (in general position) mixture of spherical Gaussians are *easy* to learn. This result is also based on the method of moments, and only uses second and third moments. Several follow-up works ([25, 12]) use higher order moments to get better dependencies on n and k .

However, the algorithm in [57] as well as in the follow-ups all make strong requirements on the covariance matrices. In particular, most of them only apply to learning mixture of spherical Gaussians. For mixture of Gaussians with general co-

¹ This is different from the problem of density estimation considered in [45, 33]

² In fact, it is in the order of $O(e^{O(k)^k})$ as shown in Theorem 11.3 in [120].

variance matrices, the best known result is still [23] and [85], which algorithms are not polynomial in the number of components k . This leads to the following natural question:

Question: *Is it possible to learn most mixture of Gaussians in polynomial time using a polynomial number of samples?*

Our Results We give an algorithm that learns *most* mixture of Gaussians in high dimensional space (when $n \geq \Omega(k^2)$), and the argument is formalized under the *smoothed analysis* framework first proposed in [107].

In the smoothed analysis framework, the adversary first choose an arbitrary mixture of Gaussians. Then the mean vectors and covariance matrices of this Gaussian mixture are randomly *perturbed* by a small amount ρ ³. The samples are then generated from the Gaussian mixture model with the perturbed parameters. The goal of the algorithm is to learn the perturbed parameters from the samples.

The smoothed analysis framework is a natural bridge between worst-case and average-case analysis. On one hand, it is similar to worst-case analysis, as the adversary chooses the initial instance, and the perturbation allowed is small. On the other hand, even with small perturbation, we may hope that the instance be different enough from degenerate cases. A successful algorithm in the smoothed analysis setting suggests that the bad instances must be very “sparse” in the parameter space: they are highly unlikely in any small neighborhood of any instance. Recently, the smoothed analysis framework has also motivated several research work ([65] [25]) in analyzing learning algorithms.

In the smoothed analysis setting, we show that it is easy to learn most Gaussian mixtures:

Theorem 3.1. *(informal statement of Theorem 3.4) In the smoothed analysis setting, when $n \geq \Omega(k^2)$, given samples from the perturbed n -dimensional Gaussian mixture model with k components, there is an algorithm that learns the correct parameters up to accuracy ϵ with high probability, using polynomial time and number of samples.*

³See Definition 3.2 in Section 3.2 for the details.

An important step in our algorithm is to learn Gaussian mixture models whose components all have mean zero, which is also a problem of independent interest ([128]). Intuitively this is also a “hard” case, as there is no separation in the means. Yet algebraically, this case gives rise to a novel tensor decomposition algorithm. The ideas for solving this decomposition problem are then generalized to tackle the most general case.

Theorem 3.2. *(informal statement of Theorem 3.5) In the smoothed analysis setting, when $n \geq \Omega(k^2)$, given samples from the perturbed mixture of zero-mean n -dimensional Gaussian mixture model with k components, there is an algorithm that learns the parameters up to accuracy ϵ with high probability, using polynomial running time and number of samples.*

Organization We first focus on learning mixtures of zero-mean Gaussians. The proposed algorithm for this special case contains most of the new ideas and techniques. In Section 1.3.1 we introduce the notations for matrices and tensors which are used to handle higher order moments throughout the discussion. Then in Section 3.2 we introduce the smoothed analysis model for learning mixture of Gaussians and discuss the moment structure of mixture of Gaussians, then we formally state our main theorems. Section 3.3.1 outlines our algorithm for learning zero-mean mixture of Gaussians. The details of the steps are presented in Section 3.3.2. The detailed proofs for the correctness and the robustness are deferred to Appendix (Sections 3.4.1 to 3.4.3). In Section 3.3.3 we briefly discuss how the ideas for zero-mean case can be generalized to learning mixture of nonzero Gaussians, for which the detailed algorithm and the proofs are deferred to Appendix 3.4.5.

3.2 Main results

In this section, we first formally introduce the smoothed analysis framework for our problem and state our main theorems. Then we will discuss the structure of the moments of Gaussian mixtures, which is crucial for understanding our method of

moments based algorithm.

Smoothed Analysis for Learning Mixture of Gaussians Let $\mathcal{G}_{n,k}$ denote the class of Gaussian mixtures with k components in \mathbb{R}^n . A distribution in this family is specified by the following parameters: the mixing weights ω_i , the mean vectors $\mu^{(i)}$ and the covariance matrices $\Sigma^{(i)}$, for $i \in [k]$.

$$\mathcal{G}_{n,k} := \left\{ \mathcal{G} = \{(\omega_i, \mu^{(i)}, \Sigma^{(i)})\}_{i \in [k]} : \omega_i \in \mathbb{R}_+, \sum_{i=1}^k \omega_i = 1, \mu^{(i)} \in \mathbb{R}^n, \Sigma^{(i)} \in \mathbb{R}_{sym}^{n \times n}, \Sigma^{(i)} \succeq 0 \right\}.$$

As an interesting special case of the general model, we also consider the mixture of “zero-mean” Gaussians, which has $\mu^{(i)} = 0$ for all components $i \in [k]$.

A sample x from a mixture of Gaussians is generated in two steps:

1. Sample $h \in [k]$ from a multinomial distribution, with probability $\Pr[h = i] = \omega_i$ for $i \in [k]$.
2. Sample $x \in \mathbb{R}^n$ from the h -th Gaussian distribution $\mathcal{N}(\mu^{(h)}, \Sigma^{(h)})$.

The learning problem asks to estimate the parameters of the underlying mixture of Gaussians:

Definition 3.1 (Learning mixture of Gaussians). *Given N samples x_1, x_2, \dots, x_N drawn i.i.d. from a mixture of Gaussians $\mathcal{G} = \{(\omega_i, \mu^{(i)}, \Sigma^{(i)})\}_{i \in [k]}$, an algorithm learns the mixture of Gaussians with accuracy ϵ , if it outputs an estimation $\widehat{\mathcal{G}} = \{(\widehat{\omega}_i, \widehat{\mu}^{(i)}, \widehat{\Sigma}^{(i)})\}_{i \in [k]}$ such that there exists a permutation π on $[k]$, and for all $i \in [k]$, we have $|\widehat{\omega}_i - \omega_{\pi(i)}| \leq \epsilon$, $\|\widehat{\mu}^{(i)} - \mu^{(\pi(i))}\| \leq \epsilon$ and $\|\widehat{\Sigma}^{(i)} - \Sigma^{(\pi(i))}\| \leq \epsilon$.*

In the worst case, learning mixture of Gaussians is a information theoretically hard problem ([85]). There exists worst-case examples where the number of samples required for learning the instance is at least exponential in the number of components k ([82]). The non-convexity arises from the hidden variable h : without knowing h we cannot determine which Gaussian component each sample comes from.

The smoothed analysis framework provides a way to circumvent the worst case instances, yet still studying this problem in its most general form. The basic idea

is that, with high probability over the small random perturbation to any instance, the instance will not be a “worst-case” instance, and actually has reasonably good condition for the algorithm.

Next, we show how the parameters of the mixture of Gaussians are *perturbed* in our setup.

Definition 3.2 (ρ -smooth mixture of Gaussian). *For $\rho < 1/n$, a ρ -smooth n -dimensional k -component mixture of Gaussians $\tilde{\mathcal{G}} = \{(\tilde{\omega}_i, \tilde{\mu}^{(i)}, \tilde{\Sigma}^{(i)})\}_{i \in [k]} \in \mathcal{G}_{n,k}$ is generated as follows:*

1. *Choose an arbitrary (could be adversarial) instance $\mathcal{G} = \{(\omega_i, \mu^{(i)}, \Sigma^{(i)})\}_{i \in [k]} \in \mathcal{G}_{n,k}$. Scale the distribution such that $0 \preceq \Sigma^{(i)} \preceq \frac{1}{2}I_n$ and $\|\mu^{(i)}\| \leq \frac{1}{2}$ for all $i \in [k]$.*
2. *Let $\Delta_i \in \mathbb{R}_{sym}^{n \times n}$ be a random symmetric matrix with zeros on the diagonals, and the upper-triangular entries are independent random Gaussian variables $\mathcal{N}(0, \rho^2)$. Let $\delta_i \in \mathbb{R}^n$ be a random Gaussian vector with independent Gaussian variables $\mathcal{N}(0, \rho^2)$.*
3. *Set $\tilde{\omega}_i = \omega_i$, $\tilde{\mu}^{(i)} = \mu^{(i)} + \delta_i$, $\tilde{\Sigma}^{(i)} = \Sigma^{(i)} + \Delta_i$.*
4. *Choose the diagonal entries of $\tilde{\Sigma}^{(i)}$ arbitrarily, while ensuring the positive semi-definiteness of the covariance matrix $\tilde{\Sigma}^{(i)}$, and the diagonal entries are upper bounded by 1. The perturbation procedure fails if this step is infeasible⁴.*

A ρ -smooth zero-mean mixture of Gaussians is generated using the same procedure, except that we set $\tilde{\mu}^{(i)} = \mu^{(i)} = 0$, for all $i \in [k]$.

Remark 3.3. *When the original matrix is of low rank, a simple random perturbation may not lead to a positive semidefinite matrix, which is why our procedure of perturbation is more restricted in order to guarantee that the perturbed matrix is still a valid covariance matrix.*

⁴ Note that by standard random matrix theory, with high probability the 4-th step is feasible and the perturbation procedure in Definition 3.2 succeeds. Also, with high probability we have $\|\tilde{\mu}^{(i)}\| \leq 1$ and $0 \preceq \tilde{\Sigma}^{(i)} \preceq I_n$ for all $i \in [k]$.

There could be other ways of locally perturbing the covariance matrix. Our procedure actually gives more power to the adversary as it can change the diagonals after observing the perturbations for other entries. Note that with high probability if we just let the new diagonal to be $5\sqrt{n}\rho$ larger than the original ones, the resulting matrix is still a valid covariance matrix. In other words, the adversary can always keep the perturbation small if it wants to.

Instead of the worst-case problem in Definition 3.1, our algorithms work on the smoothed instance. Here the model first gets perturbed to $\tilde{\mathcal{G}} = \{(\tilde{\omega}_i, \tilde{\mu}^{(i)}, \tilde{\Sigma}^{(i)})\}_{i \in [k]}$, the samples are drawn according to the perturbed model, and the algorithm tries to learn the perturbed parameters. We give a polynomial time algorithm in this case:

Theorem 3.4 (Main theorem). *Consider a ρ -smooth mixture of Gaussians $\tilde{\mathcal{G}} = \{(\tilde{\omega}_i, \tilde{\mu}^{(i)}, \tilde{\Sigma}^{(i)})\}_{i \in [k]} \in \mathcal{G}_{n,k}$ for which the number of components is at least ⁵ $k \geq C_0$ and the dimension $n \geq C_1 k^2$, for some fixed constants C_0 and C_1 . Suppose that the mixing weights $\tilde{\omega}_i \geq \omega_o$ for all $i \in [k]$. Given N samples drawn i.i.d. from $\tilde{\mathcal{G}}$, there is an algorithm that learns the parameters of $\tilde{\mathcal{G}}$ up to accuracy ϵ , with high probability over the randomness in both the perturbation and the samples. Furthermore, the running time and number of samples N required are both upper bounded by $\text{poly}(n, k, 1/\omega_o, 1/\epsilon, 1/\rho)$.*

To better illustrate the algorithmic ideas for the general case, we first present an algorithm for learning mixtures of zero-mean Gaussians. Note that this is not just a special case of the general case, as with the smoothed analysis, the zero mean vectors are not perturbed.

Theorem 3.5 (Zero-mean). *Consider a ρ -smooth mixture of zero-mean Gaussians $\tilde{\mathcal{G}} = \{(\tilde{\omega}_i, 0, \tilde{\Sigma}^{(i)})\}_{i \in [k]} \in \mathcal{G}_{n,k}$ for which the number of components is at least $k \geq C_0$ and the dimension $n \geq C_1 k^2$, for some fixed constants C_0 and C_1 . Suppose that the mixing weights $\tilde{\omega}_i \geq \omega_o$ for all $i \in [k]$. Given N samples drawn i.i.d. from $\tilde{\mathcal{G}}$, there is an algorithm that learns the parameters of $\tilde{\mathcal{G}}$ up to accuracy ϵ , with*

⁵Note that the algorithms of [23] and [85] run in polynomial time for fixed k .

high probability over the randomness in both the perturbation and the samples. Furthermore, the running time and number of samples N are both upper bounded by $\text{poly}(n, k, 1/\omega_o, 1/\epsilon, 1/\rho)$.

Throughout the discussion we always assume that $n \geq C_1 k^2$ and $\tilde{\omega}_i \geq \omega_o$.

Moment Structure of Mixture of Gaussians Our algorithm is also based on the method of moments, and we only need to estimate the 3-rd, the 4-th and the 6-th order moments. In this part we briefly discuss the structure of 4-th and 6-th moments in the zero-mean case (3-rd moment is always 0 in the zero-mean case). These structures are essential to the proposed algorithm. For more details, and discussions on the general case see Appendix 3.4.6.

The m -th order moments of the *zero-mean* Gaussian mixture model $\mathcal{G} \in \mathcal{G}_{n,k}$ are given by the following m -th order symmetric tensor $M_m \in \mathbb{R}_{sym}^{n \times \dots \times n}$:

$$[M_m]_{j_1, \dots, j_m} := \mathbb{E}[x_{j_1} \dots x_{j_m}] = \sum_{i=1}^k \omega_i \mathbb{E}[y_{j_1}^{(i)} \dots y_{j_m}^{(i)}], \quad \forall j_1, \dots, j_m \in [n],$$

where $y^{(i)}$ corresponds to the n -dimensional zero-mean Gaussian distribution $\mathcal{N}(0, \Sigma^{(i)})$. The moments for each Gaussian component are characterized by Isserlis's theorem as below:

Theorem 3.6 (Isserlis' Theorem). *Let (y_1, \dots, y_{2t}) be a multivariate zero-mean Gaussian random vector $\mathcal{N}(0, \Sigma)$, then*

$$\mathbb{E}[y_1 \dots y_{2t}] = \sum \prod \Sigma_{u,v},$$

where the summation is taken over all distinct ways of partitioning y_1, \dots, y_{2t} into t pairs, which correspond to all the perfect matchings in a complete graph.

Ideally, we would like to obtain the following quantities (recall $n_2 = \binom{n+1}{2}$):

$$X_4 = \sum_{i=1}^k \omega_i \text{vec}(\Sigma^{(i)}) \otimes^2 \in \mathbb{R}^{n_2 \times n_2}, \quad X_6 = \sum_{i=1}^k \omega_i \text{vec}(\Sigma^{(i)}) \otimes^3 \in \mathbb{R}^{n_2 \times n_2 \times n_2}. \quad (3.1)$$

Note that the entries in X_4 and X_6 are quadratic and cubic monomials of the covariance matrices, respectively. If we have X_4 and X_6 , the tensor decomposition algorithm in [7] can be immediately applied to recover ω_i 's and $\Sigma^{(i)}$'s under mild conditions. It is easy to verify that those conditions are indeed satisfied with high probability in the smoothed analysis setting.

By Isserlis's theorem, the entries of the moments M_4 and M_6 are indeed quadratic and cubic functions of the covariance matrices, respectively. However, the structure of the true moments M_4 and M_6 have more symmetries, consider for example,

$$[M_4]_{1,2,3,4} = \sum_{i=1}^k \omega_i (\Sigma_{1,2}^{(i)} \Sigma_{3,4}^{(i)} + \Sigma_{1,3}^{(i)} \Sigma_{2,4}^{(i)} + \Sigma_{1,4}^{(i)} \Sigma_{2,3}^{(i)}), \quad \text{while } [X_4]_{(1,2),(3,4)} = \sum_{i=1}^k \omega_i \Sigma_{1,2}^{(i)} \Sigma_{3,4}^{(i)}.$$

Note that due to symmetry, the number of distinct entries in M_4 ($\binom{n+3}{4} \approx n^4/24$) is much smaller than the number of distinct entries in X_4 ($\binom{n^2+1}{2} \approx n^4/8$). Similar observation can be made about M_6 and X_6 .

Therefore, it is not immediate how to find the desired X_4 and X_6 based on M_4 and M_6 . We call the moments M_4, M_6 the *folded moments* as they have more symmetry, and the corresponding X_4, X_6 the *unfolded moments*. One of the key steps in our algorithm is to unfold the true moments M_4, M_6 to get X_4, X_6 by exploiting special structure of M_4, M_6 .

In some cases, it is easier to restrict our attention to the entries in M_4 with indices corresponding to distinct variables. In particular, we define

$$\overline{M}_4 = [[M_4]_{j_1, j_2, j_3, j_4} : 1 \leq j_1 < j_2 < j_3 < j_4 \leq n] \in \mathbb{R}^{n_4}, \quad (3.2)$$

where $n_4 = \binom{n}{4}$ is the number of 4-tuples with indices corresponding to distinct variables. We define $\overline{M}_6 \in \mathbb{R}^{n_6}$ similarly where $n_6 = \binom{n}{6}$. We will see that these entries are nice as they are *linear projections* of the desired unfolded moments X_4 and X_6 (Lemma 3.1 below), also such projections satisfy certain "symmetric off-diagonal" properties which are convenient for the proof (see Definition 3.3 in Section 3.4.2).

Lemma 3.1. *For a zero-mean Gaussian mixture model, there exist two fixed and*

known linear mappings $\mathcal{F}_4 : \mathbb{R}^{n_2 \times n_2} \rightarrow \mathbb{R}^{n_4}$ and $\mathcal{F}_6 : \mathbb{R}^{n_2 \times n_2 \times n_2} \rightarrow \mathbb{R}^{n_6}$ such that:

$$\overline{M}_4 = \sqrt{3}\mathcal{F}_4(X_4), \quad \overline{M}_6 = \sqrt{15}\mathcal{F}_6(X_6). \quad (3.3)$$

Moreover \mathcal{F}_4 is a projection from a $\binom{n_2+1}{2}$ -dimensional subspace to a n_4 -dimensional subspace, and \mathcal{F}_6 is a projection from a $\binom{n_2+2}{3}$ -dimensional subspace to a n_6 -dimensional subspace.

3.3 Outline of our algorithm

3.3.1 Learning Mixture of Zero-Mean Gaussians

In this section, we present our algorithm for learning zero-mean Gaussian mixture model. The algorithmic ideas and the analysis are at the core of this work. Later we show that it is relatively easy to generalize the basic ideas and the techniques to handle the general case.

For simplicity we state our algorithm using the exact moments \widetilde{M}_4 and \widetilde{M}_6 , while in implementation the empirical moments \widehat{M}_4 and \widehat{M}_6 obtained with the samples are used. In later sections, we verify the correctness of the algorithm and show that it is robust: the algorithm learns the parameters up to arbitrary accuracy using polynomial number of samples.

Step 1. Span Finding: Find the span of covariance matrices .

(a) For a set of indices $\mathcal{H} \subset [n]$ of size $|\mathcal{H}| = \sqrt{n}$, find the span:

$$\mathcal{S} = \text{span} \left\{ \widetilde{\Sigma}_{[i,j]}^{(i)} : i \in [k], j \in \mathcal{H} \right\} \subset \mathbb{R}^n. \quad (3.4)$$

(b) Find the span of the covariance matrices with the columns projected onto \mathcal{S}^\perp , namely,

$$\mathcal{U}_\mathcal{S} = \text{span} \left\{ \text{vec}(\text{Proj}_{\mathcal{S}^\perp} \widetilde{\Sigma}^{(i)}) : i \in [k] \right\} \subset \mathbb{R}^{n^2}. \quad (3.5)$$

(c) For two disjoint sets of indices \mathcal{H}_1 and \mathcal{H}_2 , repeat Step 1 (a) and Step 1 (b) to obtain \mathcal{U}_1 and \mathcal{U}_2 , namely the span of covariance matrices projected onto two subspaces \mathcal{S}_1^\perp and \mathcal{S}_2^\perp . Merge \mathcal{U}_1 and \mathcal{U}_2 to obtain the span of covariance matrices \mathcal{U} :

$$\mathcal{U} = \text{span} \left\{ \tilde{\Sigma}^{(i)} : i \in [k] \right\} \subset \mathbb{R}^{n_2}. \quad (3.6)$$

Step 2. Unfolding: Recover the unfolded moments \tilde{X}_4, \tilde{X}_6 .

Given the folded moments $\overline{M}_4, \overline{M}_6$ as defined in (3.2), and given the subspace $U \in \mathbb{R}^{n_2 \times k}$ from Step 1, let $\tilde{Y}_4 \in \mathbb{R}_{\text{sym}}^{k \times k}$ and $\tilde{Y}_6 \in \mathbb{R}_{\text{sym}}^{k \times k \times k}$ be the unknowns, solve the following systems of linear equations.

$$\overline{M}_4 = \sqrt{3}\mathcal{F}_4(U\tilde{Y}_4U^\top), \quad \overline{M}_6 = \sqrt{15}\mathcal{F}_6(\tilde{Y}_6(U^\top, U^\top, U^\top)). \quad (3.7)$$

The unfolded moments \tilde{X}_4, \tilde{X}_6 are then given by $\tilde{X}_4 = U\tilde{Y}_4U^\top, \tilde{X}_6 = \tilde{Y}_6(U^\top, U^\top, U^\top)$.

Step 3. Tensor Decomposition: learn $\tilde{\omega}_i$ and $\tilde{\Sigma}^{(i)}$ from \tilde{Y}_4 and \tilde{Y}_6 .

Given U , and given \tilde{Y}_4 and \tilde{Y}_6 which are relate to the parameters as follows:

$$\tilde{Y}_4 = \sum_{i=1}^k \tilde{\omega}_i (U^\top \tilde{\Sigma}^{(i)}) \otimes^2, \quad \tilde{Y}_6 = \sum_{i=1}^k \tilde{\omega}_i (U^\top \tilde{\Sigma}^{(i)}) \otimes^3,$$

we apply tensor decomposition techniques to recover $\tilde{\Sigma}^{(i)}$'s and $\tilde{\omega}_i$'s.

3.3.2 Implementing the Steps for Zero-Mean Algorithm

In this part we show how to accomplish each step of the algorithm outlined in Section 3.3.1 and sketch the proof ideas.

For each step, we first explain the detailed algorithm, and list the deterministic conditions on the underlying parameters as well as on the *exact* moments for the step to work correctly. Then we show that these deterministic conditions are satisfied with high probability over the ρ -perturbation of the parameters in the smoothed analysis setting. In order to analyze the sample complexity, we further show that when we

are given the *empirical* moments which are close to the exact moments, the output of the step is also close to that in the exact case.

In particular we show the correctness and the stability of each step in the algorithm with two main lemmas: the first lemma shows that with high probability over the random perturbation of the covariance matrices, the exact moments satisfy the deterministic conditions that ensure the correctness of each step; the second lemma shows that when the algorithm for each step works correctly, it is actually stable even when the moments are estimated from finite samples and have only inverse polynomial accuracy to the exact moments.

The detailed proofs are deferred to Section 3.4.1 to 3.4.3 in the appendix.

Step 1: Span Finding. Given the 4-th order moments \widetilde{M}_4 , Step 1 finds the span of covariance matrices \mathcal{U} as defined in (3.6). Note that by definition of the unfolded moments \widetilde{X}_4 in (3.1), the subspace \mathcal{U} coincides with the column span of the matrix \widetilde{X}_4 .

By Lemma 3.1, we know that the entries in \widetilde{M}_4 are linear mappings of entries in \widetilde{X}_4 . Since the matrix \widetilde{X}_4 is of low rank ($k \ll n_2$), this corresponds to the *matrix sensing* problem first studied in [99]. In general, matrix sensing problems can be hard even when we have many linear observations ([53]). Previous works ([99, 54, 61]) showed that if the linear mapping satisfy *matrix RIP* property, one can uniquely recover \widetilde{X}_4 from \widetilde{M}_4 .

However, properties like RIP do not hold in our setting where the linear mapping is determined by Isserlis' Theorem. We can construct two different mixtures of Gaussians with different unfolded moments \widetilde{X}_4 , but the same folded moment \widetilde{M}_4 (see Section 3.4.6). Therefore the existing matrix recovery algorithm cannot be applied, and we need to develop new tools by exploiting the special moment structure of Gaussian mixtures.

Step 1 (a). Find the Span of a Subset of Columns of the Covariance Matrices. The key observation for this step is that if we hit \widetilde{M}_4 with three basis vectors, we get a vector that lies in the span of the columns of the covariance matrices:

Claim 3.1. For a mixture of zero-mean Gaussians $\mathcal{G} = \{(\omega_i, 0, \Sigma^{(i)})\}_{i \in [k]} \in \mathcal{G}_{n,k}$, the one-dimensional slices of the 4-th order moments M_4 are given by:

$$M_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, \mathbf{e}_{j_3}, I) = \sum_{i=1}^k \omega_i \left(\Sigma_{j_1, j_2}^{(i)} \Sigma_{[:, j_3]}^{(i)} + \Sigma_{j_1, j_3}^{(i)} \Sigma_{[:, j_2]}^{(i)} + \Sigma_{j_2, j_3}^{(i)} \Sigma_{[:, j_1]}^{(i)} \right), \quad \forall j_1, j_2, j_3 \in [n]. \quad (3.8)$$

In particular, if we pick the indices j_1, j_2, j_3 in the index set \mathcal{H} , we know that the vector $M_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, \mathbf{e}_{j_3}, I)$ lies in the desired span $\mathcal{S} = \left\{ \Sigma_{[:, j]}^{(i)} : i \in [k], j \in \mathcal{H} \right\}$.

We shall partition the set \mathcal{H} into three disjoint subsets $\mathcal{H}^{(i)}$ of equal size $\sqrt{n}/3$, and pick $j_i \in \mathcal{H}^{(i)}$ for $i = 1, 2, 3$. In this way, we have $(|\mathcal{H}|/3)^3 = \Omega(n^{1.5})$ such one-dimensional slices of M_4 , which all lie in the desired subspace \mathcal{S} . Moreover, the dimension of the subspace \mathcal{S} is at most $k|\mathcal{H}| \ll n^{1.5}$. Therefore, with the ρ -perturbed parameters $\tilde{\Sigma}^{(i)}$'s, we can expect that with high probability the slices of \tilde{M}_4 span the entire subspace \mathcal{S} .

Condition 3.7 (Deterministic condition for Step 1 (a)). Let $\tilde{Q}_S \in \mathbb{R}^{n \times (|\mathcal{H}|/3)^3}$ be the matrix whose columns are the vectors $\tilde{M}_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, \mathbf{e}_{j_3}, I)$ for $j_i \in \mathcal{H}^{(i)}$. If the matrix \tilde{Q}_S achieves its maximal column rank $k|\mathcal{H}|$, we can find the desired span \mathcal{S} defined in (3.4) by the column span of matrix \tilde{Q}_S .

We first show that this deterministic condition is satisfied with high probability by bounding the $k|\mathcal{H}|$ -th singular value of \tilde{Q}_S with smoothed analysis.

Lemma 3.2 (Correctness). Given the exact 4-th order moments \tilde{M}_4 , for any index set \mathcal{H} of size $|\mathcal{H}| = \sqrt{n}$, With high probability, the $k|\mathcal{H}|$ -th singular value of \tilde{Q}_S is at least $\Omega(\omega_o \rho^2 n)$.

The proof idea involves writing the matrix \tilde{Q}_S as a product of three matrices, and using the results on spectral properties of random matrices [101] to show that with high probability the smallest singular value of each factor is lower bounded.

Since this step only involves the singular value decomposition of the matrix \tilde{Q}_S , we then use the standard matrix perturbation theory to show that this step is stable:

Lemma 3.3 (Stability). *Given the empirical estimator of the 4-th order moments $\widehat{M}_4 = \widetilde{M}_4 + E_4$, suppose that the entries of E_4 have absolute value at most δ . Let the columns of matrix $\widetilde{S} \in \mathbb{R}^{n \times k|\mathcal{H}|}$ be the left singular vector of \widetilde{Q}_S , and let \widehat{S} be the corresponding matrix obtained with \widehat{M}_4 . When δ is inverse polynomially small, the distance between the two projections $\|\text{Proj}_{\widehat{S}} - \text{Proj}_{\widetilde{S}}\|$ is upper bounded by $O\left(n^{1.25}\delta/\sigma_{k|\mathcal{H}|}(\widetilde{Q}_S)\right)$.*

Remark 3.8. *Note that we need the high dimension assumption ($n \gg k$) to guarantee the correctness of this step: in order to span the subspace \mathcal{S} , the number of distinct vectors should be equal or larger than the dimension of the subspace, namely $|\mathcal{H}|^3 \geq k|\mathcal{H}|$; and the subspace should be non-trivial, namely $k|\mathcal{H}| < n$. These two inequalities suggest that we need $n \geq \Omega(k^{1.5})$. However, we used the stronger assumption $n \geq \Omega(k^2)$ to obtain the lower bound of the smallest singular value in the proof.*

Step 1 (b). Find the Span of Projected Covariance Matrices. In this step, we continue to use the structural properties of the 4-th order moments. In particular, we look at the two-dimensional slices of M_4 obtained by hitting it with two basis vectors:

Claim 3.2. *For a mixture of zero-mean Gaussians $\mathcal{G} = \{(\omega_i, 0, \Sigma^{(i)})\}_{i \in [k]} \in \mathcal{G}_{n,k}$, the two-dimensional slices of the 4-th order moments M_4 are given by:*

$$M_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, I, I) = \sum_{i=1}^k \omega_i \left(\Sigma_{j_1, j_2}^{(i)} \Sigma^{(i)} + \Sigma_{[:, j_1]}^{(i)} (\Sigma_{[:, j_2]}^{(i)})^\top + \Sigma_{[:, j_2]}^{(i)} (\Sigma_{[:, j_1]}^{(i)})^\top \right), \quad \forall j_1, j_2 \in [n]. \quad (3.9)$$

Note that if we take the indices j_1 and j_2 in the index set \mathcal{H} , the slice $M_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, I, I)$ is *almost* in the span of the covariance matrices, except $2k$ additive rank-one terms in the form of $\Sigma_{[:, j_1]}^{(i)} (\Sigma_{[:, j_2]}^{(i)})^\top$. These rank-one terms can be eliminated by projecting the slice to the subspace \mathcal{S}^\perp obtained in Step 1 (a), namely,

$$\text{vec}(\text{Proj}_{\mathcal{S}^\perp} M_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, I, I)) = \sum_{i=1}^k \omega_i \Sigma_{j_1, j_2}^{(i)} \text{vec}(\text{Proj}_{\mathcal{S}^\perp} \Sigma^{(i)}), \quad \forall j_1, j_2 \in \mathcal{H},$$

and this projected two-dimensional slice lies in the desired span \mathcal{U}_S as defined in (3.5).

Moreover, there are $\binom{|\mathcal{H}|+1}{2} = \Omega(n)$ such projected two-dimensional slices, while the dimension of the desired span \mathcal{U}_S is at most k .

Condition 3.9 (Deterministic condition for Step 1 (b)). *Let $\tilde{Q}_{U_S} \in \mathbb{R}^{n_2 \times |\mathcal{H}|(|\mathcal{H}|+1)/2}$ be a matrix whose (j_1, j_2) -th column for is equal to the projected two-dimensional slice $\text{vec}(\text{Proj}_{S^\perp} \tilde{M}_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, I, I))$, for $j_1 \leq j_2$ and $j_1, j_2 \in \mathcal{H}$. If the matrix \tilde{Q}_{U_S} achieves its maximal column rank k , the desired span \mathcal{U}_S defined in (3.5) is given by the column span of the matrix \tilde{Q}_{U_S} .*

We show that this deterministic condition is satisfied by bounding the k -th singular value of \tilde{Q}_{U_S} in the smoothed analysis setting:

Lemma 3.4 (Correctness). *Given the exact 4-th order moments \tilde{M}_4 , with high probability, the k -th singular value of \tilde{Q}_{U_S} is at least $\Omega(\omega_o \rho^2 n^{1.5})$.*

Similar to Lemma 3.2, the proof is based on writing the matrix Q_{U_S} as a product of three matrices, then bound their k -th singular values using random matrix theory. The stability analysis also relies on the matrix perturbation theory.

Lemma 3.5 (Stability). *Given the empirical 4-th order moments $\widehat{M}_4 = \tilde{M}_4 + E_4$, assume that the absolute value of entries of E_4 are at most δ_2 . Also, given the output $\text{Proj}_{\widehat{S}^\perp}$ from Step 1 (a), and assume that $\|\text{Proj}_{\widehat{S}^\perp} - \text{Proj}_{\tilde{S}^\perp}\| \leq \delta_1$. When δ_1 and δ_2 are inverse polynomially small, we have $\|\text{Proj}_{\widehat{U}_S} - \text{Proj}_{\tilde{U}_S}\| \leq O\left(n^{2.5}(\delta_2 + 2\delta_1) / \sigma_k(\tilde{Q}_{U_S})\right)$.*

Step 1 (c). Merge $\mathcal{U}_1, \mathcal{U}_2$ to get the span of covariance matrices \mathcal{U} . Note that for a given index set \mathcal{H} , the span \mathcal{U}_S obtained in Step 1 (b) only gives partial information about the span of the covariance matrices. The idea of getting the span of the full covariance matrices is to obtain two sets of such partial information and then merge them.

In order to achieve that, we repeat Step 1 (a) and Step 1 (b) for two *disjoint* sets \mathcal{H}_1 and \mathcal{H}_2 , each of size \sqrt{n} . The two subspace S_1 and S_2 thus correspond to the span of two disjoint sets of covariance matrix columns. Therefore, we can hope that U_1 and U_2 , the span of covariance matrices projected to S_1^\perp and S_2^\perp contain enough information to recover the full span U .

In particular, we prove the following claim:

Condition 3.10 (Deterministic condition for Step 1 (c)). *Let the columns of two (unknown) matrices $V_1 \in \mathbb{R}^{n \times k}$ and $V_2 \in \mathbb{R}^{n \times k}$ form two basis of the same k -dimensional (unknown) subspace $\mathcal{U} \subset \mathbb{R}^n$, and let U denote an arbitrary orthonormal basis of \mathcal{U} . Given two s -dimensional subspaces S_1 and S_2 , denote $S_3 = S_1^\perp \cup S_2^\perp$. Given two projections of \mathcal{U} onto the two subspaces S_1^\top and S_2^\top : $U_1 = \text{Proj}_{S_1^\perp} V_1$ and $U_2 = \text{Proj}_{S_2^\perp} V_2$. If $\sigma_{2s}([S_1, S_2]) > 0$ and $\sigma_k(\text{Proj}_{S_3} U) > 0$, there is an algorithm for finding \mathcal{U} robustly.*

The main idea in the proof is that since s is not too large, the two subspaces S_1^\perp and S_2^\perp have a large intersection. Using this intersection we can “align” the two basis V_1 and V_2 and obtain $V_1^\top V_2$, and then it is easy to merge the two projections of the same matrix (instead of a subspace).

Moreover, we show that when applying this result to the projected span of covariance matrices, we have $s = k|\mathcal{H}| \leq n/3$, and the two deterministic conditions $\sigma_{2s}([S_1, S_2]) > 0$ and $\sigma_k(\text{Proj}_{S_3} V_1) > 0$ are indeed satisfied with high probability over the parameter perturbation. The detailed smoothed analysis (Lemma 3.13 and 3.14) and the stability analysis (Lemma 3.12) are provided in Section 3.4.1 in the appendix.

Step 2. Unfold the moments to get \tilde{X}_4 and \tilde{X}_6 . We show that given the span of covariance matrices \mathcal{U} obtained from Step 1, finding the unfolded moments \tilde{X}_4, \tilde{X}_6 is reduced to solving two systems of linear equations.

Recall that the challenge of recovering \tilde{X}_4 and \tilde{X}_6 is that the two linear mappings \mathcal{F}_4 and \mathcal{F}_6 defined in (3.3) are *not linearly invertible*. The key idea of this step is to make use of the span \mathcal{U} to *reduce the number of variables*. Note that given the basis $U \in \mathbb{R}^{n_2 \times k}$ of the span of the covariance matrices, we can represent each vectorized covariance matrix as $\tilde{\Sigma}^{(i)} = U\tilde{\sigma}^{(i)}$. Now Let $\tilde{Y}_4 \in \mathbb{R}_{sym}^{k \times k}$ and $\tilde{Y}_6 \in \mathbb{R}_{sym}^{k \times k \times k}$ denote the unfolded moments in this new coordinate system:

$$\tilde{Y}_4 := \sum_{i=1}^k \tilde{\omega}_i \tilde{\sigma}^{(i)} \otimes^2, \quad \tilde{Y}_6 := \sum_{i=1}^k \tilde{\omega}_i \tilde{\sigma}^{(i)} \otimes^3.$$

Note that once we know \tilde{Y}_4 and \tilde{Y}_6 , the unfolded moments \tilde{X}_4 and \tilde{X}_6 are given by

$\tilde{X}_4 = U\tilde{Y}_4U^\top$ and $\tilde{X}_6 = \tilde{Y}_6(U^\top, U^\top, U^\top)$. Therefore, after changing the variable, we need to solve the two linear equation systems given in (3.7) with the variables \tilde{Y}_4 and \tilde{Y}_6 .

This change of variable significantly reduces the number of unknown variables. Note that the number of distinct entries in \tilde{Y}_4 and \tilde{Y}_6 are $k_2 = \binom{k+1}{2}$ and $k_3 = \binom{k+2}{3}$, respectively. Since $k_2 \leq n_4$ and $k_3 \leq n_6$, we can expect that the linear mapping from \tilde{Y}_4 to \widetilde{M}_4 and the one from \tilde{Y}_6 to \widetilde{M}_6 are linearly invertible. This argument is formalized below.

Condition 3.11 (Deterministic condition for Step 2). *Rewrite the two systems of linear equations in (3.7) in their canonical form and let $\tilde{H}_4 \in \mathbb{R}^{n_4 \times k_2}$ and $\tilde{H}_6 \in \mathbb{R}^{n_6 \times k_3}$ denote the coefficient matrices. We can obtain the unfolded moments \tilde{X}_4 and \tilde{X}_6 if the coefficient matrices have full column rank.*

We show with smoothed analysis that the smallest singular value of the two coefficient matrices are lower bounded with high probability:

Lemma 3.6 (Correctness). *With high probability over the parameter random perturbation, the k_2 -th singular value of the coefficient matrix \tilde{H}_4 is at least $\Omega(\rho^2 n/k)$, and the k_3 -th singular value of the coefficient matrix \tilde{H}_6 is at least $\Omega(\rho^3(n/k)^{1.5})$.*

To prove this lemma we rewrite the coefficient matrix as product of two matrices and bound their smallest singular values separately. One of the two matrices corresponds to a projection of the Kronecker product $\tilde{\Sigma} \otimes_{kr} \tilde{\Sigma}$. In the smoothed analysis setting, this matrix is not necessarily incoherent. In order to provide a lower bound to its smallest singular value, we further apply a carefully designed projection to it, and then we use the concentration bounds for Gaussian chaoses to show that after the projection its columns are incoherent, finally we apply Gershgorin's Theorem to bound the smallest singular value ⁶.

⁶Note that the idea of unfolding using system of linear equations also appeared in the work of [62]. However, in order to show the system of linear equations in their setup is robust, i.e., the coefficient matrix has full rank, they heavily rely on the *incoherence* assumption, which we do not impose in the smoothed analysis setting.

When implementing this step with the empirical moments, we solve two least squares problems instead of solving the system of linear equations. Again using results in matrix perturbation theory and using the lower bound of the smallest singular values of the two coefficient matrices, we show the stability of the solution to the least squares problems:

Lemma 3.7 (Stability). *Given the empirical moments $\widehat{M}_4 = \widetilde{M}_4 + E_4$, $\widehat{M}_6 = \widetilde{M}_6 + E_6$, and suppose that the absolute value of entries of E_4 and E_6 are at most δ_1 . Let \widehat{U} , the output of Step 1, be the estimation for the span of the covariance matrices, and suppose that $\|\widehat{U} - \widetilde{U}\| \leq \delta_2$. Let \widehat{Y}_4 and \widehat{Y}_6 be the least squares solution respectively. When δ_1 and δ_2 are inverse polynomially small, we have $\|\widetilde{Y}_4 - \widehat{Y}_4\|_F \leq O(\sqrt{n_4}(\delta_1 + \delta_2/\sigma_{\min}(\widetilde{H}_4)))$ and $\|\widetilde{Y}_6 - \widehat{Y}_6\|_F \leq O(\sqrt{n_6}(\delta_1 + \delta_2/\sigma_{\min}(\widetilde{H}_6)))$.*

Step 3. Tensor Decomposition.

Claim 3.3. *Given \widetilde{Y}_4 , \widetilde{Y}_6 and \widetilde{U} , the symmetric tensor decomposition algorithm can correctly and robustly find the mixing weights $\widetilde{\omega}_i$'s and the vectors $\widetilde{\sigma}_i$'s, up to some unknown permutation over $[k]$, with high probability over both the randomized algorithm and the parameter perturbation.*

The algorithm and its analysis mostly follow the algorithm of symmetric tensor decomposition in [7], and the details are provided in Section 3.4.3 in the appendix.

Proof Sketch for the Main Theorem of Zero-mean Case. Theorem 3.5 follows from the previous smoothed analysis and stability analysis lemmas for each step.

First, exploiting the randomness of parameter perturbation, the smoothed analysis lemmas show that the deterministic conditions, which guarantee the correctness of each step, are satisfied with high probability. Then using concentration bounds of Gaussian variables, we show that with high probability over the random samples, the empirical moments \widehat{M}_4 and \widehat{M}_6 are entrywise δ -close to the exact moments \widetilde{M}_4 and \widetilde{M}_6 . In order to achieve ϵ accuracy in the parameter estimation, we choose δ

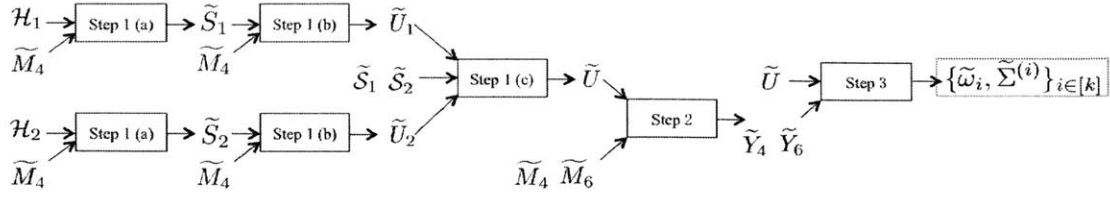


Figure 3-1: Flow of the algorithm for learning mixture of zero-mean Gaussians.

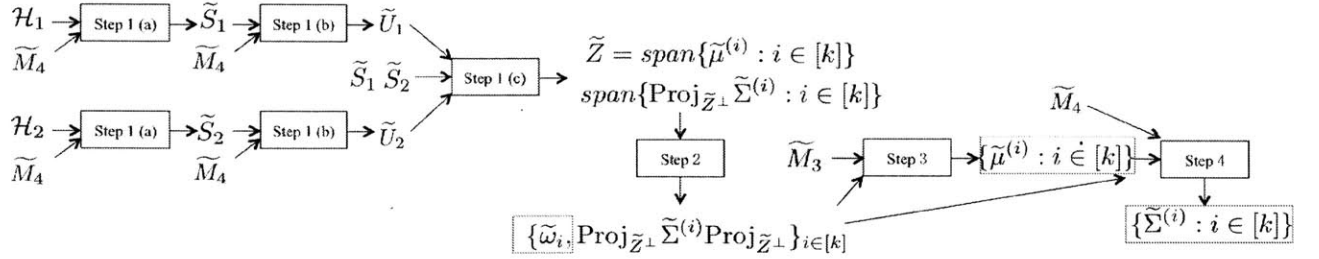


Figure 3-2: Flow of the algorithm for learning mixtures of general Gaussians.

to be inverse polynomially small, and therefore the number of samples required will be polynomial in the relevant parameters. The stability lemmas show how the errors propagate only “polynomially” through the steps of the algorithm, which is visualized in Figure 3-1.

A more detailed illustration is provided in Section 3.4.4 in the appendix.

3.3.3 Learning Mixture of General Gaussians

In this section, we briefly discuss the algorithm for learning mixture of *general* Gaussians. Figure 3-2 shows the inputs and outputs of each step in this algorithm. Many steps share similar ideas to those of the algorithm for the zero-mean case in previous sections. We only highlight the basic ideas and defer the details to Section 3.4.5 in the appendix.

Step 1. Find $\tilde{Z} = \text{span}\{\tilde{\mu}^{(i)} : i \in [k]\}$ and $\tilde{\Sigma}_o = \text{span}\{\mathbf{Proj}_{\tilde{Z}^\perp} \tilde{\Sigma}^{(i)} \mathbf{Proj}_{\tilde{Z}^\perp} : i \in [k]\}$.

Similar to Step 1 in the zero-mean case, this step makes use of the structure of the 4-th order moments \tilde{M}_4 , and is achieved in three small steps:

(a) For a subset $\mathcal{H} \subset [n]$ of size $|\mathcal{H}| = \sqrt{n}$, find the span:

$$\mathcal{S} = \text{span} \left\{ \tilde{\mu}^{(i)}, \tilde{\Sigma}_{[i,j]}^{(i)} : i \in [k], j \in \mathcal{H} \right\} \subset \mathbb{R}^n. \quad (3.10)$$

(b) Find the span of the covariance matrices with the columns projected onto \mathcal{S}^\perp , namely,

$$\mathcal{U}_\mathcal{S} = \text{span} \left\{ \text{vec}(\text{Proj}_{\mathcal{S}^\perp} \tilde{\Sigma}^{(i)}) : i \in [k] \right\} \subset \mathbb{R}^{n^2}. \quad (3.11)$$

(c) For disjoint subsets \mathcal{H}_1 and \mathcal{H}_2 , repeat Step 1 (a) and Step 1 (b) to obtain \mathcal{U}_1 and \mathcal{U}_2 , the span of the covariance matrices projected onto the subspaces \mathcal{S}_1^\perp and \mathcal{S}_2^\perp . The intersection of the two subspaces \mathcal{U}_1 and \mathcal{U}_2 gives the span of the mean vectors $\tilde{Z} = \text{span} \{ \tilde{\mu}^{(i)}, i \in [k] \}$. Merge the two subspaces \mathcal{U}_1 and \mathcal{U}_2 to obtain the span of the covariance matrices projected to the subspace orthogonal to \tilde{Z} , namely $\tilde{\Sigma}_o = \text{span} \left\{ \text{Proj}_{\tilde{Z}^\perp} \tilde{\Sigma}^{(i)} \text{Proj}_{\tilde{Z}^\perp} : i \in [k] \right\}$.

Step 2. Find the Covariance Matrices in the Subspace \tilde{Z}^\perp and the Mixing Weights $\tilde{\omega}_i$'s. The key observation of this step is that when the samples are projected to the subspace orthogonal to all the mean vectors, they are equivalent to samples from a mixture of zero-mean Gaussians with covariance matrices $\tilde{\Sigma}_o^{(i)} = \text{Proj}_{\tilde{Z}^\perp} \tilde{\Sigma}^{(i)} \text{Proj}_{\tilde{Z}^\perp}$ and with the same mixing weights $\tilde{\omega}_i$'s. Therefore, projecting the samples to \tilde{Z}^\perp , the subspace orthogonal to the mean vectors, and use the algorithm for the zero-mean case, we can obtain $\tilde{\Sigma}_o^{(i)}$'s, the covariance matrices projected to this subspace, as well as the mixing weights $\tilde{\omega}_i$'s.

Step 3. Find the means With simple algebra, this step extracts the projected covariance matrices $\tilde{\Sigma}_o^{(i)}$'s from the 3-rd order moments \tilde{M}_3 , the mixing weights $\tilde{\omega}_i$ and the projected covariance matrices $\tilde{\Sigma}_o^{(i)}$'s obtained in Step 2.

Step 4. Find the full covariance matrices In Step 2, we obtained $\tilde{\Sigma}_o^{(i)}$, the covariance matrices projected to the subspace orthogonal to all the means. Note

that they are equal to matrices $(\tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)}(\tilde{\mu}^{(i)})^\top)$ projected to the same subspace. We claim that if we can find the span of these matrices $((\tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)}(\tilde{\mu}^{(i)})^\top)$'s), we can get each matrix $(\tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)}(\tilde{\mu}^{(i)})^\top)$, and then subtracting the known rank-one component to find the covariance matrix $\tilde{\Sigma}^{(i)}$. This is similar to the idea of merging two projections of the same subspace in Step 1 (c) for the zero-mean case.

The idea of finding the desired span is to construct a 4-th order tensor:

$$\tilde{M}'_4 = \tilde{M}_4 + 2 \sum_{i=1}^k \tilde{\omega}_i (\tilde{\mu}^{(i)} \otimes^4),$$

which corresponds to the 4-th order moments of a mixture of zero-mean Gaussians with covariance matrices $\tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)}(\tilde{\mu}^{(i)})^\top$ and the same mixing weights $\tilde{\omega}_i$'s. Then we can then use Step 1 of the algorithm for the zero-mean case to obtain the span of the new covariance matrices, i.e. $\text{span}\{\tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)}(\tilde{\mu}^{(i)})^\top : i \in [k]\}$.

3.4 Proofs for Chapter 3

3.4.1 Step 1 of Zero-Mean Case: Span Finding

Recall that in Step 1 of the algorithm for learning mixture of zero-mean Gaussians, we find the span of the covariance matrices in three small steps. In this section, we prove the correctness and the robustness of each step with smoothed analysis.

For completeness we restate each substep and highlight the key properties we need, followed by the detailed proofs.

Step 1(a). Finding \mathcal{S} , the span of a subset of columns of $\tilde{\Sigma}^{(i)}$'s.

In Step 1 (a), for any set \mathcal{H} of size \sqrt{n} , we want to show that the one-dimensional slices of M_4 span the entire subspace $\mathcal{S} = \text{span}\left\{\tilde{\Sigma}_{[i,j]}^{(i)} : i \in [k], j \in \mathcal{H}\right\}$, which is the span of a subset of the columns in the covariance matrices.

Algorithm 5: FindColumnSpan

Input: 4-th order moments M_4 , set of indices \mathcal{H} .

Output: $\text{span}\{\Sigma_j^{(i)} : i \in [k], j \in \mathcal{H}\}$, represented by an orthonormal matrix $S \in \mathbb{R}^{n \times k|\mathcal{H}|}$.

Let Q be a matrix of dimension $n \times |\mathcal{H}|^3$ whose columns are all of $M_4(e_{i_1}, e_{i_2}, e_{i_3}, I)$, for $i_1, i_2, i_3 \in \mathcal{H}$.

Compute the SVD of Q : $Q = UDV^\top$.

Return: The first $k|\mathcal{H}|$ left singular vectors $S = [U_{[:,1]}, \dots, U_{[:,k|\mathcal{H}|]}]$.

Recall that in Claim 3.1 we showed:

$$\widetilde{M}_4(e_{j_1}, e_{j_2}, e_{j_3}, I) = \sum_{i=1}^k \widetilde{\omega}_i \left(\widetilde{\Sigma}_{j_1, j_2}^{(i)} \widetilde{\Sigma}_{[:, j_3]}^{(i)} + \widetilde{\Sigma}_{j_1, j_3}^{(i)} \widetilde{\Sigma}_{[:, j_2]}^{(i)} + \widetilde{\Sigma}_{j_2, j_3}^{(i)} \widetilde{\Sigma}_{[:, j_1]}^{(i)} \right), \quad \forall j_1, j_2, j_3 \in [n].$$

This in particular means when $j_1, j_2, j_3 \in \mathcal{H}$, the vector $\widetilde{M}_4(e_{j_1}, e_{j_2}, e_{j_3}, I)$ is in \mathcal{S} .

We need to show that the columns of the matrix Q indeed span the entire subspace \mathcal{S} .

It is sufficient to show that a subset of the column span the entire subspace. Form a three-way even partition of the set \mathcal{H} , i.e., $|\mathcal{H}^{(1)}| = |\mathcal{H}^{(2)}| = |\mathcal{H}^{(3)}| = |\mathcal{H}|/3 = \sqrt{n}/3$, and only consider the one-dimensional slices of \widetilde{M}_4 corresponding to the indices $j_i \in \mathcal{H}^{(i)}$ for $i = 1, 2, 3$. In particular, we define matrix \widetilde{Q}_S with these one-dimensional slices of \widetilde{M}_4 :

$$\widetilde{Q}_S = \left[\left[\widetilde{M}_4(e_{j_1}, e_{j_2}, e_{j_3}, I) : j_3 \in \mathcal{H}^{(3)} \right] : j_2 \in \mathcal{H}^{(2)} \right] : j_1 \in \mathcal{H}^{(1)} \in \mathbb{R}^{n \times (|\mathcal{H}|/3)^3}. \quad (3.12)$$

Define matrix \widetilde{P}_S with the corresponding columns of the covariance matrices, forming a basis (although not orthogonal) of the desired subspace \mathcal{S} :

$$\widetilde{P}_S = \left[\left[\left[\widetilde{\Sigma}_{[:, j]}^{(i)} : i \in [k] \right] : j \in \mathcal{H}^{(l)} \right] : l = 1, 2, 3 \right] = \left[\widetilde{\Sigma}_{[:, \mathcal{H}^{(1)}}], \widetilde{\Sigma}_{[:, \mathcal{H}^{(2)}}], \widetilde{\Sigma}_{[:, \mathcal{H}^{(3)}}] \right] \in \mathbb{R}^{n \times k|\mathcal{H}|}. \quad (3.13)$$

In the following two lemmas, we show that with high probability over the random perturbation, the column span of \widetilde{Q}_S is exactly equal to the column span of \widetilde{P}_S , and

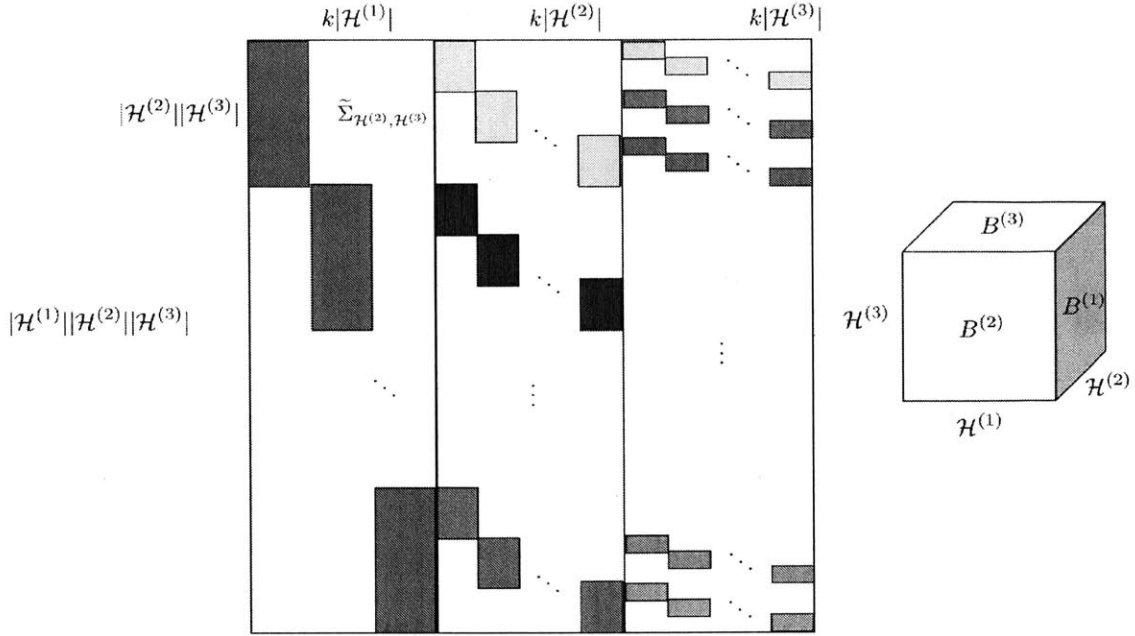


Figure 3-3: Structure of the matrix B_S

robustly so.

Lemma 3.8 (Lemma 3.2 restated). *Given \widetilde{M}_4 , the exact 4-th order moment of the ρ -smooth mixture of zero-mean Gaussians, for any subset $\mathcal{H} \in [n]$ with cardinality $|\mathcal{H}| = \sqrt{n}$, let \widetilde{Q}_S be the matrix defined as in (3.12) with the one-dimensional slices of \widetilde{M}_4 . For any $\epsilon > 0$, and for some absolute constant $C_1, C_2, C_3 > 0$, with probability at least $1 - (C_1\epsilon)^{C_2n}$, the $k|\mathcal{H}|$ -th singular value of \widetilde{Q}_S is bounded below by:*

$$\sigma_{k|\mathcal{H}|}(\widetilde{Q}_S) \geq C_3\omega_o\epsilon^2\rho^2n. \quad (3.14)$$

In order to prove this lemma, we first write \widetilde{Q}_S as the product of three matrices.

Claim 3.4 (Structural). *Under the same assumptions of Lemma 3.8, the matrix \widetilde{Q}_S can be written as*

$$\widetilde{Q}_S = \widetilde{P}_S (D_{\bar{\omega}} \otimes_{kr} I_{|\mathcal{H}|}) (\widetilde{B}_S)^\top, \quad (3.15)$$

where $\widetilde{P}_S \in \mathbb{R}^{n \times k|\mathcal{H}|}$ as defined in Equation (3.13 has columns equal to the columns

in $\tilde{\Sigma}_{[\cdot, \mathcal{H}]}^{(i)}$; the diagonal matrix in the middle is the Kronecker product of two diagonal matrices and depends only on the mixing weights $\tilde{\omega}_i$'s.

With the observation that the columns of \tilde{P}_S form a basis of the subspace \mathcal{S} , and each column of \tilde{Q}_S is a linear combination of the columns in \tilde{P}_S , the rows of $\tilde{B}_S \in \mathbb{R}^{(|\mathcal{H}|/3)^3 \times k|\mathcal{H}|}$ can be viewed as the coefficients for the linear combinations, and has some special structures. In particular, it consists of three blocks: $\tilde{B}_S = [\tilde{B}^{(1)}, \tilde{B}^{(2)}, \tilde{B}^{(3)}]$. The first tall matrix $\tilde{B}^{(1)} \in \mathbb{R}^{(|\mathcal{H}|/3)^3 \times k(|\mathcal{H}|/3)}$, corresponding to the coefficient of the linear combinations on the subset of basis $\tilde{\Sigma}_{[\cdot, \mathcal{H}^{(1)}]}$. By the indexing order of the columns in \tilde{Q}_S , the matrix $\tilde{B}^{(1)}$ is block diagonal with identical blocks equal to $\tilde{\Sigma}_{\mathcal{H}^{(2)}, \mathcal{H}^{(3)}}$, defined as follows:

$$\tilde{\Sigma}_{\mathcal{H}^{(2)}, \mathcal{H}^{(3)}} = \left[[\tilde{\Sigma}_{j_1, j_2}^{(i)} : j_1 \in \mathcal{H}^{(2)}, j_2 \in \mathcal{H}^{(3)}]^\top : i \in [k] \right] \in \mathbb{R}^{(|\mathcal{H}|/3)^2 \times k}.$$

With some fixed and known row permutation $\pi^{(2)}$ and $\pi^{(3)}$, the matrix $\tilde{B}^{(2)}$ and $\tilde{B}^{(3)}$ can be made block diagonal with identical blocks equal to $\tilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}$ and $\tilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(2)}}$, respectively. Note that the three parts $\tilde{B}^{(1)}, \tilde{B}^{(2)}, \tilde{B}^{(3)}$ do not have any common entry, nor do they involve any diagonal entry of the covariance matrices, therefore the three parts are independent when the covariances are randomly perturbed in the smoothed analysis.

It is easier to understand the structure by picture, see Figure 3-3. The rows of the matrix should be indexed by $(j_1, j_2, j_3) \in \mathcal{H}^{(1)} \times \mathcal{H}^{(2)} \times \mathcal{H}^{(3)}$, which can also be interpreted as a cube (in the right). The block structure in the first part $\tilde{B}^{(1)}$ correspond to a slice in $\mathcal{H}^{(2)} \times \mathcal{H}^{(3)}$ direction (for each block, the element in $\mathcal{H}^{(1)}$ is fixed, the elements in $\mathcal{H}^{(2)}$ and $\mathcal{H}^{(3)}$ take all possible values). Similarly for $\tilde{B}^{(2)}$ and $\tilde{B}^{(3)}$ (as shown in figure).

Proof. (of Claim 3.4) The proof of this claim is using Claim 3.1, the definition of matrices and the rule of matrix multiplication. Consider the column in \tilde{Q}_S corresponding to the index (j_1, j_2, j_3) for $j_1 \in \mathcal{H}^{(1)}, j_2 \in \mathcal{H}^{(2)}, j_3 \in \mathcal{H}^{(3)}$, and the row of \tilde{B}_S together with the mixing weights specifies how this column is formed as a linear combination of $3k$ columns of \tilde{P}_S . By the structure of M_4 in Claim 3.1, the (j_1, j_2, j_3) -th row of $\tilde{B}^{(1)}$

has exactly k entries corresponding to $\tilde{\Sigma}_{j_2, j_3}^{(i)}$ for $i \in [k]$, these entries are multiplied by $\tilde{\omega}_i$ in the middle (diagonal) matrix. Therefore, these directly correspond to the k terms in Claim 3.1. Similarly the entries in $\tilde{B}^{(2)}$ and $\tilde{B}^{(3)}$ correspond to the other $2k$ terms. \square

Using Claim 3.4, we need to bound the smallest singular value for each of the matrices in order to bound the $k|\mathcal{H}|$ -th singular value of \tilde{Q}_S , this is deferred to the end of this part. The most important tool is a corollary (Lemma 3.32) of the random matrix result proved in [101], which gives a lowerbound on the smallest singular value of perturbed rectangular matrices.

By Lemma 3.8, we know \tilde{Q}_S has exactly rank $k|\mathcal{H}|$, and is robust in the sense that its $k|\mathcal{H}|$ -th singular value is large (polynomial in the amount of perturbation ρ). By standard matrix perturbation theory, if we get \hat{Q}_S close to \tilde{Q}_S up to a high accuracy (inverse polynomial in the relevant parameters), the top $k|\mathcal{H}|$ singular vectors will span a subspace that is very close to the span of \tilde{Q}_S . We formalize this in the following lemma.

Lemma 3.9 (Lemma 3.3 restated). *Given the empirical estimator of the 4-th order moments $\widehat{M}_4 = \widetilde{M}_4 + E_4$. and suppose that the absolute value of entries of E_4 are at most δ . Let the columns of matrix $\tilde{S} \in \mathbb{R}^{n \times k|\mathcal{H}|}$ be the left singular vector of \tilde{Q}_S , and let \hat{S} be the corresponding matrix obtained with \widehat{M}_4 . Conditioned on the high probability event $\sigma_{k|\mathcal{H}|}(\tilde{Q}_S) > 0$, for some absolute constant C we have:*

$$\|Proj_{\hat{S}} - Proj_{\tilde{S}}\| \leq \frac{Cn^{1.25}}{\sigma_{k|\mathcal{H}|}(\tilde{Q}_S)} \delta. \quad (3.16)$$

Proof. Note that the columns of S are the leading left singular vectors of Q_S . We apply the standard matrix perturbation bound of singular vectors. Recall that S is defined to be the first $k|\mathcal{H}|$ left singular vector of Q_S , and we have

$$\|\hat{Q}_S - \tilde{Q}_S\| \leq \|\hat{Q}_S - Q_S\|_F \leq \sqrt{n(|\mathcal{H}|/3)^3 \delta^2}.$$

Therefore by Wedin's Theorem (in particular the corollary Lemma 1.5), we can con-

clude (3.16). □

Next, we prove Lemma 3.8.

Proof of Lemma 3.8 We first use Claim 3.4 to write $\tilde{Q}_S = \tilde{P}_S (D_{\tilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|}) (\tilde{B}_S)^\top$, note that the matrix $(D_{\tilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|})$ has dimension $k|\mathcal{H}| \times k|\mathcal{H}|$, therefore we just need to show with high probability each of the three factor matrix has large $k|\mathcal{H}|$ -th singular value, and that implies a bound on the $k|\mathcal{H}|$ -th singular value of \tilde{Q}_S by union bound. The smallest singular value of \tilde{P}_S and \tilde{B}_S are bounded below by the following two Claims.

Claim 3.5. *With high probability $\sigma_{k|\mathcal{H}|}(\tilde{P}_S) \geq \Omega(\rho\sqrt{n})$.*

Proof. This claim is easy as $\tilde{P}_S \in \mathbb{R}^{n \times k|\mathcal{H}|}$ is a tall matrix with $n \geq 5k|\mathcal{H}|$ rows. In particular, let \tilde{P}'_S be the block of \tilde{P}_S with rows restricted to $\mathcal{H}^C = [n] \setminus \mathcal{H}$. Note that \tilde{P}'_S is a linear projection of \mathcal{P}_S , and by basic property of singular values in Lemma 3.28, the $k|\mathcal{H}|$ singular values of \tilde{P}'_S provide lower bounds for the corresponding ones of \tilde{P}_S . We only consider the restricted rows so that \tilde{P}'_S does not involve any diagonal elements of the covariance matrices, which are not randomly perturbed in our smoothed analysis framework.

Now \tilde{P}'_S is a randomly perturbed rectangular matrix, whose smallest singular value can be lower bounded using Lemma 3.32, and we conclude that with probability at least $1 - (C\epsilon)^{0.25n}$,

$$\sigma_{k|\mathcal{H}|}(\tilde{P}_S) \geq \epsilon\rho\sqrt{n}.$$

□

Next, we bound the smallest singular value of \tilde{B}_S .

Claim 3.6. *With high probability $\sigma_{k|\mathcal{H}|}(\tilde{B}_S) \geq \Omega(\rho\sqrt{n})$.*

Proof. We make use of the special structure of the three blocks of \tilde{B}_S to lower bound its smallest singular value.

First, we prove that the block diagonal matrix $\tilde{B}^{(1)}$ has large singular values, even after projecting to the orthogonal subspace of the column span of $\tilde{B}^{(2)}$ and $\tilde{B}^{(3)}$. This idea appeared several times in our proof and is abstracted in Lemma 3.29. Apply the lemma and we have:

$$\begin{aligned}
& \sigma_{k|\mathcal{H}|}(\tilde{B}_S) \\
& \geq \min \left\{ \sigma_{k(2|\mathcal{H}|/3)}([\tilde{B}^{(2)}, \tilde{B}^{(3)}]), \sigma_k(\text{Proj}_{([\tilde{B}^{(2)}, \tilde{B}^{(3)}]_{\{j\} \times \mathcal{H}^{(2)} \times \mathcal{H}^{(3)}})^\perp} \tilde{\Sigma}_{\mathcal{H}^{(2)}, \mathcal{H}^{(3)}}) : j \in \mathcal{H}^{(1)} \right\} \\
& \geq \min \left\{ \sigma_{k(2|\mathcal{H}|/3)}([\tilde{B}^{(2)}, \tilde{B}^{(3)}]), \right. \\
& \quad \left. \sigma_k(\text{Proj}_{([\tilde{B}^{(2)}, \tilde{B}^{(3)}]_{\{j\} \times \mathcal{H}^{(2)} \times \mathcal{H}^{(3)}})^\perp} \text{Proj}_{\Sigma_{\mathcal{H}^{(2)}, \mathcal{H}^{(3)}}}^\perp \tilde{\Sigma}_{\mathcal{H}^{(2)}, \mathcal{H}^{(3)}}) : j \in \mathcal{H}^{(1)} \right\}, \quad (3.17)
\end{aligned}$$

where the j -th block of $[\tilde{B}^{(2)}, \tilde{B}^{(3)}]$ has dimension $(|\mathcal{H}|/3)^2 \times 2k|\mathcal{H}|/3$. Since

$$(|\mathcal{H}|/3)^2 - k - 2k|\mathcal{H}|/3 = \Omega(n/9 - k - 2kn^{0.5}/3) \geq \Omega(n),$$

this means for each block, even after projection it has more than $3k$ rows. Note that by definition the three blocks $\tilde{B}^{(1)}$, $\tilde{B}^{(2)}$ and $\tilde{B}^{(3)}$ are independent and do not involve any diagonal elements of the covariance matrices, so each block after the two projections is again a rectangular random matrix. We can apply Lemma 3.31, for any j , for some absolute constant C_1, C_2, C_3 (not fixed throughout the discussion), with probability at least $1 - (C_1\epsilon)^{C_2n}$ over the randomness of $\tilde{\Sigma}_{\mathcal{H}^{(2)}, \mathcal{H}^{(3)}}$, we have:

$$\sigma_k(\text{Proj}_{([\tilde{B}^{(2)}, \tilde{B}^{(3)}]_{\{j\} \times \mathcal{H}^{(2)} \times \mathcal{H}^{(3)}})^\perp} \text{Proj}_{\Sigma_{\mathcal{H}^{(2)}, \mathcal{H}^{(3)}}}^\perp \tilde{\Sigma}_{\mathcal{H}^{(2)}, \mathcal{H}^{(3)}}) \geq \epsilon\rho\sqrt{C_3n}. \quad (3.18)$$

Now we can take a union bound over the blocks and conclude that with high probability, the smallest singular value of each block is large.

In order to bound $\sigma_{k(2|\mathcal{H}|/3)}([\tilde{B}^{(2)}, \tilde{B}^{(3)}])$, we use the same strategy. Note that $\tilde{B}^{(2)}$ also has a block structure that corresponds to the $\mathcal{H}^{(1)} \times \mathcal{H}^{(3)}$ faces (see Figure 3-3). Again check the condition on dimension $(|\mathcal{H}|/3)^2 - k - k|\mathcal{H}|/3 \geq \Omega(n) > 3k$, we can apply Lemma 3.29 again to show that for any j , with probability at least $1 - (C_1\epsilon)^{C_2n}$

over the randomness of $\tilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}$, we have:

$$\begin{aligned} \sigma_{k(2|\mathcal{H}|/3)}([\tilde{B}^{(2)}, \tilde{B}^{(3)}]) &\geq \min \left\{ \sigma_{k(|\mathcal{H}|/3)}(\tilde{B}^{(3)}), \right. \\ &\quad \left. \sigma_k(\text{Proj}_{([\tilde{B}^{(3)}]_{\mathcal{H}^{(1)} \times \{j\} \times \mathcal{H}^{(3)}})^{\perp}} \text{Proj}_{\Sigma_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}^{\perp}}^{\perp} \tilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}) : j \in \mathcal{H}^{(2)} \right\}. \end{aligned} \quad (3.19)$$

Again by Lemma 3.31, for any j , with probability at least $1 - (C_1\epsilon)^{C_2n}$ over the randomness of $\tilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}$, we have:

$$\sigma_k(\text{Proj}_{([\tilde{B}^{(3)}]_{\mathcal{H}^{(1)} \times \{j\} \times \mathcal{H}^{(3)}})^{\perp}} \text{Proj}_{\Sigma_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}^{\perp}}^{\perp} \tilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}) \geq \epsilon\rho\sqrt{C_3n}. \quad (3.20)$$

Finally, for $\tilde{B}^{(3)}$ it is a block diagonal structure with blocks correspond to $\mathcal{H}^{(1)} \times \mathcal{H}^{(2)}$ faces (see Figure 3-3). Each block is a perturbed rectangular matrix, therefore we apply Lemma 3.31 to have that with high probability over the randomness of $\tilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(2)}}$,

$$\sigma_{k(|\mathcal{H}|/3)}(\tilde{B}^{(3)}) \geq \sigma_k(\tilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(2)}}) \geq \epsilon\rho\sqrt{n}. \quad (3.21)$$

Now plug in the lower bounds in (3.18) (3.20) (3.21) into the inequalities in (3.17) and (3.19). By union bound we conclude that with high probability:

$$\sigma_{k|\mathcal{H}|}(\tilde{B}_S) \geq \epsilon\rho\sqrt{C_3n}.$$

□

Finally, the diagonal matrix in the middle is given by the Kronecker product of $I_{|\mathcal{H}|}$ and $D_{\tilde{\omega}}$. Recall that $D_{\tilde{\omega}}$ is the diagonal matrix with the mixing weights $\tilde{\omega}_i$'s on its diagonal. By property of Kronecker product and the assumption on the mixing weights, the smallest diagonal element of $D_{\tilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|}$ is at least ω_0 . Therefore $\sigma_{k|\mathcal{H}|}(D_{\tilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|}) \geq \omega_0$.

We have shown that the smallest singular value of all the three factor matrices are large with high probability. Therefore, apply union bound, we conclude that with

probability at least $1 - \exp(-\Omega(n))$,

$$\sigma_{k|\mathcal{H}}(\tilde{Q}_S) \geq \sigma_{k|\mathcal{H}}(\tilde{P}_S) \sigma_{k|\mathcal{H}}(D_{\tilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|}) \sigma_{k|\mathcal{H}}(\tilde{B}_S) \geq O(\omega_o \rho^2 n).$$

Step 1 (b). Finding \mathcal{U}_S , the span of $\tilde{\Sigma}^{(i)}$'s with columns projected to \mathcal{S}^\perp .

Algorithm 6: FindProjectedSigmaSpan

Input: 4-th order moments M_4 , set of indices \mathcal{H} , subspace $\mathcal{S} \subset \mathbb{R}^n$

Output: $\text{span}\{\text{vec}(\text{Proj}_{\mathcal{S}^\perp} \Sigma^{(i)}) : i \in [k]\}$, represented by an orthonormal matrix $U_S \in \mathbb{R}^{n^2 \times k}$.

Let Q be a matrix whose columns are $\text{vec}(\text{Proj}_{\mathcal{S}^\perp} M_4(e_i, e_j, I, I))$ for all $i, j \in \mathcal{H}, i \neq j$.

Compute the SVD of Q : $Q = UDV^\top$.

Return: The first k left singular vectors $U_S = [U_{[:,1]}, \dots, U_{[:,k]}]$.

In Step 1 (b), given the subset of indices \mathcal{H} and the subspace \mathcal{S} obtained in Step 1 (a), we want to show that the projected two-dimensional slices of \tilde{M}_4 span the subspace \mathcal{U}_S defined in (3.5), which is the span of the covariance matrices with the columns projected the subspace \mathcal{S}^\perp :

$$\mathcal{U}_S = \text{span} \left\{ \text{vec}(\text{Proj}_{\mathcal{S}^\perp} \tilde{\Sigma}^{(i)}) : i \in [k] \right\} \subset \mathbb{R}^{n^2}.$$

Recall that in Claim 3.2, we characterized the two dimensional slices of the 4-th moments M_4 of mixture of zero-mean Gaussians as below:

$$\tilde{M}_4(e_{j_1}, e_{j_2}, I, I) = \sum_{i=1}^k \tilde{\omega}_i \left(\tilde{\Sigma}_{j_1, j_2}^{(i)} \tilde{\Sigma}^{(i)} + \tilde{\Sigma}_{[:, j_1]}^{(i)} (\tilde{\Sigma}_{[:, j_2]}^{(i)})^\top + \tilde{\Sigma}_{[:, j_2]}^{(i)} (\tilde{\Sigma}_{[:, j_1]}^{(i)})^\top \right), \quad \forall j_1, j_2 \in [n]. \quad (3.22)$$

For notational convenience, we let \mathcal{J} denote the set $\mathcal{J} = \{(j_1, j_2) : j_1 \leq j_2, j_1, j_2 \in \mathcal{H}\}$, and note that the cardinality is $|\mathcal{J}| = \binom{|\mathcal{H}|+1}{2} = (n + \sqrt{n})/2$. First, we define the matrix $\tilde{Q}_{U_S} \in \mathbb{R}^{n^2 \times |\mathcal{J}|}$ whose columns are the vectorized two-dimensional slices of \tilde{M}_4

with the columns projected to the subspace \mathcal{S}^\perp :

$$\tilde{Q}_{U_S} = \left[\text{vec}(\text{Proj}_{\mathcal{S}^\perp} \tilde{M}_4(e_{j_1}, e_{j_2}, I, I)) : (j_1, j_2) \in \mathcal{J} \right]. \quad (3.23)$$

Similarly we define $\tilde{Q}_{U_0} \in \mathbb{R}^{n^2 \times |\mathcal{J}|}$ with the slices without the projection:

$$\tilde{Q}_{U_0} = \left[\text{vec}(\tilde{M}_4(e_{j_1}, e_{j_2}, I, I)) : (j_1, j_2) \in \mathcal{J} \right].$$

Observe the structure in (3.22) and we see the columns of \tilde{Q}_{U_0} is “almost” in the span of covariance matrices, except for some additive rank one terms. Note that all the rank one terms lie in the subspace \mathcal{S} obtained from Step 1 (a), and they vanish if we project the slice to the orthogonal subspace \mathcal{S}^\perp . In particular, $\text{Proj}_{\mathcal{S}^\perp} \tilde{\Sigma}_{[i,j]}^{(i)} = 0$ for all $j \in S$. Let the columns of the matrix $\tilde{P}_{U_S} \in \mathbb{R}^{n^2 \times k}$ be the vectorized and projected covariance matrices as below:

$$\tilde{P}_{U_S} = \left[\text{vec}(\text{Proj}_{\mathcal{S}^\perp} \tilde{\Sigma}^{(i)}) : i \in [k] \right]. \quad (3.24)$$

In the following claim, we show that the columns of \tilde{Q}_{U_S} indeed lie in the column span of \tilde{P}_{U_S} :

Claim 3.7. *Given S obtained in Step 1(a), the span of $\tilde{\Sigma}_{[i,j]}^{(i)}$ for $j \in \mathcal{H}$ and for all i , then for $j_1, j_2 \in \mathcal{H}$, we have:*

$$\text{Proj}_{\mathcal{S}^\perp} \tilde{M}_4(e_{j_1}, e_{j_2}, I, I) = \sum_{i=1}^k \tilde{\omega}_i \tilde{\Sigma}_{j_1, j_2}^{(i)} \text{Proj}_{\mathcal{S}^\perp} \tilde{\Sigma}^{(i)}, \quad \forall j_1, j_2 \in [n].$$

Similar as in Step 1(a), in the next lemma we show that the columns of \tilde{Q}_{U_S} indeed span the entire column span of \tilde{P}_{U_S} . Since the dimension of the column span of \tilde{P}_{U_S} is no larger than k , it is enough to the k -th singular value of \tilde{Q}_{U_S} :

Lemma 3.10 (Lemma 3.4 restated). *Given \tilde{M}_4 , the exact 4-th order moment of the ρ -smooth mixture of Gaussians, define the matrix \tilde{Q}_{U_S} as in (3.23) with the two-dimensional slices of \tilde{M}_4 . For any $\epsilon > 0$, and for some absolute constant $C_1, C_2, C_3 >$*

0, with probability at least $1 - 2(C_1\epsilon)^{C_2n}$, the k -th singular value of \tilde{Q}_{U_S} is bounded below by:

$$\sigma_k(\tilde{Q}_{U_S}) \geq C_3\omega_o(\epsilon\rho)^2n^{1.5}.$$

Similar as before, we first examine the structure of the matrix \tilde{Q}_{U_S} :

Claim 3.8 (Structural). *Under the same assumption as Lemma 3.10, we can write \tilde{Q}_{U_S} in the following matrix product form:*

$$\tilde{Q}_{U_S} = \tilde{P}_{U_S} D_{\tilde{\omega}} \tilde{\Sigma}_J^\top. \quad (3.25)$$

The columns of the matrix $\tilde{P}_{U_S} \in \mathbb{R}^{n^2 \times k}$ are the vectorized and projected covariance matrices as defined in (3.24); $D_{\tilde{\omega}}$ is the diagonal matrix with the mixing weights $\tilde{\omega}_i$ on its diagonal; and the matrix $\tilde{\Sigma}_J$ is defined as:

$$\tilde{\Sigma}_J = \left[\text{vec}[\tilde{\Sigma}_{(j_1, j_2)}^{(i)}] : (j_1, j_2) \in \mathcal{J} : i \in [k] \right] \in \mathbb{R}^{|\mathcal{J}| \times k}.$$

Proof. This claim follows from Claim 3.7, and the rule of matrix product. The coefficients $\tilde{\omega}_i \tilde{\Sigma}_{j_1, j_2}^{(i)}$ for the linear combinations of $\text{vec}(\text{Proj}_{\mathcal{J}^\perp} \tilde{\Sigma}^{(i)})$ are given by the columns of the product $D_{\tilde{\omega}} \tilde{\Sigma}_J^\top$. The coefficients are then multiplied by \tilde{P}_{U_S} to select the correct columns. \square

To prove Lemma 3.10, similar to the proof ideas of Lemma 3.8, we lower bound the k -th singular value of all the three factors.

Proof of Lemma 3.10 By the structural Claim 3.8, we know the matrix \tilde{Q}_{U_S} can be written as a product of the three matrices as $\tilde{Q}_{U_S} = \tilde{P}_{U_S} D_{\tilde{\omega}} \tilde{\Sigma}_J^\top$.

We lower bound the k -th singular value of each of the three factors. It is easy for the last two matrices. Note that by assumption $\sigma_k(D_{\tilde{\omega}}) \geq \omega_o$, and since $\tilde{\Sigma}_J^\top$ is just a perturbed rectangular matrix, we can apply Lemma 3.31 and with high probability we have $\sigma_k(\tilde{\Sigma}_J) \geq \Omega(\rho\sqrt{n})$.

The first matrix \tilde{P}_{U_S} is more subtle. Let us define the projection $D_{S^\perp} = \text{Proj}_{S^\perp} \otimes_{kr} I_n \in \mathbb{R}^{n^2 \times n^2}$. This is just a way of saying “apply the projection Proj_{S^\perp} to all columns” and then vectorize the matrix. In particular, for any matrix A we have $D_{S^\perp} \text{vec}(A) = \text{vec}(\text{Proj}_{S^\perp} A)$, therefore by definition of \tilde{P}_{U_S} we can write $\tilde{P}_{U_S} = D_{S^\perp} \tilde{\Sigma}$.

However, we cannot apply the same trick to directly bound the smallest singular value of D_{S^\perp} and $\text{Proj}_{D_{S^\perp}} \tilde{\Sigma}$ separately. The problem here is that D_{S^\perp} and $\tilde{\Sigma}$ are not independent, as the subspace S obtained in Step 1(a) also depends on the perturbation on $\tilde{\Sigma}$, therefore $\text{Proj}_{D_{S^\perp}} \tilde{\Sigma}$ is not simply a projected perturbed matrix. Instead, we show that even conditioned on the part of randomness that is common in S and $\tilde{\Sigma}$, $\tilde{\Sigma}$ still has sufficient randomness due to the high dimensions, and we can still extract a tall random matrix out of it. This is elaborated in the following claim:

Claim 3.9. *Under the assumptions of Lemma 3.10, with high probability the matrix $\tilde{P}_{U_S} = D_{S^\perp} \tilde{\Sigma}$ has smallest singular value at least $\Omega(\rho n)$.*

Let \mathcal{L} be the set of the (j_1, j_2) -th entries of $\tilde{\Sigma}^{(i)}$ for all i and one of j_1, j_2 is in the set \mathcal{H} . By Step 1(a), the subspace $\mathcal{S}' = \text{span}(S, e_j : j \in \mathcal{H})$ is only dependent on the entries in \mathcal{L} . Here we need to include the span of e_j 's for $j \in \mathcal{H}$ because the *diagonal* entries can depend on the other random perturbations. By adding the span of the vector e_j 's for $j \in \mathcal{H}$ the subspace remains invariant no matter how the diagonal entries change.

Let $\mathcal{Z} = \text{span}(\Sigma, \mathcal{S}' \otimes_{kr} I_n)$, and recall that the columns of Σ are the factorization of the unperturbed covariance matrices. The subspace \mathcal{Z} has dimension no larger than $|\mathcal{H}|((k+1)n + k) \leq n^2/10$, and depends on the randomness of \mathcal{L} .

Let $\tilde{\Sigma} = \Sigma + E$ where E is the random perturbation matrix. Now we condition on the randomness in \mathcal{L} . By definition the subspace \mathcal{Z} is deterministic conditional on \mathcal{L} . However, even if we only consider entries of $E \setminus \mathcal{L}$ there are still at least $\binom{n-k|\mathcal{H}|}{2} \geq n^2/4$ independent random variables. We shall show the randomness is enough to guarantee that the smallest singular value of $\text{Proj}_{D_{S^\perp}} \tilde{\Sigma}$ is lower bounded with high probability

conditioned on \mathcal{L} :

$$\begin{aligned}
\sigma_k(\tilde{P}_{U_S}) &= \sigma_k(D_{S^\perp} \tilde{\Sigma}) \\
&\geq \sigma_k(\text{Proj}_{\mathcal{Z}^\perp} \tilde{\Sigma}) \\
&= \sigma_k(\text{Proj}_{\mathcal{Z}^\perp} \Sigma + \text{Proj}_{\mathcal{Z}^\perp} E) \\
&= \sigma_k(\text{Proj}_{\mathcal{Z}^\perp} E).
\end{aligned}$$

Here we used the fact that projection to a subspace cannot increase the singular values (Lemma 3.28).

Conditioned on the randomness of entries in \mathcal{L} , $E \setminus \mathcal{L}$ still has at least $n^2/4$ random directions, while the dimension of the deterministic subspace \mathcal{Z} is at most $n^2/10$. Therefore we can apply Lemma 3.31 again to argue that conditionally, for every $\epsilon > 0$, with probability at least $1 - (C_1\epsilon)^{C_2n^2}$ we have:

$$\sigma_k(\tilde{P}_{U_S}) \geq \epsilon \rho \sqrt{C_3 n^2}.$$

In summary, apply union bound and we can conclude that with probability at least $1 - (C_1\epsilon)^{C_2n}$,

$$\sigma_k(\tilde{Q}_{U_S}) = \sigma_k(\tilde{P}_{U_S}) \sigma_k(D_{\tilde{\omega}}) \sigma_k(\tilde{\Sigma}_J) \geq C_3 \omega_o(\epsilon \rho)^2 n^{1.5}.$$

□

Next, we again use matrix perturbation bounds to prove the robustness of this step, which depends on the singular value decomposition of the matrix \tilde{Q}_{U_S} .

Lemma 3.11 (Lemma 3.7 restated). *Given the empirical 4-th order moments $\widehat{M}_4 = \widetilde{M}_4 + E_4$, and given the output $\text{Proj}_{\widehat{S}^\perp}$ from Step 1 (a). Suppose that $\|\text{Proj}_{\widehat{S}^\perp} - \text{Proj}_{\widetilde{S}^\perp}\| \leq \delta_1$, and suppose that the absolute value of entries of E_4 are at most δ_2 for $\delta_2 \leq \|\tilde{Q}_{U_S}\|_F / \sqrt{n^3}$. Conditioned on the high probability event $\sigma_k(\tilde{Q}_{U_S}) > 0$, we have:*

$$\|\text{Proj}_{\widehat{U}_S} - \text{Proj}_{\widetilde{U}_S}\| \leq \frac{n^{2.5} (1 + 2\delta_1/\delta_2)}{\sigma_k(\tilde{Q}_{U_S})} \delta_2. \quad (3.26)$$

Proof of Lemma 3.11 Note that the columns of U_S are the leading left singular vectors of \tilde{Q}_{U_S} . We want to apply the perturbation bound of singular vectors.

Similar to the proof of Lemma 3.9, we first need to bound the spectral distance between \hat{Q}_{U_S} and \tilde{Q}_{U_S} . In fact we will even bound the Frobenius norm difference:

$$\begin{aligned}
\|\hat{Q}_{U_S} - \tilde{Q}_{U_S}\|_F &= \|\hat{D}_{S^\perp} \hat{Q}_{U_0} - \tilde{D}_{S^\perp} \tilde{Q}_{U_0}\|_F \\
&= \|\tilde{D}_{S^\perp}(\hat{Q}_{U_0} - \tilde{Q}_{U_0}) + (\hat{D}_{S^\perp} - \tilde{D}_{S^\perp})\tilde{Q}_{U_0} + (\hat{D}_{S^\perp} - \tilde{D}_{S^\perp})(\hat{Q}_{U_0} - \tilde{Q}_{U_0})\|_F \\
&\leq \|\tilde{D}_{S^\perp}\|_F \|\hat{Q}_{U_0} - \tilde{Q}_{U_0}\|_F + 2\|\hat{D}_{S^\perp} - \tilde{D}_{S^\perp}\|_F \|\tilde{Q}_{U_0}\|_F \\
&\leq \sqrt{n^2} \|\tilde{D}_{S^\perp}\|_2 \|\hat{Q}_{U_0} - \tilde{Q}_{U_0}\|_F + 2\sqrt{n} \|\text{Proj}_{\hat{S}^\perp} - \text{Proj}_{\tilde{S}^\perp}\|_F \|\tilde{Q}_{U_0}\|_F \\
&\leq n\sqrt{n^2|\mathcal{J}|\delta_2^2} + 2\sqrt{n}\sqrt{n^2|\mathcal{J}|} \|\text{Proj}_{\hat{S}^\perp} - \text{Proj}_{\tilde{S}^\perp}\|_F \\
&\leq n^2 \frac{|\mathcal{H}|}{\sqrt{2}} (1 + 2\|\text{Proj}_{\hat{S}^\perp} - \text{Proj}_{\tilde{S}^\perp}\|_2 / \delta_2) \delta_2,
\end{aligned}$$

where we used the assumption $\|\tilde{\Sigma}^{(i)}\| \leq 1$ to bound $\|\tilde{Q}_{U_0}\|_F$, used the upperbound on $\|\hat{Q}_{U_0} - \tilde{Q}_{U_0}\|_F$ to bound the term $\|(\hat{D}_{S^\perp} - \tilde{D}_{S^\perp})(\hat{Q}_{U_0} - \tilde{Q}_{U_0})\|_F \leq \|(\hat{D}_{S^\perp} - \tilde{D}_{S^\perp})\|_F \delta_2 \sqrt{n^2|\mathcal{J}|} \leq \|(\hat{D}_{S^\perp} - \tilde{D}_{S^\perp})\|_F \|\tilde{Q}_{U_0}\|_F$, and used the fact that Frobenius norm is sub-multiplicative. Apply Wedin's Theorem (in particular the corollary Lemma 1.5), we can conclude (3.26). \square

Step 1 (c). Finding U by Merging the Two Projected Span

Algorithm 7: MergeProjections

Input: two subspaces $S_1, S_2 \in \mathbb{R}^{n \times ks}$, two subspaces $U_1, U_2 \in \mathbb{R}^{n^2 \times k}$ (the span of covariance matrices projected to the corresponding S_1^\perp, S_2^\perp).

Output: $\text{span}\{\Sigma^{(i)} : i \in [k]\}$, represented by an orthonormal matrix $U \in \mathbb{R}^{n^2 \times k}$.

Let A be the first $2ks$ left singular vectors of $[S_1, S_2]$.

Let S_3 be the first $(n - 2ks)$ left singular vectors of $I - AA^\top$.

Let $Q = [I_{n^2}, \text{Proj}_{(S_3 \otimes_{kr} I_n)} \text{Proj}_{U_1}]^\top U_2$, compute the SVD of Q .

Return: matrix U , whose columns are the first k left singular vectors Q .

Pick two disjoint sets of indices $\mathcal{H}_1, \mathcal{H}_2$, and repeat Step 1 (a) and Step 1 (b) on each of them to get \tilde{S}_j^\perp and \tilde{U}_j for $j = 1, 2$. In Step 1 (c), we merge the two span \tilde{U}_1

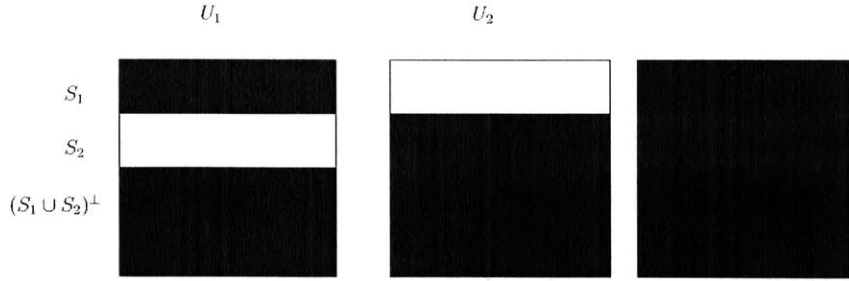


Figure 3-4: Step 1(c): Merging two subspaces.

and \tilde{U}_2 to get U .

If we are given two projections $\text{Proj}_{S_1^\perp} U$ and $\text{Proj}_{S_2^\perp} U$ of a *matrix* U , and if the union of the two subspaces S_1^\perp and S_2^\perp have full rank, namely $\dim(S_1 \cup S_2) = n$, then we can recover U by:

$$U = \begin{bmatrix} \text{Proj}_{S_1^\perp} \\ \text{Proj}_{S_2^\perp} \end{bmatrix}^\dagger \begin{bmatrix} \text{Proj}_{S_1^\perp} U \\ \text{Proj}_{S_2^\perp} U \end{bmatrix}.$$

However, it is slightly different if we are given two projections of a *subspace* \mathcal{U} , since a subspace can be equivalently represented by different orthonormal basis up to linear transformation.

In particular, in our setting for $j = 1, 2$, we can write $\tilde{U}_j = (\text{Proj}_{S_j^\perp} \otimes_{kr} I_n) \tilde{\Sigma} W_j$ for some fixed but unknown full rank matrix W_j (which makes the columns of matrix $\tilde{\Sigma} W_j$ an orthonormal basis of \mathcal{U}). Recall that we define $\tilde{\Sigma} \equiv [\text{vec}(\tilde{\Sigma}^{(i)}) : i \in [k]]$, and $D_{S_j^\perp} \equiv \text{Proj}_{S_j^\perp} \otimes_{kr} I_n$ for $j = 1, 2$.

The following Lemma shows that we can still *robustly* recover the *subspace* \mathcal{U} if the two projections have sufficiently large overlapping. The basic idea is to use the overlapping part to align the two basis of the subspace which the two projections act on.

Lemma 3.12 (Robustly merging two projections of an unknown subspace). *This is the detailed statement of Condition 3.10.*

Let the columns of two fixed but unknown matrices $V_1 \in \mathbb{R}^{n \times k}$ and $V_2 \in \mathbb{R}^{n \times k}$ form two basis (not necessarily orthonormal) of the same k -dimensional fixed but unknown

subspace \mathcal{U} in \mathbb{R}^n .

For two s -dimensional known subspaces S_1 and S_2 , Let the columns of A be the first $2s$ singular vectors of $[S_1, S_2]$, and let the columns of S_3 correspond to the first $(n - 2s)$ singular vectors of $(I_n - \text{Proj}_A)$, therefore $S_3 \subset (S_1 \cup S_2)^\perp$. Suppose that $\sigma_k(\text{Proj}_{S_3}U) > 0$ and that $\sigma_{2s}([S_1, S_2]) > 0$. Define matrices $U_1 = \text{Proj}_{S_1^\perp}V_1$ and $U_2 = \text{Proj}_{S_2^\perp}V_2$ and we know that $U_1^\top U_1 = U_2^\top U_2 = I_k$.

We are given $\widehat{S}_1, \widehat{S}_2$ and $\widehat{U}_1, \widehat{U}_2$, and suppose that for $j = 1, 2$, we have $\|\widehat{S}_j - S_j\|_F \leq \delta_s$ and $\|\widehat{U}_j - U_j\|_F \leq \delta_u$, for $\delta_s \leq 1, \delta_u \leq 1$.

Let the columns of \widehat{A} be the first $2s$ singular vectors of $[\widehat{S}_1, \widehat{S}_2]$, and let the columns of \widehat{S}_3 be the first $(n - 2s)$ singular vectors of $(I_n - \text{Proj}_{\widehat{A}})$. Define matrix $\widehat{U} \in \mathbb{R}^{n \times 2k}$ to be:

$$\widehat{U} = \left[\widehat{U}_2, \widehat{U}_1(\widehat{S}_3^\top \widehat{U}_1)^\dagger(\widehat{S}_3^\top \widehat{U}_2) \right] \quad (3.27)$$

If $\sigma_k(\text{Proj}_{S_3}U) > 0$ and $\sigma_{2s}([S_1, S_2]) > 0$, then for some absolute constant C we have:

$$\|\text{Proj}_{\widehat{U}} - \text{Proj}_U\| \leq \frac{C\sqrt{k}(\delta_u + \delta_s/\sigma_{2s}([S_1, S_2]))}{\sigma_k(\text{Proj}_{S_3}U)^2\sigma_{2s}([S_1, S_2])^3}.$$

Proof. The proof will proceed in two steps, we first show that if we are given the exact inputs, namely $\delta_s = \delta_u = 0$, then the column span of \widehat{U} defined in (3.27) is identical to the desired subspace \mathcal{U} . Then we give a stability result using matrix perturbation bounds.

1. Solving the problem using exact inputs.

Given the exact inputs S_1, S_2, U_1, U_2 , first we show that under the conditions $\sigma_{2s}([S_1, S_2]) > 0$ and $\sigma_k(\text{Proj}_{S_3}U) > 0$, then we show that the column span of the matrix $[U_2, U_1(S_3^\top U_1)^\dagger(S_3^\top U_2)]$ is indeed identical to $\mathcal{U} = \text{span}(V_1) = \text{span}(V_2)$.

Claim 3.10. Under the same assumptions of Lemma 3.12, given a matrix $V \in \mathbb{R}^{k \times k}$ such that $V = V_1^\dagger V_2$, let Proj_{U_0} be the projection to the column span of $U_0 = [U_2, U_1 V]$, then we have $\text{Proj}_{U_0} = \text{Proj}_U$.

Proof. Given $V = V_1^\dagger V_2$, then $U_1 V = \text{Proj}_{S_1^\perp} V_1 V = \text{Proj}_{S_1^\perp} V_2$. Recall that by def-

inition $U_2 = \text{Proj}_{S_2^\perp} V_2$, then the problem is now reduced to the simple problem of merging two projections ($U_2 = \text{Proj}_{S_2^\perp} V_2$ and $U_1 V = \text{Proj}_{S_1^\perp} V_2$) of the same matrix (V_2). Therefore, to show that the columns of $U_0 = [U_2, U_1 V]$ indeed span V_2 and thus the desired subspace U , we only need to show that $[\text{Proj}_{S_1^\perp}, \text{Proj}_{S_2^\perp}]$ has full column span. We show this by bounding the smallest singular value of it:

$$\begin{aligned}
\sigma_n([\text{Proj}_{S_2^\perp}, \text{Proj}_{S_1^\perp}]) &\geq \sigma_{2s}([\text{Proj}_{S_2^\perp}, \text{Proj}_{S_1^\perp}] \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix}) \\
&= \sigma_{2s}(\begin{bmatrix} (I_n - S_2 S_2^\top) S_1, & (I_n - S_1 S_1^\top) S_2 \end{bmatrix}) \\
&= \sigma_{2s}(\begin{bmatrix} S_1, S_2 \end{bmatrix} \begin{bmatrix} I_s & -S_1^\top S_2 \\ -S_2^\top S_1 & I_s \end{bmatrix}) \\
&= \sigma_{2s}(\begin{bmatrix} S_1, S_2 \end{bmatrix} \begin{bmatrix} S_1^\top \\ -S_2^\top \end{bmatrix} \begin{bmatrix} S_1, -S_2 \end{bmatrix}) \\
&= \sigma_{2s}(\begin{bmatrix} S_1, S_2 \end{bmatrix} \begin{bmatrix} S_1, -S_2 \end{bmatrix}^\top \begin{bmatrix} S_1, -S_2 \end{bmatrix}) \\
&= \sigma_{2s}([S_1, S_2])^3 \\
&> 0, \tag{3.28}
\end{aligned}$$

where the last inequality is by the assumption that $\sigma_{2s}([S_1, S_2]) > 0$. \square

Next, we show that in the exact case, the matrix $V = V_1^\dagger V_2$ can be computed by $V = (S_3^\top U_1)^\dagger (S_3^\top U_2)$. The basic idea is to use the overlapping part of the two projections U_1 and U_2 to align the two basis V_1 and V_2 . Recall that by its construction, $S_3 = (S_1 \cup S_2)^\perp = S_1^\perp \cap S_2^\perp$, and $\text{Proj}_{S_3} = \text{Proj}_{S_1^\perp \cap S_2^\perp}$. Then for $j = 1$ and 2 , we have:

$$S_3^\top U_j = S_3^\top \text{Proj}_{S_j^\perp} V_j = S_3^\top (\text{Proj}_{S_3} \text{Proj}_{S_j^\perp} + \text{Proj}_{S_3} \text{Proj}_{S_j^\perp}) V_j = S_3^\top (0 + \text{Proj}_{S_3}) V_j = S_3^\top V_j.$$

Moreover, since $U_j = \text{Proj}_{S_j^\perp} V_j$ is an orthonormal matrix, we have that all singular values of V_j are equal or greater than 1. Also note that U is an orthonormal matrix, so we have that $\sigma_k(\text{Proj}_{S_3} V_j) \geq \sigma_k(\text{Proj}_{S_3} U) > 0$. In other words, $S_3^\top V_j$ has full

column rank k . Therefore,

$$\begin{aligned}
V &= (S_3^\top U_1)^\dagger (S_3^\top U_2) \\
&= (S_3^\top V_1)^\dagger (S_3^\top V_2) \\
&= (V_1^\top S_3 S_3^\top V_1)^{-1} V_1^\top S_3 (S_3^\top V_2) \\
&= (V_1^\top S_3 S_3^\top V_1)^{-1} V_1^\top S_3 S_3^\top V_1 V_1^\dagger V_2 \\
&= V_1^\dagger V_2
\end{aligned}$$

where the third equality is the Moore-Penrose definition, the fourth equality is because V_1 and V_2 are basis of the same subspace, there exists some full rank matrix $X \in \mathbb{R}^{k \times k}$ such that $V_2 = V_1 X$, so we have $V_1 V_1^\dagger V_2 = V_1 V_1^\dagger V_1 X = V_1 X = V_2$.

2. Stability result.

Given $\widehat{S}_1, \widehat{S}_2$ and $\widehat{U}_1, \widehat{U}_2$ which are close to the exact S_1, S_2, U_1 and U_2 , we then need to bound the distance $\|\text{Proj}_{\widehat{U}} - \text{Proj}_U\|$. This follows the standard perturbation analysis. In order to apply Lemma 1.5 we need to bound the distance between $\|\widehat{U} - U_0\|_F$, and lower bound the smallest singular value of U_0 , namely $\sigma_k(U_0)$. Recall that we define U_0 same as in (3.27) for the exact case with $\delta_s = \delta_u = 0$.

First, we bound $\|\widehat{U} - U_0\|_F$. Note that we can write U_0^\top as $U_0^\top = U_2 B$, where $B = [I, \quad U_1 (S_3^\top U_1)^\dagger S_3]^\top$.

Recall that $S_3 = (S_1 \cup S_2)^\perp$, apply Lemma 1.5 and we have:

$$\|\widehat{S}_3 - S_3\| \leq \|\text{Proj}_{\widehat{S}_1 \cup \widehat{S}_2} - \text{Proj}_{S_1 \cup S_2}\| \leq \sqrt{2} \frac{\|[\widehat{S}_1, \widehat{S}_2] - [S_1, S_2]\|_F}{\sigma_{2s}([S_1, S_2])} \leq \frac{2\sqrt{2}\delta_s}{\sigma_{2s}([S_1, S_2])}.$$

Next, note that $\|\widehat{S}_3 - S_3\| < 1$ and $\|\widehat{U}_1 - U_1\| \leq \delta_u < 1$, apply Lemma 1.6 we have:

$$\|\widehat{S}_3^\top \widehat{U}_1 - S_3^\top U_1\| \leq 2(\|\widehat{S}_3 - S_3\| + \|\widehat{U}_1 - U_1\|).$$

Next, note that $\sigma_k(S_3^\top U_1) = \sigma_k(\text{Proj}_{S_3} V_1) > 0$ by assumption. Apply Lemma 1.7, we

have:

$$\|(\widehat{S}_3^\top \widehat{U}_1)^\dagger - (S_3^\top U_1)^\dagger\| \leq \frac{2\sqrt{2}\|\widehat{S}_3^\top \widehat{U}_1 - S_3^\top U_1\|}{\sigma_k(\text{Proj}_{S_3} V_1)^2}.$$

Next, apply Lemma 1.6 again we can bound the perturbation of matrix product:

$$\begin{aligned} \|\widehat{U} - U_0\| &= \|\widehat{U}_2 \widehat{B} - U_2 B\| \\ &\leq 2(\|\widehat{U}_2 - U_2\| + \|\widehat{B} - B\|) \\ &= 2(\|\widehat{U}_2 - U_2\| + \|\widehat{U}_1(\widehat{S}_3^\top \widehat{U}_1)^\dagger \widehat{S}_3 - U_1(S_3^\top U_1)^\dagger S_3\|) \\ &\leq 2(\|\widehat{U}_2 - U_2\| + 4(\|\widehat{U}_1 - U_1\| + \|(\widehat{S}_3^\top \widehat{U}_1)^\dagger - (S_3^\top U_1)^\dagger\| + \|\widehat{S}_3 - S_3\|)). \\ &\leq \frac{C(\delta_u + \delta_s/\sigma_{2s}([S_1, S_2]))}{\sigma_k(\text{Proj}_{S_3} V_1)^2}, \end{aligned}$$

where C is some absolute constant, and the last inequality summarizes the previous three inequalities, and used the fact that $\sigma_k(\text{Proj}_{S_3} V_1) < 1$. Note that $\|\widehat{U} - U_0\|_F \leq \sqrt{k}\|\widehat{U} - U_0\|$.

We are left to bound $\sigma_k(U_0)$. Recall that $\sigma_k(V_2) \geq \sigma_k(U_2) = 1$, and we have shown that in the exact case $U_0 = [\text{Proj}_{S_2^\perp} V_2, \text{Proj}_{S_1^\perp} V_2]$. Then we can bound the smallest singular value of U_0 following the inequality in (3.28):

$$\sigma_k(U_0) \geq \sigma_n([\text{Proj}_{S_2^\perp}, \text{Proj}_{S_1^\perp}]) \geq \sigma_{2s}([S_1, S_2])^3.$$

Finally we can apply Lemma 1.5 to bound the distance between the projections by:

$$\|\text{Proj}_{\widehat{U}} - \text{Proj}_{U_0}\| \leq \frac{\sqrt{2}\|\widehat{U} - U_0\|_F}{\sigma_k(U_0)} \leq \frac{C\sqrt{k}(\delta_u + \delta_s/\sigma_{2s}([S_1, S_2]))}{\sigma_k(\text{Proj}_{S_3} V_1)^2 \sigma_{2s}([S_1, S_2])^3}.$$

□

In Step 1 (c), we are given the output \widetilde{U}_1 and \widetilde{U}_2 from Step 1 (b), as well as the output \widetilde{S}_1^\perp and \widetilde{S}_2^\perp from Step 1 (a). Recall that $\mathcal{U} = \text{span}\{\text{vec}(\widetilde{\Sigma}^{(i)}) : i \in [k]\}$, and for $j = 1, 2$, the matrix \widetilde{U}_j given by Step 1 (b) corresponds to the subspace \mathcal{U} projected

to the subspace $\tilde{B}_j = \tilde{S}_j^\perp \otimes_{kr} I_n$.

Let matrix $\tilde{S}_3 = \tilde{S}_1^\perp \cap \tilde{S}_2^\perp = (\tilde{S}_1 \cup \tilde{S}_2)^\perp$ (obtained by taking the singular vectors of $(I_n - AA^\top)$, where A corresponds to the first $2k|\mathcal{H}|$ singular vectors of $[\tilde{S}_1, \tilde{S}_2]$), and denote $\tilde{B}_3 = \tilde{S}_3 \otimes_{kr} I_n$. Define the matrix \tilde{Q}_U to be:

$$\tilde{Q}_U = \begin{bmatrix} \tilde{U}_2, & \tilde{U}_1(\tilde{B}_3\tilde{U}_1)^\dagger\tilde{B}_3\tilde{U}_2 \end{bmatrix}, \quad (3.29)$$

and similarly define the perturbed version \hat{Q}_U to be:

$$\hat{Q}_U = \begin{bmatrix} \hat{U}_2, & \hat{U}_1(\hat{B}_3\hat{U}_1)^\dagger\hat{B}_3\hat{U}_2 \end{bmatrix}.$$

Now we want to apply Lemma 3.12 to show that $\text{Proj}_{\tilde{Q}_U} = \text{Proj}_{\tilde{\Sigma}}$ and bound the distance $\|\text{Proj}_{\hat{Q}_U} - \text{Proj}_{\tilde{\Sigma}}\|$. In order to use the lemma, we first use smoothed analysis to show (in Lemma 3.13 and Lemma 3.14) that the conditions required by the lemma are all satisfied with high probability over the ρ -perturbation of the covariance matrices, then conclude the robustness of Step 1 (c) in Lemma 3.15.

Lemma 3.13. *With high probability, for some constant C*

$$\sigma_k(\text{Proj}_{\tilde{B}_3}\tilde{\Sigma}) \geq C\epsilon\rho n.$$

Proof. This is in fact exactly the same as Claim 3.9.

Given $\tilde{\Sigma} = \Sigma + E$, by the definition of \tilde{S}_3 and \tilde{B}_3 we know that \tilde{B}_3 only depends on the randomness of $P_j E$ for $i = 1, 2$, where

$$\mathcal{J} = \{(j_1, j_2) : j_1 \in \mathcal{H}_1 \cup \mathcal{H}_2, \text{ or } j_2 \in \mathcal{H}_1 \cup \mathcal{H}_2\},$$

and P_j denotes the mapping that only keeps the coordinates corresponding to the set \mathcal{J} . Therefore, we have:

$$\sigma_k(\text{Proj}_{\tilde{B}_3}\tilde{\Sigma}) \geq \sigma_k(\text{Proj}_{(\tilde{B}_3^\top \Sigma)^\perp} \text{Proj}_{\tilde{B}_3} E).$$

Note that the rank of \tilde{B}_3^\perp is $2nk|\mathcal{H}|$ and $|\mathcal{J}| = 2n|\mathcal{H}|$, thus $n_2 - |\mathcal{J}| - 2nk|\mathcal{H}| - k = \Omega(n^2) > 2k$. So we can apply Lemma 3.31 to conclude that for some absolute constants C_1, C_2, C_3 , with probability at least $1 - (C_1\epsilon)^{C_2n^2}$, $\sigma_k(\tilde{B}_3^\top \tilde{\Sigma}) \geq \epsilon\rho\sqrt{C_3n^2}$. \square

Lemma 3.14. *With high probability, for some constant C ,*

$$\sigma_{2k|\mathcal{H}|}([\tilde{S}_1, \tilde{S}_2]) \geq C\omega_o(\epsilon\rho)^2n^{-0.25}.$$

Proof. For $i = 1, 2$, recall that \tilde{S}_i is the singular vectors of \tilde{Q}_{S_i} , where \tilde{Q}_{S_i} is defined with the set \mathcal{H}_i as in (3.12). We can write the singular value decomposition of \tilde{Q}_{S_i} as $\tilde{Q}_{S_i} = \tilde{S}_i\tilde{D}_i\tilde{V}_i^\top$ for some diagonal matrix \tilde{D}_i and orthonormal matrix \tilde{V}_i , and

$$[\tilde{S}_1, \tilde{S}_2] = [\tilde{Q}_{S_1}, \tilde{Q}_{S_2}] \begin{bmatrix} \tilde{V}_1\tilde{D}_1^{-1} & 0 \\ 0 & \tilde{V}_2\tilde{D}_2^{-1} \end{bmatrix}.$$

Note that we can write $[\tilde{Q}_{S_1}, \tilde{Q}_{S_2}] = [\tilde{P}_{S_1}, \tilde{P}_{S_2}](\text{diag}(B_{\tilde{S}_1}, B_{\tilde{S}_2}))^\top$, and following almost exactly with the proof of Lemma 3.8, we can argue that, with probability at least $1 - (C_1\epsilon)^{C_2n}$,

$$\sigma_{2k|\mathcal{H}|}([\tilde{Q}_{S_1}, \tilde{Q}_{S_2}]) \geq C\omega_o(\epsilon\rho)^2n.$$

Moreover, by the structure of M_4 and the bounds on $\tilde{\Sigma}^{(i)} \prec \frac{1}{2}I$, we can bound $\|\tilde{Q}_{S_i}\| \leq 3\sqrt{n(|\mathcal{H}|/3)^3}$, and thus:

$$\sigma_{k|\mathcal{H}|}(V_i\tilde{D}_i^{-1}) = \frac{1}{\sigma_{\max}(\tilde{Q}_{S_i})} \geq \frac{1}{3\sqrt{n(|\mathcal{H}|/3)^3}} = \Omega(n^{-1.25}).$$

Therefore, we can conclude that, for some absolute constant C , we have:

$$\sigma_{2k|\mathcal{H}|}([\tilde{S}_1, \tilde{S}_2]) \geq C\omega_o(\epsilon\rho)^2n^{-0.25}.$$

\square

In the next lemma, we apply Lemma 3.12 to show that under perturbation, with

high probability the column span of $\text{Proj}_{\widehat{Q}_U} = \text{Proj}_{\widehat{\Sigma}}$ and this step is robust.

Lemma 3.15. *Given the output $\widehat{S}_1, \widehat{S}_2$ and $\widehat{U}_1, \widehat{U}_2$ from Step 1 (a) and (b) based on the empirical moments \widehat{M}_4 . Suppose that for $i = 1, 2$, $\|\widehat{S}_i - \widetilde{S}_i\|_F \leq \delta_s$, $\|\widehat{U}_i - \widetilde{U}_i\|_F \leq \delta_u$ for $\delta_s, \delta_u < 1$. Let the columns of $\widetilde{U} \in \mathbb{R}^{n^2 \times k}$ be the k leading singular vectors of \widetilde{Q}_U defined in (3.29). Then for some absolute constants C , with high probability,*

$$\|\text{Proj}_{\widehat{U}} - \text{Proj}_{\widetilde{U}}\| \leq \frac{C\sqrt{k}(\delta_u + \delta_s n^{0.75}/(\omega_o \epsilon^2 \rho^2))}{\omega_o^3 \epsilon^8 \rho^8 n^{1.25}}. \quad (3.30)$$

Note that $\sigma_{2k|\mathcal{H}|n}([\widetilde{B}_1, \widetilde{B}_2]) = \sigma_{2k|\mathcal{H}|}([\widetilde{S}_1, \widetilde{S}_2])$, and for $i = 1, 2$, we have $\|\widehat{B}_i - \widetilde{B}_i\|_F \leq \sqrt{n}\|\widehat{S}_i - \widetilde{S}_i\|_F \leq \sqrt{n}\delta_s$. Therefore, with the above two smoothed analysis Lemmas showing polynomial bound of $\sigma_{2k|\mathcal{H}|}([\widetilde{S}_1, \widetilde{S}_2])$ and $\sigma_k(\text{Proj}_{\widetilde{B}_3}(\widetilde{\Sigma}))$, the proof of Lemma 3.15 follows by applying Lemma 3.12.

3.4.2 Step 2 of Zero-Mean Case. Moments Unfolding

Algorithm 8: Estimate $Y_4 Y_6$

Input: 4-th order moments $\overline{M}_4 \in \mathbb{R}^{n^4}$, 6-th order moments $\overline{M}_6 \in \mathbb{R}^{n^6}$, the span of (vectorized with distinct entries) covariance matrices $U \in \mathbb{R}^{n^2 \times k}$.

Output: Unfolded moments in the coordinate system of U :

$$Y_4 \in \mathbb{R}_{sym}^{k \times k}, Y_6 \in \mathbb{R}_{sym}^{k \times k \times k}.$$

Let Y_4 be the solution to $\min_{Y_4 \in \mathbb{R}_{sym}^{k \times k}} \|\sqrt{3}\mathcal{F}_4(UY_4U^\top) - \overline{M}_4\|_F^2$.

Let Y_6 be the solution to $\min_{Y_6 \in \mathbb{R}_{sym}^{k \times k \times k}} \|\sqrt{15}\mathcal{F}_6(Y_6(U^\top, U^\top, U^\top)) - \overline{M}_6\|_F^2$.

Return: Y_4, Y_6 .

In the second step of the algorithm, we solve two systems of linear equations to recover the unfolded moments.

Unfolding the 4-th Order Moments

Recall the first system of linear equations is

$$\overline{M}_4 = \sqrt{3}\mathcal{F}_4 \circ \mathcal{X}_4^U(Y_4).$$

In the equation, $Y_4 \in \mathbb{R}_{sym}^{k \times k}$ is the unknown variable which can be viewed as a $k \times k$ symmetric matrix. Given $U \in \mathbb{R}^{n_2 \times k}$, the column span of $\tilde{\Sigma}$ that we learned in Step 1, the first linear transformation \mathcal{X}_4^U is simply $\mathcal{X}_4^U(Y_4) = UY_4U^\top$. It is supposed to transform Y_4 into the unfolded moments $X_4 \in \mathbb{R}_{sym}^{n_2 \times n_2}$, which is defined to be $\sum_{i=1}^k w_i \text{vec}(\tilde{\Sigma}^{(i)}) \text{vec}(\tilde{\Sigma}^{(i)})^\top$. The next transformation $\sqrt{3}\mathcal{F}_4$ maps the unfolded moments X_4 to the folded moments $\overline{M}_4 \in \mathbb{R}^{n_4}$. As we showed in Lemma 3.1, the mapping \mathcal{F}_4 is a projection.

Since U is the column span matrix of $\tilde{\Sigma}$, there must exist a Y_4 such that $X_4 = \tilde{\Sigma} D_{\tilde{\omega}} \tilde{\Sigma}^\top = UY_4U^\top$ (recall that $D_{\tilde{\omega}}$ is the diagonal matrix with entries $\tilde{\omega}_i$), so the system must have at least one solution.

Rewrite the system of linear equations $\overline{M}_4/\sqrt{3} = \mathcal{F}_4 \circ \mathcal{X}_4^U(Y_4)$ in the canonical form: $\overline{M}_4\sqrt{3} = H_4 \text{vec}(Y_4)$ where the variable $\text{vec}(Y_4) \in \mathbb{R}^{k_2}$, and the coefficient matrix $H_4 \in \mathbb{R}^{n_4 \times k_2}$ is a function of U and therefore also a function of the parameter Σ (recall $n_4 = \binom{n}{4}$ and $k_2 = \binom{k+1}{2}$). The system has a *unique* solution if the smallest singular value of the coefficient matrix H_4 is greater than zero.

The main theorem of this section shows that with high probability over the ρ -perturbation the system has a *unique* solution:

Theorem 3.12. *With high probability over the ρ -perturbation of $\tilde{\Sigma}$, the smallest singular value of the coefficient matrix \tilde{H}_4 is lower bounded by $\sigma_{\min}(\tilde{H}_4) \geq \Omega(\rho^2 n/k)$. As a corollary, the system has a unique solution.*

In order to prove this theorem, we first need the following structural lemma:

Lemma 3.16. *The coefficient matrix \tilde{H}_4 is equal to $\tilde{A}_4 \tilde{B}_4$. The first matrix $\tilde{A}_4 \in \mathbb{R}^{n_4 \times k_2}$ has columns indexed by pair $\{(i, j) : 1 \leq i \leq j \leq k\}$, and the (i, j) -th column is equal to $C_{i,j} \mathcal{F}_4(\text{vec}(\tilde{\Sigma}^{(i)}) \odot \text{vec}(\tilde{\Sigma}^{(j)}))$. Here $C_{i,j} = 1$ if $i = j$ and $C_{i,j} = 2$ if $i < j$. The second matrix $\tilde{B}_4 \in \mathbb{R}^{k_2 \times k_2}$ transforms a $k \times k$ symmetric matrices Y_4 into:*

$$\tilde{B}_4 \text{vec}(Y_4) = \text{vec}((\tilde{\Sigma}^\dagger U) Y_4 (\tilde{\Sigma}^\dagger U)^\top).$$

Next we need to prove the bounds on the smallest singular values for \tilde{A}_4 and \tilde{B}_4 . The first matrix \tilde{A}_4 is essentially a projection of the Kronecker product $(\tilde{\Sigma} \otimes_{kr} \tilde{\Sigma})$.

In particular, this projection satisfy the “symmetric off-diagonal” property defined below:

Definition 3.3 (symmetric off-diagonal). *Let the columns of matrix $P \in \mathbb{R}^{n_2^2 \times d_2}$ form an (arbitrary) basis of the subspace \mathcal{P} , and index the rows of P by pair $(i, j) \in [n_2] \times [n_2]$. The subspace \mathcal{P} and the matrix P is called symmetric off-diagonal, if (i, i) -th row of P is 0 (“off-diagonal”), and the (i, j) -th row and (j, i) -th row are identical (“symmetric”).*

Remark 3.13. *Since symmetric off-diagonal is a property on the structure of rows of the basis P . If one basis of the subspace \mathcal{P} is symmetric off-diagonal, then any basis is too. Moreover, any orthogonal basis of the subspace \mathcal{P} will still be symmetric off-diagonal.*

Consider a Kronecker product of the same matrix $E \in \mathbb{R}^{n_2 \times k}$. The columns of $E \otimes_{kr} E$ are indexed by pair $(i, j) \in [k] \times [k]$. Consider applying a symmetric off-diagonal projection P^\top to the Kronecker product. By the property of symmetry the projection will map two columns $E_{[:,i]} \odot E_{[:,j]}$ and $E_{[:,j]} \odot E_{[:,i]}$ to the same vector. Therefore the projected Kronecker product $P^\top(E \otimes_{kr} E)$ will not have full column rank k^2 . However, we will show that the k_2 “unique” columns after the projection are linearly independent.

To formalize this, we define the matrix $(E \otimes_{kr} E)_{uniq} \in \mathbb{R}^{n_2^2 \times k_2}$ with the “unique” columns of $E \otimes_{kr} E$ labeled by pairs $\{(i, j) : 1 \leq i \leq j \leq k\}$. In particular,

$$[(E \otimes_{kr} E)_{uniq}]_{[:,(i,j)]} = E_{[:,i]} \odot E_{[:,j]}.$$

In the following main lemma, we show even after projection to any symmetric off-diagonal space with sufficiently many dimensions, the “unique” columns of a Kronecker product of random matrices still has good condition number.

Lemma 3.17. *Let $E \in \mathbb{R}^{n_2 \times k}$ be a Gaussian random matrix (each entry distributed as $\mathcal{N}(0, 1)$). Let $P \in \mathbb{R}^{n_2^2 \times d_2}$ be a symmetric off-diagonal subspace of dimension*

$d_2 = \Omega(n_2^2)$. Then for any constant $C > 0$, when $n_2 \geq k^{2+C}$ we have with high probability $\sigma_{\min}(P^\top(E \otimes_{kr} E)_{\text{uniq}}) \geq \Omega(n_2)$.

Let us first see how Theorem 3.12 follows from the two lemmas (Lemma 3.16 and Lemma 3.17).

Proof. (of Theorem 3.12) Using the structural Lemma 3.16, we know we only need to bound the smallest singular value of \tilde{A}_4 and \tilde{B}_4 separately. The following two claims directly imply the theorem.

Claim 3.11. $\sigma_{\min}(\tilde{A}_4) \geq \Omega(\rho^2 n_2)$.

Claim 3.12. $\sigma_{\min}(\tilde{B}_4) \geq 1/(4\|\tilde{\Sigma}\|^2) \geq 1/(4nk)$.

Next we prove the above two claims.

We apply Lemma 3.17 to prove Claim 3.11. Note that the ρ -perturbed covariances $\tilde{\Sigma}$ is not a random Gaussian matrix, yet it is equal to the unperturbed matrix Σ plus a random Gaussian matrix $E_\Sigma = \rho E^\top$. Since we consider arbitrary Σ , the columns of $\tilde{\Sigma}$ as well as the columns \tilde{A}_4 may not be incoherent.

Instead, we project \tilde{A}_4 to a subspace to strip away the terms involving the original matrix Σ . Let S be the range space corresponding to the projection \mathcal{F}_4 . Recall that $|S| = n_4 = \Omega(n_2^2)$, and by the definition of \mathcal{F}_4 , S is symmetric off-diagonal. Define the subspace $S' = \text{span}(S^\perp, \Sigma \otimes_{kr} I_{n_2}, I_{n_2} \otimes_{kr} \Sigma)$. Let $P = (S')^\perp$. By construction $|P| \geq |S| - 2kn_2 = \Omega(n_2^2)$. Also, since $P = (S')^\perp$ is a subspace of S , it must also be symmetric off-diagonal (see Remark 3.13). After projecting \tilde{A}_4 to P , we know that the (i, j) -th column ($1 \leq i \leq j \leq k$) of $P^\top \tilde{A}_4$ is given by:

$$\begin{aligned} P^\top [\tilde{A}_4]_{[:,(i,j)]} &= C_{i,j} P^\top (\Sigma_{[:,i]} \odot \Sigma_{[:,j]} + \rho E_{[:,i]} \odot \Sigma_{[:,j]} + \rho \Sigma_{[:,i]} \odot E_{[:,j]} + \rho^2 E_{[:,i]} \odot E_{[:,j]}) \\ &= C_{i,j} \rho^2 P^\top E_{[:,i]} \odot E_{[:,j]}. \end{aligned}$$

⁷Note that the diagonal entries are then arbitrarily perturbed, but we will project on a symmetric off-diagonal subspace so changes on diagonal entries do not change the result.

Thus in $P^\top \tilde{A}_4$ all the terms involving Σ disappears. Therefore

$$\begin{aligned}\sigma_{\min}(\tilde{A}_4) &\geq \sigma_{\min}(P^\top \tilde{A}_4) = \sigma_{\min}(P^\top (\tilde{\Sigma} \otimes_{kr} \tilde{\Sigma})_{\text{uniq}}) \\ &= \rho^2 \sigma_{\min}(P^\top (E \otimes_{kr} E)_{\text{uniq}}) \geq \Omega(\rho^2 n_2),\end{aligned}$$

where the first inequality is because the smallest singular value cannot become larger after projection, the first equality is by definition, the second equality is by the property of P , and the final step uses Lemma 3.17⁸.

For Claim 3.12. Pick any $Y_4 \in \mathbb{R}_{\text{sym}}^{k \times k}$, we have

$$\begin{aligned}\|\tilde{B}_4(Y_4)\| &= \|\text{vec}((\tilde{\Sigma}^\dagger U)Y_4(\tilde{\Sigma}^\dagger U)^\top)\| = \|(\tilde{\Sigma}^\dagger U)Y_4(\tilde{\Sigma}^\dagger U)^\top\|_F \\ &\geq \|Y_4\|_F \sigma_{\min}(\tilde{\Sigma}^\dagger U)^2 = \|Y_4\|_F / \|\tilde{\Sigma}\|^2,\end{aligned}$$

where the inequality is because $\|AB\|_F \geq \sigma_{\min}(A)\|B\|_F$ if $A \in \mathbb{R}^{m \times n}$ and $m \geq n$. Since $\|\text{vec}(Y_4)\|$ is within a factor of $\sqrt{2}$ to $\|Y_4\|_F$, and by the assumption $\tilde{\Sigma}^{(i)} \prec \frac{1}{2}I$ we can bound $\|\tilde{\Sigma}\| \leq \Omega(\sqrt{nk})$, we have the desired bound for $\sigma_{\min}(\tilde{B}_4)$. \square

Structure of the Coefficient Matrix Next we prove the structural Lemma 3.16.

Proof. (of Lemma 3.16) First, assume we know the true $\tilde{\Sigma}$ matrix, then in order to get the unfolded moments X_4 , we only need to solve the equation $\mathcal{F}_4(\tilde{\Sigma} D_4 \tilde{\Sigma}^\top) = \overline{M}_4$ with the $k \times k$ symmetric variable D_4 , and the solution should be equal to the diagonal matrix $D_{\tilde{\omega}}$.

However, we only know U which is the column span of $\tilde{\Sigma}$, so we can only use UY_4U^\top and let $UY_4U^\top = \tilde{\Sigma} D_4 \tilde{\Sigma}^\top$. Note that there is a one-to-one correspondence between Y_4 and D_4 . In particular we know $D_4 = (\tilde{\Sigma}^\dagger U)Y_4(\tilde{\Sigma}^\dagger U)^\top$, this is exactly the second part \tilde{B}_4 .

Now the first matrix \tilde{A}_4 should map $\text{vec}(D_4)$ to M_4 . By construction, the (i, j) -th column ($i < j$) of \tilde{A}_4 is equal to $\mathcal{F}_4(\tilde{\Sigma}^{(i)} \odot \tilde{\Sigma}^{(j)} + \tilde{\Sigma}^{(j)} \odot \tilde{\Sigma}^{(i)}) = 2\mathcal{F}_4(\tilde{\Sigma}^{(i)} \odot \tilde{\Sigma}^{(j)})$, since \mathcal{F}_4 is symmetric off-diagonal we know $\mathcal{F}_4(v_1 \odot v_2) = \mathcal{F}_4(v_2 \odot v_1)$ for any two vectors

⁸Note that although diagonal entries are not perturbed, we also have $P_{[i,i]} = 0$ so we can still apply the lemma.

v_1, v_2 . For the (i, i) -th column, by construction they are equal to $\mathcal{F}_4(\tilde{\Sigma}^{(i)} \odot \tilde{\Sigma}^{(i)})$ as we wanted. \square

Main Lemma on Projection of Kronecker Product In this part we prove Lemma 3.17.

The singular values of Kronecker Product between two matrices are well-understood: they are just the products of the singular values of the two matrices. Therefore, the Kronecker product of two rank k matrices will have rank k^2 . However, in our case the problem becomes more complicated because we only look at a projection of the resulting matrix. The projected Kronecker product may no longer have rank k^2 because of symmetry. Here we are able to show that even with projection to a low dimensional space, the rank of the new matrix is still as large as $\binom{k+1}{2}$.

The basic idea of the proof is to consider the inner-products between columns, and show that the columns are incoherent even after projection.

Proof. (of Lemma 3.17) Consider the matrix $(E \otimes_{kr} E)_{uniq}^\top P P^\top (E \otimes_{kr} E)_{uniq}$, we shall show the matrix is *diagonally dominant* and hence its smallest singular value must be large. In order to do that we need to prove the following two claims:

Claim 3.13. For any $i, j \leq k$, $i \leq j$, with high probability $\|P^\top(E_{[:,i]} \odot E_{[:,j]})\|^2 \geq \Omega(n_2^2)$.

Claim 3.14. For any $i, j \leq k$, $i \leq j$, with high probability

$$\sum_{1 \leq i' \leq j' \leq k, (i,j) \neq (i',j')} |\langle P^\top(E_{[:,i]} \odot E_{[:,j]}), P^\top(E_{[:,i']} \odot E_{[:,j']}) \rangle| \leq o(n_2^2).$$

With this two claims, we can apply Gershgorin's Disk Theorem 1.6 to conclude that $\sigma_{min}((E \otimes_{kr} E)_{uniq}^\top P P^\top (E \otimes_{kr} E)_{uniq}) \geq \Omega(n_2^2)$. Therefore $\sigma_{min}(P^\top(E \otimes_{kr} E)_{uniq}) \geq \Omega(n_2)$.

Now we prove the two claims. For Claim 3.13, it essentially says the projection of a random vector to a fixed subspace should have large norm. If the vector has independent entries, this is first shown in [113]. Recently [124] general-

ized the result to K -concentrated vectors, see Lemma 3.33. By Lemma 3.34 we know conditioned on $\|E_{[:,i]}\|, \|E_{[:,j]}\| \leq 2\sqrt{n_2}$, $(E_{[:,i]} \odot E_{[:,j]})_{p,q}$ ($p \neq q$) is $O(\sqrt{n_2})$ -concentrated. By assumption P ignores all the $(E_{[:,i]} \odot E_{[:,j]})_{p,p}$ entries. Therefore $\Pr[|\|P^\top(E_{[:,i]} \odot E_{[:,j]})\|^2 - d_2| \geq 2t\sqrt{d_2} + t^2] \leq Ce^{-\Omega(t^2/n_2)} + e^{-\Omega(n_2)}$. We then pick $t = \sqrt{d_2}/5 \geq \Omega(n_2)$, which implies $\Pr[\|P(E_{[:,i]} \odot E_{[:,j]})\|^2 \leq d_2/2] \leq Ce^{-\Omega(n_2)}$. This is what we need for Claim 3.13.

We need to bound terms of the form $\langle P^\top(E_{[:,i]} \odot E_{[:,j]}), P^\top(E_{[:,i']} \odot E_{[:,j']}) \rangle$ in order to show Claim 3.14. These are degree-4 Gaussian chaoses and are well-studied in [74].

We break the terms according to how many of i', j' appears in i, j .

Case 1: $i', j' \notin \{i, j\}$. In this case we first randomly pick $E_{[:,i]}, E_{[:,j]}$, and condition on the high probability event that $\|E_{[:,i]}\|, \|E_{[:,j]}\| \leq 2\sqrt{n_2}$. In this case the inner-product can be rewritten as $\langle PP^\top(E_{[:,i]} \odot E_{[:,j]}), (E_{[:,i']} \odot E_{[:,j']}) \rangle$, and we know $\|PP^\top(E_{[:,i]} \odot E_{[:,j]})\| \leq 4n_2$. Also, since P is symmetric off-diagonal we know in this degree-2 Gaussian chaos (only $E_{[:,i']}$ and $E_{[:,j']}$ are random now) there are no ‘‘diagonal’’ terms. Therefore the Decoupling Theorem 3.20 shows without loss of generality we can assume $i' \neq j'$. Apply Theorem 3.19 we know this term is bounded by $O(n_2^{1+\epsilon})$ with high probability for any $\epsilon > 0$.

Case 2: One of i', j' is in $\{i, j\}$. Without loss of generality assume $i' \in \{i, j\}$ (the other case is symmetric). Again we first randomly pick $E_{[:,i]}, E_{[:,j]}$ and condition on the high probability event that $\|E_{[:,i]}\|, \|E_{[:,j]}\| \leq 2\sqrt{n_2}$ (but this will also determine $E_{[:,i']}$). After the conditioning, only $E_{[:,j']}$ is still random, and the inner-product can be rewritten as $\langle \text{mat}(PP^\top(E_{[:,i]} \odot E_{[:,j]})E_{[:,i']}, E_{[:,j']}) \rangle$ where the fixed vector $\text{mat}(PP^\top(E_{[:,i]} \odot E_{[:,j]})E_{[:,i']})$ has norm bounded by $\|PP^\top(E_{[:,i]} \odot E_{[:,j]})\| \|E_{[:,i']}\| \leq 8n_2^{3/2}$. By property of Gaussian with high probability the inner-product is bounded by $O(n_2^{3/2+\epsilon})$ for any $\epsilon > 0$.

Case 3: $i', j' \in \{i, j\}$. Since i', j' cannot be equal to i, j , there is only one possibility: i', j' are both equal to one of i, j and $i \neq j$. Without loss of generality assume $i' = j' = i \neq j$. We can swap i, j with i', j' and this actually becomes Case 2. By the same argument we know this term is bounded by $O(n_2^{3/2+\epsilon})$ for any $\epsilon > 0$.

There are $O(k^2)$ terms in Case 1, $O(k)$ terms in Case 2 and $O(1)$ terms in Case 3.

Therefore by union bound we know the sum is bounded by $O(kn_2^{3/2+\epsilon} + k^2n_2^{1+\epsilon})$ with high probability. Recall we are assuming $n_2 \geq k^{2+C}$ (which only requires $n \geq k^{1+C/2}$). Choose ϵ to be a small enough constant depending on C gives the result. \square

Unfolding 6-th Order Moments

Recall the second system of linear equations is

$$\overline{M}_6/\sqrt{15} = \mathcal{F}_6 \circ \mathcal{X}_6^U(Y_6).$$

In the equation, $Y_6 \in \mathbb{R}_{sym}^{k \times k \times k}$ is the unknown variable which can be viewed as a $k \times k \times k$ symmetric tensor. The first linear transformation \mathcal{X}_6^U transforms Y_6 into the unfolded moments $X_6 \in \mathbb{R}_{sym}^{n_2 \times n_2 \times n_2}$, which is supposed to be equal to $\sum_{i=1}^k \tilde{w}_i \text{vec}(\tilde{\Sigma}^{(i)}) \otimes^3$. The transformation is simply $X_6 = \mathcal{X}_6^U(Y_6) = Y_4(U^\top, U^\top, U^\top)$ where $U \in \mathbb{R}^{n_2 \times k}$ is the column span of $\tilde{\Sigma}$ that we learned in the previous section.

The next transformation \mathcal{F}_6 maps the unfolded moments X_6 to the folded moments $\overline{M}_6 \in \mathbb{R}^{n_6}$, which as we showed in Lemma 3.1 is a projection. Recall that $n_6 = \binom{n}{6}$.

Rewrite the system of linear equations $\overline{M}_6/\sqrt{15} = \mathcal{F}_6 \circ \mathcal{X}_6^U(Y_6)$ in the canonical form: $\overline{M}_6/\sqrt{15} = \tilde{H}_6 \text{vec}(Y_6)$ where the coefficient matrix $\tilde{H}_6 \in \mathbb{R}^{n_6 \times k_3}$ is a function of U and therefore is a function of $\tilde{\Sigma}$ (recall $k_3 = \binom{k+2}{3}$).

The second system of linear equations tries to unfold the 6-th order moment \overline{M}_6 to get Y_6 . Similar to Theorem 3.12 the following theorem guarantees that with high probability over the perturbation the system has a unique solution.

Theorem 3.14. *With high probability over the perturbation, the coefficient matrix \tilde{H}_6 has smallest singular value $\sigma_{\min}(\tilde{H}_6) \geq \Omega(\rho^3(n/k)^{1.5})$. As a corollary, the system has a unique solution.*

The proof of this theorem is very similar to the proof of Theorem 3.12. Here we list the important steps and highlight the differences.

As before the theorem relies on a structural lemma (Lemma 3.18), and a main lemma about the symmetric off-diagonal projection of a Kronecker product of three identical matrices (Lemma 3.19).

Lemma 3.18. *The coefficient matrix \tilde{H}_6 is equal to $\tilde{A}_6\tilde{B}_6$. The first matrix $\tilde{A}_6 \in \mathbb{R}^{n_6 \times k_3}$ has columns indexed by triples (i_1, i_2, i_3) for $1 \leq i_1 \leq i_2 \leq i_3 \leq k$, and are given by:*

$$[\tilde{A}_6]_{[:,(i_1, i_2, i_3)]} = C_{i_1, i_2, i_3} \mathcal{F}_6(\text{vec}(\tilde{\Sigma}^{(i_1)}) \odot \text{vec}(\tilde{\Sigma}^{(i_2)}) \odot \text{vec}(\tilde{\Sigma}^{(i_3)})),$$

where C_{i_1, i_2, i_3} is a constant depending only on multiplicity of the indices (i_1, i_2, i_3) .

The second matrix $\tilde{B}_6 \in \mathbb{R}^{k_3 \times k_3}$ transforms a $k \times k \times k$ symmetric tensor Y_6 into:

$$\tilde{B}_6(Y_6) = Y_6((\tilde{\Sigma}^\dagger U)^\top, (\tilde{\Sigma}^\dagger U)^\top, (\tilde{\Sigma}^\dagger U)^\top).$$

Before stating the main lemma, we update the definition of symmetric off-diagonal subspace.

Definition 3.4. *Let the columns of matrix $P \in \mathbb{R}^{n_2^3 \times d_3}$ form a basis of a subspace \mathcal{P} . Index the rows of P by triples $(i_1, i_2, i_3) \in [n_2] \times [n_2] \times [n_2]$. The matrix P and the subspace \mathcal{P} are called symmetric off-diagonal if: whenever i_1, i_2, i_3 are not distinct the corresponding row is 0 (“off-diagonal”); and for any permutation π over $\{1, 2, 3\}$, the rows corresponding to (i_1, i_2, i_3) and $(i_{\pi(1)}, i_{\pi(2)}, i_{\pi(3)})$ are identical (“symmetric”).*

It is easy to verify that since the moments in \overline{M}_6 all have indices corresponding to distinct variables, the projection \mathcal{F}_6 is indeed symmetric off-diagonal. The constraints in this definition is closely related to the decoupling Theorem 3.20 of Gaussian chaoses.

Similarly, we define the “unique” columns in the 3-way Kronecker product to be the matrix $(E \otimes_{kr} E \otimes_{kr} E)_{\text{uniq}} \in \mathbb{R}^{n_2^3 \times k_3}$ whose columns are labeled by triples $(i_1, i_2, i_3) : 1 \leq i_1 \leq i_2 \leq i_3 \leq k$, and $(E \otimes_{kr} E \otimes_{kr} E)_{\text{uniq}}[:,(i_1, i_2, i_3)] = E_{[:,i_1]} \odot E_{[:,i_2]} \odot E_{[:,i_3]}$.

Lemma 3.19. *Let $E \in \mathbb{R}^{n_2 \times k}$ be a Gaussian random matrix. Let $P \in \mathbb{R}^{n_2^3 \times d_3}$ be a symmetric off-diagonal subspace of dimension $d_3 \geq \Omega(n_2^3)$. For any constant $C > 0$, if $n_2 \geq k^{2+C}$, with high probability $\sigma_{\min}(P^\top (E \otimes_{kr} E \otimes_{kr} E)_{\text{uniq}}) \geq \Omega(n_2^{3/2})$.*

The proofs of Theorem 3.14 are based on the above two lemmas. The proof of

Lemma 3.18 is essentially the same as Lemma 3.16. The proof of Lemma 3.19 is very similar to that of Lemma 3.17, and we highlight the only different case below:

Proof. (of Lemma 3.19)

As before we try to prove that the columns of $P^\top(E \otimes_{kr} E \otimes_{kr} E)_{uniq}$ are incoherent.

Recall we needed the following two claims:

Claim 3.15. *For any $1 \leq i_1 \leq i_2 \leq i_3 \leq k$, with high probability $\|P^\top(E_{[:,i_1]} \odot E_{[:,i_2]} \odot E_{[:,i_3]})\|^2 \geq \Omega(n_2^3)$.*

Claim 3.16. *For any $1 \leq i_1 \leq i_2 \leq i_3 \leq k$, with high probability*

$$\sum_{1 \leq i'_1 \leq i'_2 \leq i'_3, (i_1, i_2, i_3) \neq (i'_1, i'_2, i'_3)} \left| \langle P^\top(E_{[:,i_1]} \odot E_{[:,i_2]} \odot E_{[:,i_3]}), P^\top(E_{[:,i'_1]} \odot E_{[:,i'_2]} \odot E_{[:,i'_3]}) \rangle \right| \leq o(n_2^3).$$

The first claim can still be proved by the projection Lemma 3.33, except the vector $E_{[:,i_1]} \odot E_{[:,i_2]} \odot E_{[:,i_3]}$ is now $O(n_2)$ -concentrated (the proof is an immediate generalization of Lemma 3.34).

The second claim can be proved using similar ideas, however there is one new case. We again separate the terms according to the number of i'_1, i'_2, i'_3 that do not appear in $\{i_1, i_2, i_3\}$.

Case 1: At least one of i'_1, i'_2, i'_3 does not appear in $\{i_1, i_2, i_3\}$. Suppose there are t of i'_1, i'_2, i'_3 that do not appear in $\{i_1, i_2, i_3\}$, similar to before we first sample $E_{i_1}, E_{i_2}, E_{i_3}$ and condition on the event that they all have norm at most $2\sqrt{n_2}$. The inner-product then becomes an order t Gaussian chaos with Frobenius norm $n_2^{6-t/2}$. By Theorem 3.20 and Theorem 3.19 we know with high probability all these terms are bounded by $n_2^{6-t/2+\epsilon}$ for any constant $\epsilon > 0$.

Case 2: All of i'_1, i'_2, i'_3 appear in $\{i_1, i_2, i_3\}$. In the previous proof (of Lemma 3.17), there was only one possibility and it reduces to Case 1. However for 6-th moment we have a new case: $i = i_1 = i_2 = i'_1 < i'_2 = i'_3 = i_3 = j$ (and the symmetric case $i_1 = i'_1 = i'_2 < i_2 = i_3 = i'_3$). For this we will treat $T = PP^\top$ as a 6-th order tensor with Frobenius norm at most $n_2^{3/2}$ (as a matrix it has spectral norm 1, and rank at most n_2^3). The tensor is applied to the vectors $E_{[:,i]}$ and $E_{[:,j]}$ as $T(E_{[:,i]}, E_{[:,i]}, E_{[:,j]}, E_{[:,i]}, E_{[:,j]}, E_{[:,j]})$.

First we sample $E_{[:,i]}$, by Lemma 3.36 we know with high probability what remains will be a 3-rd order tensor $T(E_{[:,i]}, E_{[:,i]}, I, E_{[:,i]}, I, I)$ with Frobenius norm bounded by $O(n_2^{2+\epsilon})$. Notice that here it is important that Lemma 3.36 can handle diagonal entries, because $E_{[:,i]}$ appears on the 1, 2, 4-th coordinate (instead of the first three). We then apply Lemma 3.36 again on $T(E_{[:,i]}, E_{[:,i]}, I, E_{[:,i]}, I, I)(E_{[:,j]}, E_{[:,j]}, E_{[:,j]})^9$, and conclude that with high probability the term is bounded by $O(n_2^{2.5+2\epsilon})$ which is still much smaller than n_2^3 .

Finally we take the sum over all terms and choose ϵ to be small enough (depending on C), then when $k^{2+C} \leq n_2$ the sum is a lower-order term. \square

Stability Bounds

For the two linear equation systems in (3.7), we can write them in canonical form with coefficient matrices \tilde{H}_4, \tilde{H}_6 and the unknown variable $\text{vec}(Y_4), \text{vec}(Y_6)$, corresponding to the k_2, k_3 distinct elements in symmetric Y_4, Y_6 , namely:

$$\tilde{H}_4 \text{vec}(Y_4) = \overline{M}_4 / \sqrt{3}, \quad \tilde{H}_6 \text{vec}(Y_6) = \overline{M}_6 / \sqrt{15}.$$

When $\widehat{M}_4, \widehat{M}_6$, the empirical moment estimations for $\widetilde{M}_4, \widetilde{M}_6$, are used throughout the algorithm, both the coefficient matrices \tilde{H}_4, \tilde{H}_6 and the constant terms $\overline{M}_4, \overline{M}_6$ are affected by the noise from empirical estimation. In practice, instead of solving systems of linear equations, we solve the least square problem:

$$\min_{Y_4 \in \mathbb{R}_{sym}^{k \times k}} \|\sqrt{3}\mathcal{F}_4(UY_4U^\top) - \overline{M}_4\|^2, \quad \min_{Y_6 \in \mathbb{R}_{sym}^{k \times k \times k}} \|\sqrt{15}\mathcal{F}_6Y_6(U^\top, U^\top, U^\top) - \overline{M}_6\|^2. \quad (3.31)$$

and the solution to the least square problems are given by: $\text{vec}(\widehat{Y}_4) = \widehat{H}_4^\dagger \overline{M}_4$ and $\text{vec}(\widehat{Y}_6) = \widehat{H}_6^\dagger \overline{M}_6$.

⁹The notation might be confusing here: $T(E_{[:,i]}, E_{[:,i]}, I, E_{[:,i]}, I, I)$ is a 3rd order tensor, and we are applying it to $E_{[:,j]}, E_{[:,j]}, E_{[:,j]}$. The whole expression is equal to $T(E_{[:,i]}, E_{[:,i]}, E_{[:,j]}, E_{[:,i]}, E_{[:,j]}, E_{[:,j]})$.

Lemma 3.20. *Given the empirical 4-th and 6-th order moments $\widehat{M}_4 = \widetilde{M}_4 + E_4$, $\widehat{M}_6 = \widetilde{M}_6 + E_6$, and suppose that the absolute value of entries in E_4 and E_6 are at most δ_1 . Let \widehat{U} be the output of Step 1 for the span of the covariance matrices, and suppose that $\|\widehat{U} - \widetilde{U}\| \leq \delta_2$. Suppose that $\delta_1 \leq \min\{\|\widetilde{M}_4\|_F/\sqrt{n_4}, \|\widetilde{M}_6\|_F/\sqrt{n_6}\}$, and $\delta_2 \leq \min\{1, \sigma_{k_2}(\widetilde{H}_4)/2, \sigma_{k_3}(\widetilde{H}_6)/2\}$. Then, conditioned on the high probability event that both $\sigma_{k_2}(\widetilde{H}_4), \sigma_{k_3}(\widetilde{H}_6)$ are bounded below, we have:*

$$\begin{aligned}\|\widehat{Y}_4 - \widetilde{Y}_4\|_F &\leq O\left(\left(\delta_1 + \frac{\delta_2}{\sigma_{k_2}(\widetilde{H}_4)^2}\right)\sqrt{n_4}\right). \\ \|\widehat{Y}_6 - \widetilde{Y}_6\|_F &\leq O\left(\left(\delta_1 + \frac{\delta_2}{\sigma_{k_3}(\widetilde{H}_6)^2}\right)\sqrt{n_6}\right).\end{aligned}$$

Proof. We write the proof for \widehat{Y}_4 , the proof for \widehat{Y}_6 is exactly the same except changing the subscripts.

Recall that the coefficient matrix \widetilde{H}_4 corresponds to the composition of two linear mappings $\mathcal{F}_4(UY_4U^\top)$ on the variable Y_4 . Since we have showed that \mathcal{F}_4 is a projection determined by the Isserlis' Theorem and independent of the empirical estimation of the moments, we can bound the perturbation on the coefficient matrices by:

$$\|\widehat{H}_4 - \widetilde{H}_4\| \leq \|\widehat{U} \odot^2 - \widetilde{U} \odot^2\| \leq 2\|\widehat{U} - \widetilde{U}\|\|\widetilde{U}\| + \|\widehat{U} - \widetilde{U}\|_2^2 \leq 3\delta_2 \leq \|\widetilde{H}_4\|.$$

Similarly, we have $\|\widehat{H}_6 - \widetilde{H}_6\| \leq \|\widehat{U} \odot^3 - \widetilde{U} \odot^3\| \leq 7\delta_2 \leq \|\widetilde{H}_6\|$.

Therefore we can analyze the stability of the solution to the least square problems in (3.31) as follows:

$$\begin{aligned}\|\text{vec}(\widehat{Y}_4) - \text{vec}(\widetilde{Y}_4)\| &= \left\| \widehat{H}_4^\dagger \widehat{M}_4 - \widetilde{H}_4^\dagger \widetilde{M}_4 \right\| \\ &\leq O(\|\widehat{H}_4^\dagger\| \|\widehat{M}_4 - \widetilde{M}_4\| + \|\widehat{H}_4^\dagger - \widetilde{H}_4^\dagger\| \|\widetilde{M}_4\|) \\ &\leq O(\|\widehat{M}_4 - \widetilde{M}_4\| + \|\widehat{H}_4^\dagger - \widetilde{H}_4^\dagger\| \sqrt{n_4}) \\ &\leq O\left(\sqrt{n_4}(\delta_1 + \|\widehat{H}_4^\dagger\| \|\widetilde{H}_4^\dagger\| \delta_2)\right) \\ &\leq O\left(\sqrt{n_4}\left(\delta_1 + \frac{1}{\sigma_{k_2}(\widetilde{H}_4)^2} \delta_2\right)\right),\end{aligned}$$

where the first inequality is by applying Lemma 1.6 and note that $\|(\widetilde{M}_4 - \widehat{M}_4)\|_F \leq \delta_1 \sqrt{n_4} \leq \|\widehat{M}_4\|_F$, the second inequality is because $\|\widehat{M}_4\|_F \leq O(\sqrt{n_4})$, the third inequality is by applying the perturbation bound of pseudo-inverse in Theorem 1.5, the fourth inequality is by the assumption that δ_2 is sufficiently small compared to the smallest singular value of \widehat{H}_4 thus $\sigma_{k_2}(\widehat{H}_4) = O(\sigma_{k_2}(\widetilde{H}_4))$.

□

3.4.3 Step 3 of Zero-Mean Case: Tensor Decomposition

Algorithm 9: TensorDecomp

Input: the span of covariance matrices $U \in \mathbb{R}^{n_2 \times k}$ (vectorized with distinct entries), the unfolded 4-th and 6-th moments $Y_4 \in \mathbb{R}^{k \times k}$ and $Y_6 \in \mathbb{R}^{k \times k \times k}$ in the coordinate system of U .

Output: Parameters $\mathcal{G} = \{(\omega_i, \Sigma^{(i)}) : i \in [k]\}$.

Compute the SVD of Y_4 : $Y_4 = V_2 \Lambda_2 V_2^\top$.

Let $G = Y_6 (V_2 \Lambda_2^{-1/2}, V_2 \Lambda_2^{-1/2}, V_2 \Lambda_2^{-1/2})$

Find the (unique) first k orthogonal eigenvectors v_i and the corresponding eigenvalues λ_i of G , denoted by $\{(v_i, \lambda_i) : i \in [k]\}$

For all $i \in [k]$, let $\text{vec}(\Sigma^{(i)}) = \lambda_i U V_2 \Lambda_2^{1/2} v_i$, let $\omega_i = (\lambda_i)^{-2}$.

Return: $\mathcal{G} = \{(\omega_i, \Sigma^{(i)}) : i \in [k]\}$.

Given the estimations of the unfolded moments Y_4 and Y_6 from Step 2, and given the span of covariance matrices U from Step 1, Step 3 use tensor decomposition to robustly find the parameters of the mixture of zero-mean Gaussians.

Recall that in the coordinate system with basis U , the covariance matrices (vectorized with distinct entries) are given by $\widetilde{\Sigma}^{(i)} = \widetilde{U} \widetilde{\sigma}^{(i)}$ for all i . The unfolded moments in the same coordinate system are:

$$\widetilde{Y}_4 = \sum_{i=1}^k \widetilde{\omega}_i \widetilde{\sigma}^{(i)} \otimes^2, \quad \widetilde{Y}_6 = \sum_{i=1}^k \widetilde{\omega}_i \widetilde{\sigma}^{(i)} \otimes^3.$$

We will apply tensor decomposition algorithm to find the $\widetilde{\sigma}^{(i)}$'s. We restate the theorem for orthogonal symmetric tensor decomposition in Anandkumar et al. [7] below:

Theorem 3.15 (Theorem 5.1 in [7]). *Consider k orthonormal vector $v_1, \dots, v_k \in \mathbb{R}^n$'s and k positive weights $\lambda_1, \dots, \lambda_k$. Define the tensor $T = \sum_{i=1}^k \lambda_i v_i \otimes^3$. Given $\widehat{T} = T + E$ and assume that $\|E\| \leq C_1 \min\{\lambda_i\}/k$, then there is an algorithm that finds λ_i 's and v_i 's in polynomial running time with the following guarantee: with probability at least $1 - e^{-n}$, for some permutation π over $[k]$ and for all $i \in [k]$, we have:*

$$\|v_i - \widehat{v}_i\| \leq O(\|E\|/\lambda_i), \quad |\lambda_i - \widehat{\lambda}_i| \leq O(\|E\|).$$

In order to reduce our problem to the orthogonal tensor decomposition so that the tensor power method (Algorithm 1; page 21 in [7]) can be applied, we use the same “whitening” technique as in [7]. We first compute the SVD of the unfolded 4-th moments $\widetilde{Y}_4 = \widetilde{V}_2 \widetilde{\Lambda}_2 \widetilde{V}_2^T$, then use the singular vectors to transform the unfolded 6-th moments Y_6 into an orthogonal symmetric tensor $\widetilde{Y}_6(\widetilde{V}_2 \widetilde{\Lambda}_2^{-1/2}, \widetilde{V}_2 \widetilde{\Lambda}_2^{-1/2}, \widetilde{V}_2 \widetilde{\Lambda}_2^{-1/2})$.

Next we complete the stability analysis for the two-step procedure, i.e. whitening and orthogonal tensor decomposition, which was not analyzed in [7].

Theorem 3.16. *Consider k linearly independent vectors $a_1, \dots, a_k \in \mathbb{R}^n$, and k positive weights $\omega_1, \dots, \omega_k$. Define $G_2 = \sum_{i=1}^k \omega_i a_i \otimes a_i \in \mathbb{R}_{sym}^{n \times n}$ and $G_3 = \sum_{i=1}^k \omega_i a_i \otimes a_i \otimes a_i \in \mathbb{R}_{sym}^{n \times n \times n}$. Let $\gamma_{min} = \min\{\sigma_{min}(G_2), 1\}$, $\gamma_{max} = \sigma_{max}(G_2)$, and let $\omega_o = \min\{\omega_i\}$. Given $\widehat{G}_2, \widehat{G}_3$ and assume that:*

$$\|\widehat{G}_2 - G_2\|_F \leq \delta_2 \leq o\left(\frac{\gamma_{min}^{2.5}}{k\|G_3\|}\right), \quad \|\widehat{G}_3 - G_3\|_F \leq \delta_3 \leq o\left(\frac{\gamma_{min}^{1.5}}{k}\right).$$

There exists an algorithm that finds \widehat{a}_i and $\widehat{\omega}_i$ in polynomial (in all the variables $(n, k, 1/\sigma_{min}(G_2))$) running time with the following guarantee: with probability at least $1 - e^{-n}$, for some permutation π over $[k]$ and for all $i \in [k]$ we have:

$$\begin{aligned} \|\widehat{a}_{\pi(i)} - a_{\pi(i)}\| &\leq \text{poly}(\|G_3\|, 1/\sigma_{min}(G_2), 1/\omega_o)\delta_2 + \text{poly}(\|G_3\|, 1/\sigma_{min}(G_2), 1/\omega_o)\delta_3, \\ \|\widehat{\omega}_i - \omega_i\| &\leq \text{poly}(\|G_3\|, 1/\sigma_{min}(G_2))\delta_2 + \text{poly}(\|G_3\|, 1/\sigma_{min}(G_2))\delta_3. \end{aligned}$$

Proof. (to Theorem 3.16)

1. *Algorithm*

We first apply the whitening technique in [7]: Let $\widehat{G}_2 = \widehat{V}_2 \widehat{\Lambda}_2 \widehat{V}_2^\top$ be the singular value decomposition of \widehat{G}_2 , and note that the matrix $\widehat{V}_2 \widehat{\Lambda}_2^{-1/2}$ whitens G_2 in the sense that $\widehat{G}_2 (\widehat{V}_2 \widehat{\Lambda}_2^{-1/2}, \widehat{V}_2 \widehat{\Lambda}_2^{-1/2}) = I_n$. Similarly we can whiten \widehat{G}_3 with the matrix $\widehat{V}_2 \widehat{\Lambda}_2^{-1/2}$ and obtain the following symmetric 3-rd order tensor $\widehat{G} \in \mathbb{R}_{sym}^{k \times k \times k}$:

$$\widehat{G} = \widehat{G}_3 (\widehat{V}_2 \widehat{\Lambda}_2^{-1/2}, \widehat{V}_2 \widehat{\Lambda}_2^{-1/2}, \widehat{V}_2 \widehat{\Lambda}_2^{-1/2}).$$

Note that in the exact case with G_2 and G_3 , we have that:

$$G = \sum_{i=1}^k \lambda_i v_i \otimes^3,$$

where $\lambda_i = \omega_i^{-1/2}$, and the vectors $v_i = \lambda_i^{-1} V_2^\top \Lambda_2^{-1/2} a_i$ and they are orthonormal. Also note that $\lambda_{min} \geq 1$ and $\lambda_{max} \leq \omega_o^{-1/2}$. We can then apply *orthogonal tensor decomposition* (Algorithm 1 in [7]) to \widehat{G} to robustly obtain estimations of v_i 's and λ_i 's. After obtaining the estimation \widehat{v}_i and $\widehat{\lambda}_i$'s, we can further obtain the estimation of a_i 's and ω_i 's as:

$$\widehat{a}_i = \widehat{V}_2 \widehat{\Lambda}_2^{1/2} \widehat{v}_i \widehat{\lambda}_i, \quad \widehat{\omega}_i = (\widehat{\lambda}_i)^{-2} \tag{3.32}$$

2. *Stability analysis*

The estimation of the vectors and weights are given in (3.32). In order to bound the distance $\|\widehat{a}_i - a_i\|$ and $\|\widehat{\omega}_i - \omega_i\|$, we show the stability of the estimation \widehat{V}_2 , $\widehat{\Lambda}_2$, and \widehat{v}_i , $\widehat{\lambda}_i$ separately.

First, note that by assumption $\|\widehat{G}_2 - G_2\|_F \leq \delta_2$, we can apply Lemma 1.2 and Lemma 1.3 to bound the singular values and the singular vectors of \widehat{G}_2 by:

$$\|\widehat{V}_2 - V_2\| \leq \sqrt{2} \delta_2 / \gamma_{min}, \quad \|\widehat{\Lambda}_2 - \Lambda_2\| \leq \delta_2.$$

Define $X = V_2 \Lambda_2^{-1/2}$ and define $\Delta_X = \widehat{X} - X$. By the assumption that $\delta_2 \leq o(\gamma_{min})$,

we have $\|\widehat{V}_2 - V_2\| \leq 1$ and $\|\widehat{\Lambda}_2^{-1/2} - \Lambda_2^{-1/2}\| \leq \|\Lambda_2^{-1/2}\| \leq \gamma_{min}^{-1/2}$. Therefore we can apply Lemma 1.6 to bound $\|\Delta_X\|$:

$$\begin{aligned} \|\Delta_X\| &\leq O(\|\widehat{V}_2 - V_2\| \|\Lambda_2^{-1/2}\| + \|V_2\| \|\widehat{\Lambda}_2^{-1/2} - \Lambda_2^{-1/2}\|) \\ &\leq O\left(\frac{\delta_2}{\gamma_{min}^1} \gamma_{min}^{-1/2} + (\gamma_{min}^{-1/2})^2 \delta_2\right) \\ &\leq O(\delta_2 / \gamma_{min}^{1.5}). \end{aligned}$$

Moreover, since $\delta_2 \leq o(\gamma_{min})$, we also have $\|\Delta_X\| \leq \|X\| = \gamma_{min}^{-0.5}$.

Next, we bound the distance $\|\widehat{G} - G\|$. Recall that $\widehat{G} = \widehat{G}_3(\widehat{X}, \widehat{X}, \widehat{X})$. Using the fact that tensor is a multi-linear operator, and by the assumption that $\|\widehat{G}_3 - G_3\| \leq \delta_3$, we have:

$$\begin{aligned} \epsilon \equiv \|\widehat{G} - G\| &\leq \|\widehat{G}_3(\widehat{X}, \widehat{X}, \widehat{X}) - G_3(X, X, X)\|_F \\ &\leq \|G_3(\widehat{X}, \widehat{X}, \widehat{X}) - G_3(X, X, X)\| + \|\widehat{G}_3(\widehat{X}, \widehat{X}, \widehat{X}) - G_3(\widehat{X}, \widehat{X}, \widehat{X})\| \\ &\leq 3\|G_3(\Delta_X, X, X)\| + 3\|G_3(\Delta_X, \Delta_X, X)\| + \|G_3(\Delta_X, \Delta_X, \Delta_X)\| + \delta_3 \|\widehat{X}\|^3 \\ &\leq 7\|G_3\| \|X\|^2 \|\Delta_X\| + (\|X\| + \|\Delta_X\|)^3 \delta_3 \\ &\leq O\left(\frac{\|G_3\|}{\gamma_{min}^{2.5}} \delta_2 + \frac{1}{\gamma_{min}^{1.5}} \delta_3\right). \end{aligned}$$

Note that by the assumption $\delta_2 \leq o(\frac{\gamma_{min}^{2.5}}{k\|G_3\|})$, $\delta_3 \leq o(\frac{\gamma_{min}^{1.5}}{k})$, we have $\epsilon \leq o(\frac{1}{k})$. Therefore we can apply Theorem 3.16 to conclude that with probability at least $1 - e^{-n}$ (over the randomness of the randomized algorithm itself), the tensor power algorithm runs in time $\text{poly}(n, k, 1/\lambda_{min})$ and for some permutation π over $[k]$ it returns:

$$\|\widehat{v}_{\pi(i)} - v_{\pi(i)}\| \leq \frac{8\epsilon}{\lambda_{min}}, \quad |\widehat{\lambda}_i - \lambda_i| \leq 5\epsilon, \quad \forall j \in [k].$$

Finally, since we also have $5\epsilon \leq 1/2 \leq \lambda_{min}/2$ we can bound the estimation error

of \hat{a}_i and $\hat{\omega}_i$ as defined in (3.32) by:

$$\begin{aligned} \|\hat{a}_{\pi(i)} - a_i\| &\leq 3(\|\Delta_X\| \lambda_{max} + \frac{1}{\gamma_{min}^{0.5}} \frac{8\epsilon}{\lambda_{min}} \lambda_{max} + \frac{1}{\gamma_{min}^{0.5}} 5\epsilon) \\ &\leq \text{poly}(\|G_3\|, 1/\sigma_{min}(G_2), 1/\omega_o) \delta_2 + \text{poly}(\|G_3\|, 1/\sigma_{min}(G_2), 1/\omega_o) \delta_3, \\ \|\hat{\omega}_i - \omega_i\| &\leq \text{poly}(\|G_3\|, 1/\sigma_{min}(G_2)) \delta_2 + \text{poly}(\|G_3\|, 1/\sigma_{min}(G_2)) \delta_3. \end{aligned}$$

□

Now we can apply Theorem 3.16 to our case.

Lemma 3.21. *Given $\hat{Y}_4, \hat{Y}_6, \hat{U}$ and suppose that $\|\hat{Y}_4 - \tilde{Y}_4\|_F, \|\hat{Y}_6 - \tilde{Y}_6\|_F$ as well as $\|\hat{U} - \tilde{U}\|$ are bounded by some inverse $\text{poly}(n, k, 1/\omega_o, 1/\rho)\delta$. There exists an algorithm that with high probability, returns $\hat{\Sigma}^{(i)}$'s and $\hat{\omega}_i$'s such that for some permutation π over $[k]$, we have the distance $\|\hat{\Sigma}^{(i)} - \tilde{\Sigma}^{(i)}\|$ and $\|\hat{\omega}_i - \tilde{\omega}_i\|$ are bounded by δ . Moreover, the running time of the algorithm is upperbounded by $\text{poly}(n, k, 1/\omega_o, 1/\rho)$.*

Proof. (to Lemma 3.21)

We apply Theorem 3.16, and pick $G_2 = \tilde{Y}_4, G_3 = \tilde{Y}_6$. We only need to verify that $\|\tilde{Y}_6\|$ and $1/\sigma_{min}(\tilde{Y}_4)$ are polynomials of the relevant parameters. This is easy to see, since $\sigma_{min}(\tilde{Y}_4) \geq \omega_o \sigma_{min}(\tilde{\Sigma})^2$, and the matrix $\tilde{\Sigma}$ is a perturbed rectangular matrix which by Lemma 3.31 has $\sigma_{min}(\tilde{\Sigma}) \geq \Omega(\rho\sqrt{n_2})$ with high probability.

Finally, given $\hat{\sigma}^{(i)}$, and given the output of Step 2, i.e. \hat{U} , with inverse polynomial accuracy, we can recover $\hat{\Sigma}^{(i)} = \hat{U}\hat{\sigma}^{(i)}$ up to accuracy polynomial in the relevant parameters. □

3.4.4 Proof of Theorem 3.5

The results in all previous sections showed the correctness and robustness of each individual step for the algorithm for zero-mean case, In this section, we summarize those results to prove that the overall algorithm has polynomial time/sample complexity.

Lemma 3.22 (Concentration of empirical moments). *Given N samples x_1, \dots, x_N drawn i.i.d. from the n -dimensional mixture of k Gaussians, if $N \geq n^7/\delta^2$, then with*

Algorithm 10: MainAlgorithm (Zero-mean case)

Input: Samples x_i from the mixture of Gaussians, number of components k .

Output: Set of parameters $\mathcal{G} = \{(\omega_i, \Sigma^{(i)}) : i \in [k]\}$.

Estimate M_4, M_6 using the samples.

$$M_4 = \frac{1}{N} \sum_{i=1}^N x_i \otimes^4, \quad M_6 = \frac{1}{N} \sum_{i=1}^N x_i \otimes^6.$$

Let $s = 9\lceil\sqrt{n}\rceil$

(Step 1 (a) Algorithm 5)

$S_1 = \text{FindColumnSpan}(M_4, \{1, \dots, s\})$,

$S_2 = \text{FindColumnSpan}(M_4, \{s+1, \dots, 2s\})$.

(Step 1 (b) Algorithm 6)

$U_1 = \text{FindProjectedSigmaSpan}(M_4, \{1, \dots, s\}, S_1)$,

$U_2 = \text{FindProjectedSigmaSpan}(M_4, \{s+1, \dots, 2s\}, S_2)$.

(Step 1 (c) Algorithm 7)

$U = \text{MergeProjections}(S_1, U_1, S_2, U_2)$.

(Step 2 Algorithm 8)

$(Y_4, Y_6) = \text{Estimate}Y_4Y_6(M_4, M_6, U)$.

(Step 3 Algorithm 9)

$\mathcal{G} = \text{TensorDecomp}(Y_4, Y_6, U)$

Return: \mathcal{G} .

high probability, we have that for all $j_1, \dots, j_6 \in [n]$:

$$\left| [\widehat{M}_4]_{j_1, j_3, j_3, j_4} - [\widetilde{M}_4]_{j_1, j_3, j_3, j_4} \right| \leq \delta, \quad \left| [\widehat{M}_6]_{j_1, j_3, j_3, j_4, j_5, j_6} - [\widetilde{M}_6]_{j_1, j_3, j_3, j_4, j_5, j_6} \right| \leq \delta.$$

Proof. Let x denote the random vector of this mixture of Gaussians. We first truncate its tail probabilities to make all the entries ($[x]_j$ for $j \in [n]$) in the vector x be in the range $[-\sqrt{n}, \sqrt{n}]$. Apply union bound, we know that with high probability (at least $1 - O(e^{-n})$), for all indices $j_1, \dots, j_6 \in [n]$, we have $\left| [x]_{j_1} \dots [x]_{j_6} \right| \leq n^3$. Then we can apply Hoeffding's inequality to bound the empirical moments by:

$$\Pr \left[\left| \widehat{\mathbb{E}}[x_{j_1} \dots x_{j_6}] - \mathbb{E}[x_{j_1} \dots x_{j_6}] \right| \geq \delta \right] \leq \exp\left(-\frac{2\delta^2 N^2}{N(2n^3)^2}\right) + O(e^{-n}) \leq O(e^{-n}).$$

□

Proof. (of Theorem 3.5)

We show that, to achieve ϵ accuracy in the output of Step 3 in the algorithm for the zero-mean case, the number of samples we need to estimate the moments M_4 and M_6 is bounded by a polynomial of relevant parameters, namely $\text{poly}(n, k, 1/\omega_o, 1/\epsilon, 1/\rho)$, and each step of the algorithm can be done in polynomial time.

We backtrack the input-output relations from Step 3 to Step 2 and to Step 1, and we show that the estimation error in the empirical moments and the inputs / outputs only *polynomially* propagate throughout the steps.

First note that we have shown that every steps fails with negligible probability ($O(e^{-n^C})$ for any absolute constant C). Then apply union bound, we have that the entire algorithm works correctly with high probability.

1. By Lemma 3.21, in order to achieve ϵ accuracy in the final estimation of the mixing weights and the covariance matrices, we need to drive the input accuracy of Step 3 (also the output accuracy of Step 2) to be bounded by some inverse polynomial in $(n, 1/\epsilon, 1/\rho, 1/\omega_o)$. Also recall that this step has running time $\text{poly}(n, k, 1/\rho, 1/\omega_o)$.
2. Theorem 3.12 and Theorem 3.14 guarantee that with smoothed analysis $\sigma_{\min}(\tilde{H}_4)$ and $\sigma_{\min}(\tilde{H}_6)$ are lower bounded polynomially. Then by Lemma 3.20, in order to have the output accuracy of Step 2 be bounded by inverse $\text{poly}(n, 1/\epsilon, 1/\rho, 1/\omega_o)$, we need to drive the input accuracy of Step 2 (\hat{U}, \hat{M}_4) to be bounded by some other inverse polynomial. Step 2 involves solving linear systems of dimension $n_4 k_2$ and $n_6 k_3$, thus its running time is polynomial.
3. Lemma 3.13 and 3.14 guarantees that with smoothed analysis $\sigma_k(\tilde{Q}_U)$ is lower bounded polynomially. Then by Lemma 3.15, in order to have the output accuracy of Step 1 (c) (\hat{U}) be bounded by inverse polynomial, we need to drive the input accuracy (output \hat{S}_i of Step 1 (a) and output \hat{U}_i of Step 1 (b)) to be bounded by some other inverse polynomial. Step 1 (c) involves multiplications and factorization of matrices of polynomial size, and thus the running time is also polynomial.

4. Lemma 3.10 guarantees that with smoothed analysis $\sigma_k(\tilde{Q}_{U_S})$ is lower bounded polynomially. Then by Lemma 3.11, in order to have the output accuracy of Step 1 (b) (\widehat{U}_S) be bounded by inverse polynomial, we need to drive the input accuracy (output \widehat{S}_i of Step 1 (a)) to be bounded by some other inverse polynomial. Step 1 (b) involves multiplications and factorization of matrices of polynomial size, and thus the running time is also polynomial.
5. Lemma 3.8 guarantees that with smoothed analysis $\sigma_k(\tilde{Q}_S)$ is lower bounded by inverse polynomial. Then by Lemma 3.9, in order to have the output accuracy of Step 1 (a) (\widehat{S}) be bounded by inverse polynomial, we need to drive the input accuracy (the moment estimation \widehat{M}_4) to be bounded by some other inverse polynomial. Step 1 (a) involves multiplications and factorization of matrices of polynomial size, and thus the running time is also polynomial.
6. Finally, by Lemma 3.22, in order to have the accuracy of moment estimation ($\widehat{M}_4, \widehat{M}_6$) be bounded by inverse polynomial, we need the number of samples N polynomial in all the relevant parameters, including k .

□

3.4.5 Proofs for the General Case

In this section, we present the algorithm for learning mixture of Gaussians with general means. The algorithm generalizes the insights obtained from the algorithm for the zero-mean case. The steps are very similar, and we will highlight the differences.

Step 1. Span finding In this step, we find the following two subspaces:

$$\tilde{Z} = \text{span}\{\tilde{\mu}^{(i)} : i \in [k]\}, \quad \tilde{\Sigma}_o = \text{span}\{\text{Proj}_{\tilde{Z}^\perp} \tilde{\Sigma}^{(i)} \text{Proj}_{\tilde{Z}^\perp}\}.$$

This is very similar to Step 1 in the algorithm for the zero-mean case, and can be achieved in three small steps:

Algorithm 11: MainAlgorithm (General Case)

Input: Samples $\{x_i \in \mathbb{R}^n : i = 1, \dots, N\}$ from the mixture of Gaussians, number of components k .

Output: Set of parameters $\mathcal{G} = \{(\omega_i, \mu^{(i)}, \Sigma^{(i)}) : i \in [k]\}$.

Estimate M_3, M_4, M_6 using the samples

$$M_3 = \frac{1}{N} \sum_{i=1}^N x_i \otimes^3, \quad M_4 = \frac{1}{N} \sum_{i=1}^N x_i \otimes^4, \quad M_6 = \frac{1}{N} \sum_{i=1}^N x_i \otimes^6$$

- Step 1. (a) This can be accomplished similar to Algorithm 5 FindColumnSpan
Let $\mathcal{H}_1 = \{1, \dots, 12\sqrt{n}\}$, find $S_1 = \text{span}\{\tilde{\mu}^{(i)}, \tilde{\Sigma}_{[:,j]}^{(i)} : i \in [k], j \in \mathcal{H}_1\}$.
Let $\mathcal{H}_2 = \{12\sqrt{n} + 1, \dots, 24\sqrt{n}\}$, find $S_2 = \text{span}\{\tilde{\mu}^{(i)}, \tilde{\Sigma}_{[:,j]}^{(i)} : i \in [k], j \in \mathcal{H}_2\}$.
- (b) This can be accomplished similar to Algorithm 6 FindProjectedSigmaSpan
Find $U_1 = \text{span}\{\text{Proj}_{S_1^\perp} \tilde{\Sigma}^{(i)} : i \in [k]\}$.
Find $U_2 = \text{span}\{\text{Proj}_{S_2^\perp} \tilde{\Sigma}^{(i)} : i \in [k]\}$.
- (c) This can be accomplished similar to Algorithm 7 MergeProjections
Merge U_1 and U_2 to get $Z = \text{span}\{\mu^{(i)} : i \in [k]\}$,
 $U' = \text{span}\{\text{vec}(\text{Proj}_{Z^\perp} \Sigma^{(i)}) : i \in [k]\}$, and $U_o = \text{span}\{\text{Proj}_{Z^\perp} \Sigma^{(i)} \text{Proj}_{Z^\perp} : i \in [k]\}$.

Step 2. Project the samples to the subspace Z^\perp : $\text{Proj}_{Z^\perp} x = \{\text{Proj}_{Z^\perp} x_1, \dots, \text{Proj}_{Z^\perp} x_N\}$.

Apply the algorithm for zero mean case to the projected samples, let $\mathcal{G}_o = \{(\omega_i, \text{Proj}_{Z^\perp} \Sigma^{(i)} \text{Proj}_{Z^\perp}) : i \in [k]\} = \text{MainAlgorithm (Zero-mean case)}(\text{Proj}_{Z^\perp} x)$.

Step 3. Let $T = [\text{vec}(\text{Proj}_{Z^\perp} \Sigma^{(i)} \text{Proj}_{Z^\perp}) : i \in [k]]^{\dagger\top} \in \mathbb{R}^{n^2 \times k}$, and let $T^{(i)}$ for $i \in [k]$ denote the columns of T .

Let $M_{3(1)} \in \mathbb{R}^{n \times n^2}$ be the matricization of M_3 along the first dimension.

Let $\mu^{(i)} = M_{3(1)} T^{(i)} / \omega_i$ for $i \in [k]$ and let $\mu = [\mu^{(i)} : i \in [k]]$.

Step 4. Let $M'_4 = M_4 + 2 \sum_{i=1}^k \omega_i \mu^{(i)} \otimes^4$.

Find the span $S = \text{span}\{\text{vec}(\tilde{\Sigma}^{(i)}) + \tilde{\mu}^{(i)} \odot \tilde{\mu}^{(i)} : i \in [k]\}$.

This can be achieved by treating M'_4 as the 4-th moments of a mixture of zero-mean Gaussians, and apply Step 1 in the algorithm for zero-mean case to find the span of the covariance matrices, and let S denote the result.

Let $\Sigma = [\text{vec}(\Sigma^{(i)}) : i \in [k]] = (\text{Proj}_S U' - \mu \odot \mu)$.

Return: $\mathcal{G} = \{(\omega_i, \mu^{(i)}, \Sigma^{(i)}) : i \in [k]\}$.

1. Step 1 (a). For a subset \mathcal{H} of size $12\sqrt{n}$, find the span \mathcal{S} of the mean vectors and a subset of columns of the covariance matrices:

$$\mathcal{S} = \text{span}\{\tilde{\mu}^{(i)}, \tilde{\Sigma}_{[:,j]}^{(i)} : i \in [k], j \in \mathcal{H}\}.$$

2. Step 1 (b). Find the span of covariance matrices projected to the subspace S^\perp :

$$U_S = \text{span}\{\text{Proj}_{S^\perp} \tilde{\Sigma}^{(i)} : i \in [k]\}.$$

3. Step 1 (c). Run 1(a) and 1(b) on two disjoint subsets \mathcal{H}_1 and \mathcal{H}_2 . Merge the two spans U_1 and U_2 to get \tilde{Z} and $\text{span}\{\text{Proj}_{\tilde{Z}^\perp} \tilde{\Sigma}^{(i)} : i \in [k]\}$.

Next, we discuss each small step and compare it with the similar analysis of the algorithm for the zero-mean case.

Step 1 (a). Find the span \mathcal{S} of the means and a subset of the columns of the covariance matrices Similar to Step 1 (a) for the zero-mean case, in this step we want to find a subspace \mathcal{S} which contains the span of a subset of columns of $\tilde{\Sigma}^{(i)}$'s. However, with the mean vector $\tilde{\mu}^{(i)}$'s appearing in the moments, the subspace we find also contains the span of all the mean vectors. In particular, for a subset $\mathcal{H} \in [n]$ with $|\mathcal{H}| = \sqrt{n}$, we aim to find the following subspace:

$$\mathcal{S} = \text{span}\{\tilde{\mu}^{(i)}, \tilde{\Sigma}_{[:,j]}^{(i)} : i \in [k], j \in \mathcal{H}\}. \quad (3.33)$$

Similar to Claim 3.1 for the zero-mean case, the key observation for finding the subspace is the structure of the one-dimensional slices of the 4-th order moments for the general case:

Claim 3.17. *For any indices $j_1, j_2, j_3 \in [n]$, the one-dimensional slices of \tilde{M}_4 are*

given by:

$$\begin{aligned} & \widetilde{M}_4(e_{j_1}, e_{j_2}, e_{j_3}, I) \tag{3.34} \\ &= \sum_{i=1}^n \widetilde{\omega}_i \left(\widetilde{\mu}_{j_1}^{(i)} \widetilde{\mu}_{j_2}^{(i)} \widetilde{\mu}_{j_3}^{(i)} \widetilde{\mu}^{(i)} + \sum_{\pi \in \left\{ \begin{array}{l} (j_1, j_2, j_3), \\ (j_2, j_3, j_1), \\ (j_3, j_1, j_2) \end{array} \right\}} \widetilde{\Sigma}_{\pi_1, \pi_2}^{(i)} \widetilde{\Sigma}_{[:, \pi_3]}^{(i)} + \widetilde{\mu}_{\pi_1}^{(i)} \widetilde{\mu}_{\pi_2}^{(i)} \widetilde{\Sigma}_{[:, \pi_3]}^{(i)} + \widetilde{\Sigma}_{\pi_1, \pi_2}^{(i)} \widetilde{\mu}_{\pi_3}^{(i)} \widetilde{\mu}^{(i)} \right) \end{aligned}$$

Note that if we pick the indices $j_1, j_2, j_3 \in \mathcal{H}$, all such one-dimensional slice of \widetilde{M}_4 lie in the subspace \mathcal{S} . We again evenly partition the set \mathcal{H} into three disjoint subset $\mathcal{H}^{(i)}$ and take $j_i \in \mathcal{H}^{(i)}$ for $i = 1, 2, 3$. Define the matrix $\widetilde{Q}_S \in \mathbb{R}^{n \times (|\mathcal{H}|/3)^3}$ as in (3.12) whose columns are the one-dimensional slices of \widetilde{M}_4 :

$$\widetilde{Q}_S = \left[\left[\widetilde{M}_4(e_{j_1}, e_{j_2}, e_{j_3}, I) : j_3 \in \mathcal{H}^{(3)} \right] : j_2 \in \mathcal{H}^{(2)} \right] : j_1 \in \mathcal{H}^{(1)} \right] \in \mathbb{R}^{n \times (|\mathcal{H}|/3)^3}. \tag{3.35}$$

The proof of this step is similar to the Lemmas 3.8 (for smoothed analysis) and 3.9 (for stability analysis). The main difference is that in the matrix \widetilde{B} defined in the structural Claim 3.4, there is now another block $\widetilde{B}^{(0)}$ with k columns that corresponds to the $\widetilde{\mu}^{(i)}$ directions, which we can again handle with Lemma 3.29.

Lemma 3.23 shows the deterministic conditions for Step 1 (a) to correctly identify the subspace \mathcal{S} from the columns of \widetilde{Q}_S , and uses smoothed analysis to show that the conditions hold with high probability.

Lemma 3.23 (Correctness). *Given \widetilde{M}_4 of a general mixture of Gaussians, for any subset $\mathcal{H} \in [n]$ and $|\mathcal{H}| = c_2 k$ with the constant $c_2 > 9$, let \widetilde{Q}_S be the matrix defined as in (3.35). The columns of \widetilde{Q}_S give the desired span \mathcal{S} defined in (3.33) if the matrix \widetilde{Q}_S achieves the maximal column rank $k + k|\mathcal{H}|$. With probability (over the ρ -perturbation) at least $1 - C\epsilon^{0.5n}$ for some constant C , the $k(1 + |\mathcal{H}|)$ -th singular value of \widetilde{Q}_S is bounded below by:*

$$\sigma_{k(1+|\mathcal{H}|)}(\widetilde{Q}_S) \geq \rho\epsilon\sqrt{n}.$$

The proof idea is similar to that of Lemma 3.8. We construct a basis $\widetilde{P}_S \in$

$\mathbb{R}^{n \times (k+k|\mathcal{H}|)}$ for the subspace \mathcal{S} as follows.

$$\begin{aligned} \tilde{P}_S &= \left[\left[\tilde{\mu}^{(i)} : i \in [k] \right], \left[\left[\tilde{\Sigma}_{[:,j]}^{(i)} : i \in [k] \right] : j \in \mathcal{H}^{(l)} \right] : l = 1, 2, 3 \right] \\ &= \left[\tilde{\mu}, \tilde{\Sigma}_{[:,\mathcal{H}^{(1)}}], \tilde{\Sigma}_{[:,\mathcal{H}^{(2)}}], \tilde{\Sigma}_{[:,\mathcal{H}^{(3)}}] \right]. \end{aligned} \quad (3.36)$$

Note that the dimension of the subspace \mathcal{S} is at most $k(|\mathcal{H}| + 1) < n/3$. Then we show by the Claim about the moment structure that the matrix \tilde{Q}_S can be written as a product of \tilde{P}_S and some coefficient matrix \tilde{B}_S . Then we bound the smallest singular value of the two matrices \tilde{P}_S and \tilde{B}_S via smoothed analysis separately. The coefficient matrix \tilde{B}_S is slightly different than that in the zero-mean case, but has similar block-diagonal structure properties.

The detailed proof is provided below.

Proof. (of Proposition 3.23)

Similar to structural property in Claim 3.4 for the zero-mean case, we can write the matrix \tilde{Q}_S in a product form:

$$\tilde{Q}_S = \tilde{P}_S (D_{\tilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|}) (\tilde{B}_S)^\top.$$

We will bound the smallest singular value for each of the factor, and apply union bound to conclude the lower bound of $\sigma_{k(1+|\mathcal{H}|)}(\tilde{Q}_S)$.

The matrix $\tilde{P}_S \in \mathbb{R}^{n \times (k+k|\mathcal{H}|)}$ is defined in (3.36). Restricting to the rows corresponding to $[n] \setminus \mathcal{H}$, we can use Lemma 3.32 to argue that $\sigma_{k(1+|\mathcal{H}|)} \geq \epsilon \rho \sqrt{n}$ with probability at least $1 - (C\epsilon)^{0.25n}$.

In order to lower bound $\sigma_{\min}(\tilde{B}_S)$, we first analyze the structure of this coefficient matrix. The matrix \tilde{B}_S has the following block structure:

$$\tilde{B}_S = \left[\tilde{B}^{(0)}, \tilde{B}^{(1)}, \tilde{B}^{(2)}, \tilde{B}^{(3)} \right].$$

The first block $\tilde{B}^{(0)} \in \mathbb{R}^{(|\mathcal{H}|/3)^3 \times k}$ is a summation of four matrices $\tilde{B}_i^{(0)}$ for $i = 0, 1, 2, 3$, where $\tilde{B}_0^{(0)} = \tilde{\mu}_{\mathcal{H}^{(3)}} \odot \tilde{\mu}_{\mathcal{H}^{(2)}} \odot \tilde{\mu}_{\mathcal{H}^{(1)}}$, and $\tilde{B}_1^{(0)} = \tilde{\Sigma}_{\mathcal{H}^{(3)}, \mathcal{H}^{(2)}} \odot \tilde{\mu}_{\mathcal{H}^{(1)}}$. With some fixed and

known row permutation $\pi^{(2)}$ and $\pi^{(3)}$, the other two matrix blocks $\tilde{B}_2^{(0)}$ and $\tilde{B}_3^{(0)}$ are equal to $\tilde{\Sigma}_{\mathcal{H}^{(3)}, \mathcal{H}^{(1)}} \odot \tilde{\mu}_{\mathcal{H}^{(2)}}$ and $\tilde{\Sigma}_{\mathcal{H}^{(2)}, \mathcal{H}^{(1)}} \odot \tilde{\mu}_{\mathcal{H}^{(3)}}$, separately.

The block $\tilde{B}^{(1)} \in \mathbb{R}^{(|\mathcal{H}|/3)^3 \times k|\mathcal{H}|/3}$ is block diagonal with the identical block $\tilde{\Sigma}_{\mathcal{H}^{(3)}, \mathcal{H}^{(2)}} + \tilde{\mu}_{\mathcal{H}^{(3)}} \odot \tilde{\mu}_{\mathcal{H}^{(2)}}$. Similarly, with the row permutation $\pi^{(2)}$, $\pi^{(3)}$, the other two matrix blocks $\tilde{B}^{(2)}$, $\tilde{B}^{(3)}$ are equal to the block diagonal matrices with the identical block $(\tilde{\Sigma}_{\mathcal{H}^{(3)}, \mathcal{H}^{(1)}} + \tilde{\mu}_{\mathcal{H}^{(3)}} \odot \tilde{\mu}_{\mathcal{H}^{(1)}})$ and $(\tilde{\Sigma}_{\mathcal{H}^{(2)}, \mathcal{H}^{(1)}} + \tilde{\mu}_{\mathcal{H}^{(2)}} \odot \tilde{\mu}_{\mathcal{H}^{(1)}})$ respectively.

Note that we can write the block $\tilde{B}^{(0)}$ as:

$$\begin{aligned} \tilde{B}^{(0)} = & (\tilde{\mu}_{\mathcal{H}^{(3)}} \odot \tilde{\mu}_{\mathcal{H}^{(2)}} + \tilde{\Sigma}_{\mathcal{H}^{(3)}, \mathcal{H}^{(2)}}) \odot \tilde{\mu}_{\mathcal{H}^{(1)}} + (\pi^{(2)})^{-1} (\tilde{\mu}_{\mathcal{H}^{(3)}} \odot \tilde{\mu}_{\mathcal{H}^{(1)}} + \tilde{\Sigma}_{\mathcal{H}^{(3)}, \mathcal{H}^{(1)}}) \odot \tilde{\mu}_{\mathcal{H}^{(2)}} \\ & + (\pi^{(3)})^{-1} (\tilde{\mu}_{\mathcal{H}^{(2)}} \odot \tilde{\mu}_{\mathcal{H}^{(1)}} + \tilde{\Sigma}_{\mathcal{H}^{(2)}, \mathcal{H}^{(1)}}) \odot \tilde{\mu}_{\mathcal{H}^{(3)}} - 2\tilde{\mu}_{\mathcal{H}^{(3)}} \odot \tilde{\mu}_{\mathcal{H}^{(2)}} \odot \tilde{\mu}_{\mathcal{H}^{(1)}}, \end{aligned}$$

where it is easy to see the first summand $(\tilde{\mu}_{\mathcal{H}^{(3)}} \odot \tilde{\mu}_{\mathcal{H}^{(2)}} + \tilde{\Sigma}_{\mathcal{H}^{(3)}, \mathcal{H}^{(2)}}) \odot \tilde{\mu}_{\mathcal{H}^{(1)}}$ is a linear combination of the columns of the block diagonal matrix $\tilde{B}^{(1)}$, and similarly the second and third summands are linear combinations of the columns of $\tilde{B}^{(2)}$ and $\tilde{B}^{(3)}$, and the last summand is simply $-2\tilde{B}_0^{(0)}$. Therefore for some absolute constant C (the smallest singular value corresponding to the linear transformation) we have that:

$$\sigma_{\min}(\tilde{B}_S) \geq C \sigma_{\min}([\tilde{B}_0^{(0)}, \tilde{B}^{(1)}, \tilde{B}^{(2)}, \tilde{B}^{(3)}])$$

Note that $\tilde{B}_0^{(0)} = \tilde{\mu}_{\mathcal{H}^{(3)}} \odot \tilde{\mu}_{\mathcal{H}^{(2)}} \odot \tilde{\mu}_{\mathcal{H}^{(1)}}$ only depends on the randomness over the mean vectors. Note that the Khatri-Rao product is a submatrix of the Kronecker product, therefore for tall matrices Q_1 and Q_2 , we have that $\sigma_{\min}(Q_1 \odot Q_2) \leq \sigma_{\min}(Q_1 \otimes_{kr} Q_2) = \sigma_{\min}(Q_1) \sigma_{\min}(Q_2)$. In particular, we can bound the smallest singular value of $\tilde{B}_0^{(0)}$ with high probability (at least $1 - C\epsilon^{0.5n}$) as follows:

$$\sigma_k(\tilde{B}_0^{(0)}) \geq \sigma_k(\tilde{\mu}_{\mathcal{H}^{(3)}}) \sigma_k(\tilde{\mu}_{\mathcal{H}^{(2)}}) \sigma_k(\tilde{\mu}_{\mathcal{H}^{(1)}}) \geq (\rho\epsilon\sqrt{n})^3.$$

Then condition on the value of the means, we further exploit the randomness over the covariance matrices to lower bound $\sigma_{k|\mathcal{H}|}(\text{Proj}_{\tilde{B}_0^{(0)\perp}[\tilde{B}^{(1)}, \tilde{B}^{(2)}, \tilde{B}^{(3)}]})$. It is almost the same as the argument of the proof for Proposition 3.8. For example, compared

to (3.18) we have the following inequality instead:

$$\sigma_k \left(\text{Proj}_{([\tilde{B}^{(0)}, \tilde{B}^{(2)}, \tilde{B}^{(3)}]_{\{j\} \times \mathcal{H}^{(2)} \times \mathcal{H}^{(3)}})^{\perp}} \text{Proj}_{(\Sigma_{\mathcal{H}^{(2)}, \mathcal{H}^{(3)}} + \tilde{\mu}_{\mathcal{H}^{(2)}} \odot \tilde{\mu}_{\mathcal{H}^{(3)}})^{\perp}} (\tilde{\Sigma}_{\mathcal{H}^{(2)}, \mathcal{H}^{(3)}} + \tilde{\mu}_{\mathcal{H}^{(2)}} \odot \tilde{\mu}_{\mathcal{H}^{(3)}}) \right) \geq \epsilon \rho \sqrt{n},$$

and note that any block in $\tilde{B}^{(0)}$ is independent of the randomness of covariance matrices, and we have $(|\mathcal{H}|/3)^2 - k - 2k|\mathcal{H}|/3 \geq 2k$. Similar modifications apply to the inequalities in (3.20),(3.21).

Finally by the argument of Lemma 3.29 we can bound $\sigma_{\min}(\tilde{B}_S)$ with probability at least $1 - C\epsilon^{0.5n}$ (over the randomness of both the perturbed means and covariance matrices):

$$\sigma_{\min}(\tilde{B}_S) \geq \min\{(\rho\epsilon\sqrt{n})^3, \epsilon\rho\sqrt{n}\} = \epsilon\rho\sqrt{n},$$

as we assume ρ to be small perturbation and $\rho\epsilon\sqrt{n} < 1$.

□

Step 1 (b). Find the projected span of covariance matrices Given the subspace $\mathcal{S} = \text{span}\{\tilde{\mu}^{(i)}, \tilde{\Sigma}_{[\cdot, \mathcal{H}]}^{(i)} : i \in [k]\}$ obtained from Step 1 (a), Step 1(b) finds the span of the covariance matrices with the columns projected to S^{\perp} , namely:

$$\mathcal{U}_S = \text{span}\{\text{Proj}_{S^{\perp}} \tilde{\Sigma}^{(i)} : i \in [k]\}.$$

This is in parallel with Step 1 (b) for the zero-mean case, and we rely on the structure of the two-dimensional slices of \tilde{M}_4 to find the span of the projected covariance matrices. Similar to Claim 3.7 for the zero-mean case, the following claim shows how the structure of the two-dimensional slices is related to the desired span.

Claim 3.18. *For a mixture of general Gaussians, the two-dimensional slices of \tilde{M}_4*

are given by:

$$\begin{aligned} \widetilde{M}_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, I, I) &= \sum_{i=1}^k \widetilde{\omega}_i \left((\widetilde{\Sigma}_{j_1, j_2}^{(i)} + \widetilde{\mu}_{j_1}^{(i)} (\widetilde{\mu}_{j_2}^{(i)})^\top) (\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)} (\widetilde{\mu}^{(i)})^\top) \right. \\ &\quad + \widetilde{\mu}_{j_1}^{(i)} (\widetilde{\mu}^{(i)} (\widetilde{\Sigma}_{[:j_2]}^{(i)})^\top + \widetilde{\Sigma}_{[:j_2]}^{(i)} (\widetilde{\mu}^{(i)})^\top) + \widetilde{\mu}_{j_2}^{(i)} (\widetilde{\mu}^{(i)} (\widetilde{\Sigma}_{[:j_1]}^{(i)})^\top + \widetilde{\Sigma}_{[:j_1]}^{(i)} (\widetilde{\mu}^{(i)})^\top) \\ &\quad \left. + \widetilde{\Sigma}_{[:j_1]}^{(i)} (\widetilde{\Sigma}_{[:j_2]}^{(i)})^\top + \widetilde{\Sigma}_{[:j_2]}^{(i)} (\widetilde{\Sigma}_{[:j_1]}^{(i)})^\top \right), \quad \forall j_1, j_2 \in [n]. \end{aligned}$$

Note that given the set of indices \mathcal{H} we chose in Step 1 (a) and the subspace S , if we pick the indices $j_1, j_2 \in \mathcal{H}$, project the two-dimensional slice to S^\perp , all the rank one terms in the sum are eliminated and the projected slice lies in the desired span \mathcal{U}_S :

$$\text{Proj}_{S^\perp} \widetilde{M}_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, I, I) = \sum_{i=1}^k \widetilde{\omega}_i (\widetilde{\Sigma}_{j_1, j_2}^{(i)} + \widetilde{\mu}_{j_1}^{(i)} (\widetilde{\mu}_{j_2}^{(i)})^\top) \text{Proj}_{S^\perp} \widetilde{\Sigma}^{(i)}, \quad \forall j_1, j_2 \in \mathcal{H}.$$

Applying the same argument as in Lemma 3.10 for the zero-mean case, we can show that with high probability over the perturbation, all the projected slices span the subspace \mathcal{U}_S .

Step 1 (c). Merge the two projections of covariance matrices Pick two disjoint index set \mathcal{H}_1 and \mathcal{H}_2 and repeat the previous two steps 1 (a) and 1 (b), we can obtain the two spans U_1 and U_2 , corresponding to the subspace of the covariance matrices projected to \mathcal{S}_1 and \mathcal{S}_2 , respectively.

In this step, we apply similar techniques as in Step 1 (c) for the zero-mean case to merge the two spans U_1 and U_2 : we first use the overlapping part of the two projections $\text{Proj}_{S_1^\perp}$ and $\text{Proj}_{S_2^\perp}$ to align the basis of U_1 and U_2 , then merge the two spans using the same basis.

Note that for the general case, by definition the span of the mean vectors \widetilde{Z} lie in both subspaces \mathcal{S}_1 and \mathcal{S}_2 , therefore we have $\mathcal{S}_1^\perp \subset \widetilde{Z}^\perp$ and $\mathcal{S}_2^\perp \subset \widetilde{Z}^\perp$. We can show that $\mathcal{S}_1^\perp \cup \mathcal{S}_2^\perp = \widetilde{Z}^\perp$ by lower bounding $\sigma_{n-k}([\text{Proj}_{S_1^\perp}, \text{Proj}_{S_2^\perp}])$ with high probability, similar to that in (3.28). This gives us the span of the mean vectors \widetilde{Z} .

Moreover, in the general case, from merging U_1 and U_2 we are only able to find

the span of covariance matrices projected to the subspace \tilde{Z}^\perp . In particular, we can follow Lemma 3.12 and Lemma 3.15 in Step 1 (c) for the zero-mean case to show that for the general case, we can merge U_1 and U_2 to obtain the span $\text{span}\{\text{Proj}_{\tilde{Z}^\perp} \tilde{\Sigma}^{(i)} : i \in [k]\}$. By further projecting the span to \tilde{Z}^\perp from the right side, we can also obtain $\tilde{\Sigma}_o = \text{span}\{\text{Proj}_{\tilde{Z}^\perp} \tilde{\Sigma}^{(i)} \text{Proj}_{\tilde{Z}^\perp} : i \in [k]\}$.

Step 2. Find the covariance matrices in the subspace orthogonal to the means Given the subspace \tilde{Z} and $\tilde{\Sigma}_o = \text{span}\{\text{Proj}_{\tilde{Z}^\perp} \tilde{\Sigma}^{(i)} \text{Proj}_{\tilde{Z}^\perp} : i \in [k]\}$ obtained from Step 1, Step 2 applies the zero-mean case algorithm to find the covariance matrices projected to the subspace \tilde{Z}^\perp , i.e., $\text{Proj}_{\tilde{Z}^\perp} \tilde{\Sigma}^{(i)} \text{Proj}_{\tilde{Z}^\perp}$'s, as well as find the mixing weights $\tilde{\omega}_i$'s.

This follows the same arguments as in Step 2 and Step 3 for the zero mean case. Consider projecting all the samples to \tilde{Z}^\perp , the subspace orthogonal to all the means. In this subspace, the samples are like from a mixture of zero-mean Gaussians with the projected covariance matrices, and the 4-th and 6-th order moment are given by $\tilde{M}_4(\text{Proj}_{\tilde{Z}^\perp}, \text{Proj}_{\tilde{Z}^\perp}, \text{Proj}_{\tilde{Z}^\perp}, \text{Proj}_{\tilde{Z}^\perp})$ and $\tilde{M}_6(\text{Proj}_{\tilde{Z}^\perp}, \text{Proj}_{\tilde{Z}^\perp}, \text{Proj}_{\tilde{Z}^\perp}, \text{Proj}_{\tilde{Z}^\perp}, \text{Proj}_{\tilde{Z}^\perp}, \text{Proj}_{\tilde{Z}^\perp})$. Since \tilde{Z} is of dimension k , the dimension of the zero-mean Gaussian in the projected space is at least $n - k = O(n)$.

Note that the subspace \tilde{Z}^\perp only depends on the randomness of the means, and random perturbation on the covariance matrices is independent of that of $\tilde{\mu}$. The smoothed analysis for the moment unfolding in Step 2 and tensor decomposition in Step 3 for the zero-mean case, which only depend on the randomness of the covariance matrices, still go through in the projected space.

Step 3. Find the means This step finds the mean vectors based on the outputs of the previous steps. The key observation for this step is about the structure of the 3-rd order moments in the following claim:

Claim 3.19. *Let the matrix $\tilde{M}_{3(1)} \in \mathbb{R}^{n \times n^2}$ be the matricization of \tilde{M}_3 along the first*

dimension. The j -th row of $\widetilde{M}_{3(1)}$ is given by:

$$\begin{aligned} [\widetilde{M}_{3(1)}]_{[j,:]} &= \left[\mathbb{E}[x_j x_{j_1} x_{j_2}] : j_1 \in [n] : j_2 \in [n] \right] \\ &= \sum_{i=1}^k \widetilde{\omega}_i \left(\widetilde{\mu}_j^{(i)} \text{vec}(\widetilde{\Sigma}^{(i)}) + \widetilde{\mu}_j^{(i)} \widetilde{\mu}^{(i)} \odot \widetilde{\mu}^{(i)} + \widetilde{\Sigma}_{[:,j]}^{(i)} \odot \widetilde{\mu}^{(i)} + \widetilde{\mu}^{(i)} \odot \widetilde{\Sigma}_{[:,j]}^{(i)} \right)^\top \end{aligned} \quad (3.37)$$

The following lemma shows how to extract the means $\widetilde{\mu}^{(i)}$'s from $\widetilde{M}_{3(1)}$ using the information of the covariance matrices projected to the subspace orthogonal to the means, i.e. $\widetilde{\Sigma}_o$, and the mixing weights $\widetilde{\omega}_i$'s.

Lemma 3.24. *Given the mixing weights $\widetilde{\omega}_i$'s and the projected covariances $\widetilde{\Sigma}_o^{(i)}$'s, define the matrix $\widetilde{T} \in \mathbb{R}^{n^2 \times k}$ to be the pseudo-inverse of $\widetilde{\Sigma}_o$:*

$$\widetilde{T} = \left[\text{vec}(\widetilde{\Sigma}_o^{(i)}) : i \in [k] \right]^{\dagger\top}.$$

The mean $\widetilde{\mu}^{(i)}$ of the i -th component can be obtained by:

$$\widetilde{\mu}^{(i)} = \frac{1}{\widetilde{\omega}_i} \widetilde{M}_{3(1)} \widetilde{T}_{[:,i]}.$$

This step correctly finds the means if the $\widetilde{\Sigma}_o$ is full rank with good condition number, and this holds with high probability over the perturbation.

Proof. (of Lemma 3.24)

The basic idea is that since $\widetilde{\Sigma}_o$ lies in the span of $\widetilde{P} = \text{Proj}_{\widetilde{Z}^\perp} \otimes_{kr} \text{Proj}_{\widetilde{Z}^\perp}$, and the last three summands in the parenthesis in (3.37) all lie in $\text{span}\{I_n \otimes_{kr} \text{Proj}_{\widetilde{Z}}, \text{Proj}_{\widetilde{Z}} \otimes_{kr} I_n\} = \text{span}\{\widetilde{P}^\perp\}$. Therefore hitting the matrix $\widetilde{M}_{3(1)}$ with $\widetilde{\Sigma}_o^\dagger$ from the right will eliminate those summands and pull out only the mean vectors.

Recall that the columns of the matrix $\widetilde{\Sigma}_o$ are $\text{vec}(\text{Proj}_{\widetilde{Z}^\perp} \widetilde{\Sigma}^{(i)} \text{Proj}_{\widetilde{Z}^\perp}) = \widetilde{P} \text{vec}(\widetilde{\Sigma}^{(i)})$'s, and the columns of $\widetilde{\Sigma}$ are $\text{vec}(\widetilde{\Sigma}^{(i)})$'s.

Note that $\widetilde{T} = (\widetilde{P} \widetilde{\Sigma})^{\dagger\top} = \widetilde{P} \widetilde{\Sigma}^{\dagger\top}$, and the columns of \widetilde{T} lie in $\text{span}\{\widetilde{P}\}$. Also note that for all $i, j \in [k]$ the vectors $\widetilde{\mu}^{(i)} \odot \widetilde{\mu}^{(i)}$, $\widetilde{\Sigma}_{[:,j]}^{(i)} \odot \widetilde{\mu}^{(i)}$ and $\widetilde{\mu}^{(i)} \odot \widetilde{\Sigma}_{[:,j]}^{(i)}$ all lie in the subspace $\text{span}\{I_n \otimes_{kr} \text{Proj}_{\widetilde{Z}}, \text{Proj}_{\widetilde{Z}} \otimes_{kr} I_n\} = \text{span}\{\widetilde{P}^\perp\}$. Therefore these terms will

be eliminated if we multiply the columns of \tilde{T} to the right of $\tilde{M}_{3(1)}$. For the first term $\tilde{\mu}_j^{(i)} \text{vec}(\tilde{\Sigma}^{(i)})$, since $\text{vec}(\tilde{\Sigma}^{(j)})^\top \tilde{T}_{[:,i]} = (\tilde{P} \text{vec}(\tilde{\Sigma}^{(j)}))^\top \tilde{T}_{[:,i]} = 1_{[i=j]}$. Therefore, we have $\tilde{M}_{3(1)} \tilde{T}_{[:,i]} = \tilde{\omega}_i \tilde{\mu}^{(i)}$.

The smoothed analysis for the correctness of this step is easy. We only need to show that both $\tilde{\Sigma}_o$ and $\tilde{\Sigma}$ robustly have full column rank with high probability over perturbation of the covariance matrices, and thus the pseudo-inverse \tilde{T} is well defined. This follows from Lemma 3.31.

Finally, the stability analysis for this step is also straightforward using the perturbation bound for pseudo-inverse in Theorem 1.5. \square

Step 4. Find the unprojected covariance matrices Note that by definition $\tilde{Z} = \text{span}\{\tilde{\mu}^{(i)} : i \in [k]\}$, the projected covariance $\text{Proj}_{\tilde{Z}^\perp}(\tilde{\Sigma}^{(i)})$ we obtained in Step 2 is also equal to $\text{Proj}_{\tilde{Z}^\perp}(\tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)}(\tilde{\mu}^{(i)})^\top)$. In Step 4 we try to recover the missing part of the covariance matrices in the subspace \tilde{Z} . Note that since we have also obtained the means in Step 3, it is equivalent to finding $(\tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)}(\tilde{\mu}^{(i)})^\top)$ for all i . We will show that if we can find the $\text{span}\{(\tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)}(\tilde{\mu}^{(i)})^\top) : i \in [k]\}$, the projected vector $\text{Proj}_{\tilde{Z}^\perp}(\tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)}(\tilde{\mu}^{(i)})^\top)$ can be used as anchor to pin down the unprojected vector.

They key observation for finding the span of $\text{span}\{(\tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)}(\tilde{\mu}^{(i)})^\top) : i \in [k]\}$ is to first construct a 4-th order tensor \tilde{M}'_4 which corresponds to the 4-th moment of a mixture of zero-mean Gaussians with covariance matrices $(\tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)}(\tilde{\mu}^{(i)})^\top)$, and then follow Step 1 in the algorithm for zero-mean case to find the span of the covariance matrices for this new mixture of Gaussians.

The next lemma shows how to construct such 4-th order tensor:

Lemma 3.25. *Given the 4-th moment \tilde{M}_4 for a mixture of Gaussians with parameters $\{\tilde{\omega}_i, \tilde{\mu}^{(i)}, \tilde{\Sigma}^{(i)}\}$, define the 4-th order tensor \tilde{M}'_4 to be:*

$$\tilde{M}'_4 = \tilde{M}_4 + 2 \sum_{i=1}^k \tilde{\omega}_i \tilde{\mu}^{(i)} \otimes^4,$$

then \tilde{M}'_4 is equal to the 4-th moment of a mixture Gaussians with parameters $\{\tilde{\omega}_i, 0, \tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)}(\tilde{\mu}^{(i)})^\top\}$.

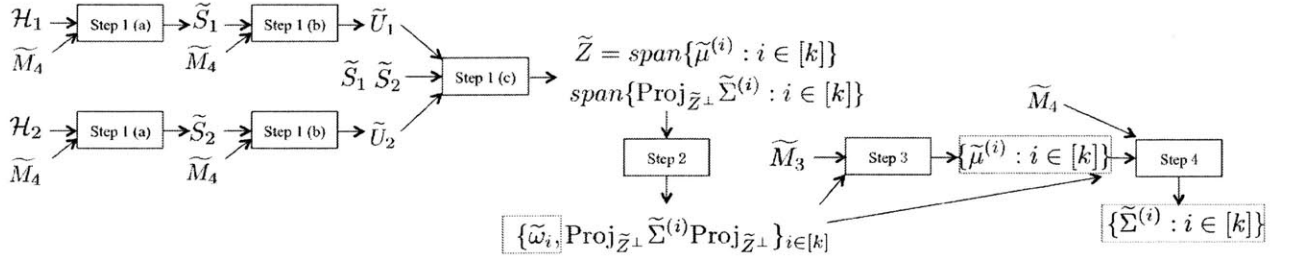


Figure 3-5: Flow of the algorithm for the general case

The proof follows directly from Isserlis' Theorem. Therefore we can repeat Step 1 in the zero-mean case here to find the span of the space $\{\text{vec}(\tilde{\Sigma}^{(i)} + \tilde{\mu}^{(i)} \odot \tilde{\mu}^{(i)}) : i \in [k]\}$. Since we also know the projection of $\tilde{\Sigma}^{(i)}$'s in a large subspace (in the subspace $\text{Proj}_{\tilde{Z}^\perp} \otimes_{kr} \text{Proj}_{\tilde{Z}^\perp}$ obtained from Step 2), we can easily recover $\tilde{\Sigma}^{(i)}$'s:

Lemma 3.26. *For any matrix $U \in \mathbb{R}^{d \times k}$ and any subspace P , given $P^\top U$ and the span S of columns of U , the matrix U can be computed as*

$$U = S(P^\top S)^\dagger(P^\top U).$$

Further, this procedure is stable if $\sigma_{\min}(P^\top S)$ is lower bounded.

Proof. This is a special case of the Step 1 (c) where we merge two projections of an unknown subspace.

The span S is equal to UV for some unknown matrix V . We can compute $V = (P^\top U)^\dagger P^\top S$, and hence $U = SV^{-1} = S(P^\top S)^\dagger(P^\top U)$. The stability analysis is similar (and simpler than) Lemma 3.12. \square

We will apply this lemma to where the subspace P is $\text{Proj}_{\tilde{Z}^\perp} \otimes_{kr} \text{Proj}_{\tilde{Z}^\perp}$. Since the perturbation of the means and the covariance matrices are independent, we can lower bound the smallest singular value of $P^\top S$.

Proof Sketch of the Main Theorem 3.4 The proof follows the same strategy as Theorem 3.5. First we apply the union bound to all the smoothed analysis lemmas, this will ensure the matrices we are inverting all have good condition number, and the whole algorithm is robust to noise.

Then in order to get the desired accuracy ϵ , we need to guarantee inverse polynomial accuracy in different steps (through the stability lemmas). The flow of the algorithm is illustrated in Figure 3-5. In the end all the requirements becomes a inverse polynomial accuracy requirement on \widehat{M}_4 and \widehat{M}_6 , which we obtain by Lemma 3.22.

3.4.6 Proofs for Moment Structures

In this section we characterize the structure of the 3-rd, 4-th and 6-th moments of Gaussians mixtures.

As described in Section 3.2, the m -th order moments of the Gaussian mixture model are given by the following m -th order symmetric tensor $M \in \mathbb{R}_{sym}^{n \times \dots \times n}$:

$$[M_m]_{j_1, \dots, j_m} := \mathbb{E}[x_{j_1} \dots x_{j_m}] = \sum_{i=1}^k \omega_i \mathbb{E}[y_{j_1}^{(i)} \dots y_{j_m}^{(i)}], \quad \forall j_1, \dots, j_m \in [n],$$

where $y^{(i)}$ corresponds to the n -dimensional Gaussian distribution $\mathcal{N}(\mu^{(i)}, \Sigma^{(i)})$.

Gaussian distribution is a highly symmetric distribution, and in the zero-mean case the higher moments are well-understood by Isserlis' Theorem:

Theorem 3.17 (Isserlis). *Let $\mathbf{y} = (y_1, \dots, y_{2t})$ be a multivariate Gaussian random vector with mean zero and covariance Σ , then*

$$\begin{aligned} \mathbb{E}[y_1 \dots y_{2t}] &= \sum \prod \Sigma_{u,v}, \\ \mathbb{E}[y_1 \dots y_{2t-1}] &= 0, \end{aligned}$$

where the summation is taken over all distinct ways of partitioning y_1, \dots, y_{2t} into t pairs, which correspond to all the perfect matchings in a complete graph. Thus there are $(2t - 1)!!$ terms in the sum, and each summand is a product of t terms.

The non-zero mean case is a direct corollary using Isserlis' Theorem and linearity of expectation.

Corollary 3.1. *Let $\mathbf{y} = (y_1, \dots, y_t)$ be a multivariate Gaussian random vector with*

mean μ and covariance Σ , then

$$\mathbb{E}[y_1 \dots y_t] = \sum \prod \Sigma_{u,v} \prod \mu_w.$$

where the summation is taken over all distinct ways of partitioning y_1, \dots, y_t into p pairs of (u, v) and s singletons of (w) , where $p \geq 0$, $s \geq 0$ and $2p + s = t$.

$$\text{As an example, } \mathbb{E}[y_1 y_2 y_3] = \mu_1 \mu_2 \mu_3 + \mu_1 \Sigma_{2,3} + \mu_2 \Sigma_{1,3} + \mu_3 \Sigma_{1,2}.$$

Proof of Lemma 3.1

We shall first prove Lemma 3.1 in Section 3.2. Recall that this lemma shows that for mixture of zero-mean Gaussians, the 4-th moments \overline{M}_4 and the 6-th moments \overline{M}_6 with distinct indices can be viewed as a linear projection of the unfolded moment X_4 and X_6 defined in (3.1).

Proof. (of Lemma 3.1)

By Isserlis Theorem 3.6, the mapping $\sqrt{3}\mathcal{F}_4$ is characterized by: $(\forall 1 \leq j_1 < j_2 < j_3 < j_4 \leq n)$

$$\begin{aligned} [M_4]_{j_1, j_2, j_3, j_4} &= \sum_{i=1}^k \omega_i (\Sigma_{j_1, j_2}^{(i)} \Sigma_{j_3, j_4}^{(i)} + \Sigma_{j_1, j_3}^{(i)} \Sigma_{j_2, j_4}^{(i)} + \Sigma_{j_1, j_4}^{(i)} \Sigma_{j_2, j_3}^{(i)}) \\ &= [X_4]_{(j_1, j_2), (j_3, j_4)} + [X_4]_{(j_1, j_3), (j_2, j_4)} + [X_4]_{(j_1, j_4), (j_2, j_3)}. \end{aligned}$$

Therefore, with the normalization constant $\sqrt{3}$, the (j_1, j_2, j_3, j_4) -th mapping of \mathcal{F}_4 is a projection of the three elements in X_4 . Similarly, we have for $\sqrt{15}\mathcal{F}_6$: $(\forall 1 \leq j_1 < j_2 < \dots < j_6 \leq n)$

$$\begin{aligned} & [M_6]_{j_1, j_2, j_3, j_4, j_5, j_6} \\ &= [X_6]_{(j_1, j_2), (j_3, j_4), (j_5, j_6)} + [X_6]_{(j_1, j_3), (j_2, j_4), (j_5, j_6)} + [X_6]_{(j_1, j_4), (j_2, j_3), (j_5, j_6)} + [X_6]_{(j_1, j_5), (j_2, j_3), (j_4, j_6)} \\ &+ [X_6]_{(j_1, j_2), (j_5, j_3), (j_4, j_6)} + [X_6]_{(j_1, j_3), (j_2, j_5), (j_4, j_6)} + [X_6]_{(j_1, j_2), (j_4, j_5), (j_3, j_6)} + [X_6]_{(j_1, j_4), (j_2, j_5), (j_3, j_6)} \\ &+ [X_6]_{(j_1, j_5), (j_2, j_4), (j_3, j_6)} + [X_6]_{(j_1, j_3), (j_4, j_5), (j_2, j_6)} + [X_6]_{(j_1, j_4), (j_3, j_5), (j_2, j_6)} + [X_6]_{(j_1, j_5), (j_3, j_2), (j_2, j_6)} \\ &+ [X_6]_{(j_2, j_3), (j_4, j_5), (j_1, j_6)} + [X_6]_{(j_2, j_4), (j_3, j_5), (j_1, j_6)} + [X_6]_{(j_2, j_5), (j_3, j_4), (j_1, j_6)}. \end{aligned}$$

Thus with the normalization constant $\sqrt{15}$, the mapping \mathcal{F}_6 is a linear projection. \square

Slices of Moments

Next we shall characterize the slices of the moments of mixture of Gaussians.

For mixture of zero-mean Gaussians, a one-dimensional slice of the 4th moment tensor is a vector in the span of corresponding columns of the covariance matrices:

Claim 3.20 (Claim 3.1 restated). *For a mixture of zero-mean Gaussians, the one-dimensional slices of the 4-th moments M_4 are given by:*

$$M_4(e_{j_1}, e_{j_2}, e_{j_3}, I) = \sum_{i=1}^k \omega_i \left(\Sigma_{j_1, j_2}^{(i)} \Sigma_{[:, j_3]}^{(i)} + \Sigma_{j_1, j_3}^{(i)} \Sigma_{[:, j_2]}^{(i)} + \Sigma_{j_2, j_3}^{(i)} \Sigma_{[:, j_1]}^{(i)} \right), \quad \forall j_1, j_2, j_3 \in [n].$$

Proof. By the definition of multilinear map, $M_4(e_{j_1}, e_{j_2}, e_{j_3}, I)$ is a vector whose p -th entry is equal to $M_4(e_{j_1}, e_{j_2}, e_{j_3}, e_p)$. We can compute this entry by Isserlis' Theorem:

$$M_4(e_{j_1}, e_{j_2}, e_{j_3}, e_p) = \sum_{i=1}^k \omega_i \left(\Sigma_{j_1, j_2}^{(i)} \Sigma_{[p, j_3]}^{(i)} + \Sigma_{j_1, j_3}^{(i)} \Sigma_{[p, j_2]}^{(i)} + \Sigma_{j_2, j_3}^{(i)} \Sigma_{[p, j_1]}^{(i)} \right),$$

this directly implies the claim. \square

For mixture of zero-mean Gaussians, a two-dimensional slice of the 4th moment M_4 is a matrix, and it is a linear combination of the covariance matrices with some additive rank one matrices:

Claim 3.21 (Claim 3.2 restated). *For a mixture of zero-mean Gaussians, the two-dimensional slices of the 4-th moment M_4 are given by:*

$$M_4(e_{j_1}, e_{j_2}, I, I) = \sum_{i=1}^k \omega_i \left(\Sigma_{j_1, j_2}^{(i)} \Sigma^{(i)} + \Sigma_{[:, j_1]}^{(i)} (\Sigma_{[:, j_2]}^{(i)})^\top + \Sigma_{[:, j_2]}^{(i)} (\Sigma_{[:, j_1]}^{(i)})^\top \right), \quad \forall j_1, j_2 \in [n].$$

Proof. Again this follows from Isserlis' theorem. By definition of multilinear map this is a matrix whose (p, q) -th entry is equal to

$$M_4(e_{j_1}, e_{j_2}, e_p, e_q) = \sum_{i=1}^k \omega_i \left(\Sigma_{j_1, j_2}^{(i)} \Sigma_{[p, q]}^{(i)} + \Sigma_{j_1, p}^{(i)} \Sigma_{[q, j_2]}^{(i)} + \Sigma_{j_2, p}^{(i)} \Sigma_{[q, j_1]}^{(i)} \right),$$

and this directly implies the claim. \square

Similarly, for mixture of general Gaussians, we prove the following claims:

Claim 3.22 (Claim 3.17 restated). *For a mixture of general Gaussians, the (j_1, j_2, j_3) -th one-dimensional slice of M_4 is given by:*

$$M_4(e_{j_1}, e_{j_2}, e_{j_3}, I) = \sum_{i=1}^n \omega_i \left(\mu_{j_1}^{(i)} \mu_{j_2}^{(i)} \mu_{j_3}^{(i)} \mu^{(i)} + \sum_{\pi \in \left\{ \begin{array}{l} (j_1, j_2, j_3), \\ (j_2, j_3, j_1), \\ (j_3, j_1, j_2) \end{array} \right\}} \left(\Sigma_{\pi_1, \pi_2}^{(i)} \Sigma_{[:, \pi_3]}^{(i)} + \mu_{\pi_1}^{(i)} \mu_{\pi_2}^{(i)} \Sigma_{[:, \pi_3]}^{(i)} + \Sigma_{\pi_1, \pi_2}^{(i)} \mu_{\pi_3}^{(i)} \mu^{(i)} \right) \right).$$

Proof. This is very similar to Claim 3.1 and follows from the corollary of Isserlis's theorem (Corollary 3.1). There are 10 ways to partition the indices $\{j_1, j_2, j_3, j_4\}$ into pairs and singletons: $((j_1), (j_2), (j_3), (j_4))$, $((j_1, j_2), (j_3), (j_4))$, $((j_1, j_3), (j_2), (j_4))$, $((j_1, j_4), (j_2), (j_3))$, $((j_2, j_3), (j_1), (j_4))$, $((j_2, j_4), (j_1), (j_3))$, $((j_3, j_4), (j_1), (j_2))$, $((j_1, j_2), (j_3, j_4))$, $((j_1, j_3), (j_2, j_4))$, $((j_1, j_4), (j_2, j_3))$. From this enumeration, we can specify each element in the vector of the one-dimensional slice. \square

Claim 3.23 (Claim 3.19 restated). *For a mixture of general Gaussians, let the matrix $M_{3(1)} \in \mathbb{R}^{n \times n^2}$ be the matricization of M_3 along the first dimension. The j -th row of $M_{3(1)}$ is given by:*

$$[M_{3(1)}]_{[j, :]} = \sum_{i=1}^k \omega_i \left(\mu_j^{(i)} \text{vec}(\Sigma^{(i)}) + \mu_j^{(i)} \mu^{(i)} \odot \mu^{(i)} + \Sigma_{[:, j]}^{(i)} \odot \mu^{(i)} + \mu^{(i)} \odot \Sigma_{[:, j]}^{(i)} \right)^\top.$$

Proof. Note that $[M_{3(1)}]_{[j, :]} = \left[\text{vec}(\mathbb{E}[x_j x x^\top]) \right] = \text{vec}(\mathbb{E}[x_j x \odot x])$. Again following the corollary of Isserlis's theorem (Corollary 3.1, there are 4 ways to partition the indices $\{j_1, j_2, j_3\}$ into pairs and singletons: $((j_1), (j_2, j_3))$, $((j_1), (j_2), (j_3))$, $((j_1, j_2), (j_3))$, $((j_2), (j_1, j_3))$, and they correspond to the four terms in the summation.) \square

Two mixtures with same M_4 but different X_4

Since M_4 gives linear observations on the symmetric low rank matrix X_4 , it is natural to wonder whether we can use matrix completion techniques to recover X_4 from M_4 . Here we show this is impossible by giving a counter example: there are two mixture of Gaussians that generates the same 4th moment M_4 , but has different X_4 (even the span of $\Sigma^{(i)}$'s are different).

By $((a, b), (c, d))$ we denote a 5×5 matrix A which has 2's on diagonals, and the only nonzero off-diagonal entries are $A_{a,b} = A_{b,a} = A_{c,d} = A_{d,c} = 1$. For example, $((1, 2), (4, 5))$ will be the following matrix:

$$\begin{pmatrix} 2 & 1 & & & \\ & 2 & & & \\ & & 2 & & \\ & & & 2 & 1 \\ & & & 1 & 2 \end{pmatrix},$$

where all the missing entries are 0's. Now we construct two mixtures of 3 Gaussians, all with mean 0 and weight $1/3$. The covariance matrices are $((1, 2), (4, 5)), ((1, 3), (2, 5)), ((1, 4), (3, 5))$ for the first mixture and $((1, 2), (3, 5)), ((1, 3), (4, 5)), ((1, 4), (2, 5))$ for the second mixture. These are clearly different mixtures with different span of $\Sigma^{(i)}$'s: in the first mixture, $\Sigma_{1,2}^{(i)} = \Sigma_{4,5}^{(i)}$ for all matrices, but this is not true for the second mixture.

These two mixture of Gaussians have the same 4th moment M_4 . This can be checked by using Isserlis' theorem to compute the moments. Intuitively, this is true because all the pairs $(1, i)$ and $(i, 5)$ appeared exactly twice in the covariance matrices for both mixtures; also, every 4-tuple $(1, i, j, 5)$ appeared exactly once in the covariance matrices for both mixtures.

3.4.7 Auxiliary Lemmas

In general, matrix perturbation bounds are the key for the perturbation lemmas, and concentration bounds are crucial for the smoothed analysis lemmas. We also prove some corollaries of known results that are very useful in our settings.

Lowerbounding the Smallest Singular Value

Sometimes, it is easier to consider the projection of a matrix. Lowerbounding the smallest singular value of a projection will imply the same lowerbound on the original matrix:

Lemma 3.27. *Suppose $A \in \mathbb{R}^{m \times n}$, let $P \in \mathbb{R}^{m \times d}$ be a subspace, then $\sigma_k(P^\top A) \leq \sigma_k(A)$.*

Proof. Observe that $(P^\top A)^\top (P^\top A) = A^\top (PP^\top) A \preceq A^\top A$ (because P is a subspace). Therefore the eigenvalues of $(P^\top A)^\top (P^\top A)$ must be dominated by the eigenvalues of $A^\top A$. Then the lemma follows from the definition of singular values. \square

As a corollary we have the following lemma:

Lemma 3.28. *Let $A \in \mathbb{R}^{m \times n}$ and suppose that $m \geq n$. For any projection Proj_S , we have that the singular values are non-increasing after the projection:*

$$\sigma_i(\text{Proj}_S(A)) \leq \sigma_i(A), \quad \text{for } i = 1, \dots, n.$$

In several places of this work we want to bound the singular value of a matrix, where part of the matrix has a block structure.

Lemma 3.29. *For given matrices $B^{(i)} \in \mathbb{R}^{m \times n}$ and $C^{(i)} \in \mathbb{R}^{m \times n'}$ for $i = 1, \dots, d$. Suppose $md > (n + n'd)$, Define the tall matrix $A \in \mathbb{R}^{md \times (n + dn')}$:*

$$A = \begin{bmatrix} B^{(1)} & C^{(1)} & 0 & \dots & 0 \\ B^{(2)} & 0 & C^{(2)} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ B^{(d)} & 0 & 0 & \dots & C^{(d)} \end{bmatrix} = [B, \text{diag}(C^{(i)})].$$

The smallest singular value is bounded by:

$$\sigma_{(n+dn')}(A) \geq \min\{\sigma_n(B), \sigma_{n'}(\text{Proj}_{(B^{(i)})^\perp} C^{(i)}) : i = 1, \dots, d\}.$$

Proof. The idea is to break the matrix into two parts $A = \text{Proj}_B A + \text{Proj}_{B^\perp} A$. Since these two spaces are orthogonal we know $\sigma_{(n+dn')}(A) \geq \min\{\sigma_n(\text{Proj}_B A), \sigma_{dn'}(\text{Proj}_{B^\perp} A)\}$.

For the first part, clearly $\sigma_n(\text{Proj}_B A) \geq \sigma_n(B)$, as B is a submatrix of $\text{Proj}_B A$.

For the second part, we actually do the projection to a smaller subspace: for each block we project to the orthogonal subspace of $B^{(i)}$. Under this projection, the block structure is preserved. The dn' -th singular value must be at least the minimum of the n' -th singular value of the blocks. In summary we have:

$$\begin{aligned} \sigma_{(n+dn')}(A) &\geq \min\{\sigma_n(B), \sigma_{dn'}(\text{Proj}_{B^\perp} \text{diag}(C^{(i)}))\} \\ &\geq \min\{\sigma_n(B), \sigma_{dn'}(\text{Proj}_{\text{diag}((B^{(i)})^\perp)} \text{diag}(C^{(i)}))\} \\ &\geq \min\{\sigma_n(B), \sigma_{dn'}(\text{diag}(\text{Proj}_{(B^{(i)})^\perp} C^{(i)}))\} \\ &\geq \min\{\sigma_n(B), \sigma_{n'}(\text{Proj}_{(B^{(i)})^\perp} C^{(i)}) : i = 1, \dots, d\}. \end{aligned}$$

□

Smallest singular value of random matrices In our analysis, we often also want to bound the smallest singular value of a matrix whose entries are Gaussian random variables. Our analysis mostly builds on the following results in random matrix theory.

For a random rectangular matrix, [101] gives the following nice result:

Lemma 3.30 (Theorem 1.1 in [101]). *Let $A \in \mathbb{R}^{m \times n}$ and suppose that $m \geq n$. Assume that the entries of A are independent standard Gaussian variable, then for every $\epsilon > 0$, with probability at least $1 - (C\epsilon)^{m-n+1} + e^{-C'n}$, where C, C' are two absolute constants, we have:*

$$\sigma_n(A) \geq \epsilon(\sqrt{m} - \sqrt{n-1}).$$

We will mostly use an immediate corollary of the above lemma with slightly simpler form:

Corollary 3.2. *Let $A \in \mathbb{R}^{m \times n}$ and suppose that $m \geq 2n$. Assume that the entries of A are independent standard Gaussian variable, then for every $\epsilon > 0$, and for some absolute constant C , with probability at least $1 - (C\epsilon)^{0.5m}$, we have:*

$$\sigma_n(A) \geq \epsilon\sqrt{m}.$$

This lemma can also be applied to a projection of a Gaussian matrix:

Lemma 3.31. *Given a Gaussian random matrix $E \in \mathbb{R}^{m \times n}$, for some set $\mathcal{J} \in [m]$ define $E_{\mathcal{J}} = [E_{[j, \cdot]} : j \in \mathcal{J}]$ and $E_{\mathcal{J}^c} = [E_{[j, \cdot]} : j \in [m]/\mathcal{J}]$. Define matrix $S \in \mathbb{R}^{n \times r}$ whose columns are orthonormal. Suppose that the matrix S is an arbitrary function of $E_{\mathcal{J}}$ and is independent of $E_{\mathcal{J}^c}$. Assume that*

$$m - |\mathcal{J}| - r \geq 2n \tag{3.38}$$

Then for any $\epsilon > 0$, we have that with probability at least $1 - (C\epsilon)^{0.5(m-|\mathcal{J}|-r)}$, for some absolute constant C , the smallest singular value of the projected random matrix is bounded by:

$$\sigma_n(\text{Proj}_{S^\perp} E) \geq \epsilon\sqrt{m - |\mathcal{J}| - r}. \tag{3.39}$$

Proof. For a matrix $A \in \mathbb{R}^{m \times n}$, define the fixed matrix $P_{\mathcal{J}^c} \in \mathbb{R}^{(m-|\mathcal{J}|) \times m}$ such that:

$$[[P_{\mathcal{J}^c}]_{[\cdot, j]} : j \in \mathcal{J}] = 0, \quad [[P_{\mathcal{J}^c}]_{[\cdot, j]} : j \in [n]/\mathcal{J}] = I_{(m-|\mathcal{J}|) \times (m-|\mathcal{J}|)},$$

which only keeps the coordinates that correspond to $[m]/\mathcal{J}$ of any vector in \mathbb{R}^m . Note

that

$$\begin{aligned}
\sigma_n(\text{Proj}_{S^\perp} E) &\geq \sigma_n(P_{J^c}(\text{Proj}_{S^\perp} E)) \\
&\geq \sigma_n(\text{Proj}_{(P_{J^c} S)^\perp} P_{J^c} \text{Proj}_{S^\perp} E) \\
&= \sigma_n(\text{Proj}_{(P_{J^c} S)^\perp} P_{J^c} E).
\end{aligned}$$

We justify the last equality below. Note that

$$\text{Proj}_{S^\perp} E = E - \text{Proj}_S E,$$

and note that the columns of $(P_{J^c} \text{Proj}_S E)$ lie in the column span of $P_{J^c} S$, therefore,

$$\begin{aligned}
\text{Proj}_{(P_{J^c} S)^\perp} P_{J^c} \text{Proj}_{S^\perp} E &= \text{Proj}_{(P_{J^c} S)^\perp} P_{J^c} E - \text{Proj}_{(P_{J^c} S)^\perp} (P_{J^c} \text{Proj}_S E) \\
&= \text{Proj}_{(P_{J^c} S)^\perp} P_{J^c} E.
\end{aligned}$$

Finally, note that $P_{J^c} S$, with column rank no more than r , is independent of $P_{J^c} E$, which is a random Gaussian matrix of size $(m - |\mathcal{J}|) \times n$, therefore we have that $\text{Proj}_{(P_{J^c} S)^\perp} P_{J^c} E$ is equivalent to a $(m - |\mathcal{J}| - r) \times n$ random Gaussian matrix. Since (3.38) is satisfied, we can apply Lemma 3.30 and conclude (3.39) with high probability. \square

However, in the smoothed analysis setting, the matrix we are interested in are often not random Gaussian matrices. Instead they are fixed matrices perturbed by Gaussian variables. We call these ‘‘perturbed rectangular matrices’’, their singular values can be bounded as follows:

Lemma 3.32 (Perturbed rectangular matrices). *Let $A \in \mathbb{R}^{m \times n}$ and suppose that $m \geq 3n$. If all the entries of A are independently ρ -perturbed to yield \tilde{A} , then for any $\epsilon > 0$, with probability at least $1 - (C\epsilon)^{0.25m}$, for some absolute constant C , the smallest singular value of \tilde{A} is bounded below by:*

$$\sigma_n(\tilde{A}) \geq \epsilon \rho \sqrt{m}.$$

Proof. The idea is to use the previous lemma and project to the orthogonal subspace of A . We have that $\tilde{A} = A + E$, where $E \in \mathbb{R}^{m \times n}$ is a random Gaussian matrix.

$$\sigma_n(\tilde{A}) \geq \sigma_n(\text{Proj}_{A^\perp} \tilde{A}) = \sigma_n(\text{Proj}_{A^\perp} E).$$

Since $m - n > 2n$, we can apply Lemma 3.31 to conclude that for any $\epsilon > 0$,

$$\sigma_n(\text{Proj}_{A^\perp} E) \geq \epsilon \rho \sqrt{m},$$

with probability at least $1 - (C\epsilon)^{0.5(m-n)} \leq 1 - (C\epsilon)^{0.25m}$. □

Projection of random vectors

In Step 2, we need to bound the norm of a random vector of the form $u \odot v$ after a projection, where u and v are two Gaussian vectors. In order to show this, we apply the result in [124] which provides a concentration bound of projection of well-behaved (K -concentrated) random vectors.

First we cite the definition of “ K -concentrated” below:

Definition 3.5. *A random vector $X = (\xi_1, \xi_2, \dots, \xi_n)$ is K -concentrated (where K may depend on n) if there are constants $C, C' > 0$ such that for any convex, 1-Lipschitz function $f : \mathbb{C}^n \rightarrow \mathbb{R}$ and for any $t > 0$, we have:*

$$\Pr[|F(X) - \text{med}(F(X))| \geq t] \leq C \exp\left(-C' \frac{t^2}{K^2}\right),$$

where $\text{med}(\cdot)$ denotes the median of a random variable (choose an arbitrary one if there are many).

Lemma 3.33 (Concentration for Random Projections (Lemma 1.2 in [124])). *Let v be a K -concentrated random vector in \mathbb{C}^n . The entries of v has expected norm 1. Then there are constants $C, C' > 0$ such that the following holds. Let Proj_S be a*

projection to a d -dimensional subspace in \mathbb{C}^n .

$$\mathbb{P}\left(\left|v^\top \text{Proj}_S v - d\right| \geq 2t\sqrt{d} + t^2\right) \leq C \exp(-C' \frac{t^2}{K^2}).$$

In order to apply this lemma in our setting, we need to prove the vectors that we are interested in is K -concentrated:

Lemma 3.34. *Conditioned on the high probability event that $\|E_{[i,i]}\|, \|E_{[i,j]}\| \leq 2\sqrt{n_2}$, the vector $[[E_{[i,i]} \odot E_{[i,j]}]_{s,s'} : s < s']$ is $2\sqrt{n_2}$ -concentrated.*

Proof. For any 1-Lipschitz function F on $[[E_{[i,i]} \odot E_{[i,j]}]_{s,s'} : s < s']$, we can define a function $G(E_{[i,i]}, E_{[i,j]}) = F([[E_{[i,i]} \odot E_{[i,j]}]_{s,s'} : s < s'])$ (if $i = j$ then the function G only takes $E_{[i,i]}$ as the variable). Under the assumption that $\|E_{[i,i]}\|, \|E_{[i,j]}\| \leq 2\sqrt{n_2}$, this new function G is $2\sqrt{n_2}$ -Lipschitz.

Now we extend G to G^* when the input $\|E_{[i,i]}\|, \|E_{[i,j]}\| > 2\sqrt{n_2}$. Define the truncation function $\text{trunc}(v) = v$ for $\|v\| \leq 2\sqrt{n_2}$, and $\text{trunc}(v) = 2\sqrt{n_2}v/\|v\|$ for $\|v\| > 2\sqrt{n_2}$. Define the extended function $G^*(E_{[i,i]}, E_{[i,j]}) = G(\text{trunc}(E_{[i,i]}), \text{trunc}(E_{[i,j]}))$, which is still $2\sqrt{n_2}$ -Lipschitz since the truncation function is 1-Lipschitz.

Note that for the two Gaussian random vectors $E_{[i,i]}, E_{[i,j]} \sim N(0, I)$, we can apply Gaussian concentration bound in Theorem 3.18 on G^* , which implies

$$\mathbb{P}[|G^*(E_{[i,i]}, E_{[i,j]}) - \text{med}(G^*(E_{[i,i]}, E_{[i,j]}))| \geq t] \leq C \exp(-C't^2/4n_2).$$

Since the probability of the event $\|E_{[i,i]}\|, \|E_{[i,j]}\| > 2\sqrt{n_2}$ is very small ($\sim \exp(-\Omega(n_2))$), we have $\delta = |\text{med}(G(E_{[i,i]}, E_{[i,j]})) - \text{med}(G^*(E_{[i,i]}, E_{[i,j]}))|$ in the order of $O(\sqrt{n_2})$.

Therefore, for $t \sim \Omega(\sqrt{n_2})$, we have

$$\begin{aligned} & \mathbb{P}[|G^*(E_{[i,i]}, E_{[i,j]}) - \text{med}(G(E_{[i,i]}, E_{[i,j]}))| \geq t] \\ & \leq \mathbb{P}[|G^*(E_{[i,i]}, E_{[i,j]}) - \text{med}(G^*(E_{[i,i]}, E_{[i,j]}))| \geq t - \delta] \\ & \leq C \exp(-C't^2/4n_2). \end{aligned}$$

Finally,

$$\begin{aligned}
& \mathbb{P} \left[\left| G(E_{[\cdot,i]}, E_{[\cdot,j]}) - \text{med}(G(E_{[\cdot,i]}, E_{[\cdot,j]})) \right| \geq t \mid \|E_{[\cdot,i]}\|, \|E_{[\cdot,j]}\| \leq 2\sqrt{n_2} \right] \\
& \leq \frac{\mathbb{P} \left[\left| G^*(E_{[\cdot,i]}, E_{[\cdot,j]}) - \text{med}(G(E_{[\cdot,i]}, E_{[\cdot,j]})) \right| \geq t \right]}{\mathbb{P} \left[\|E_{[\cdot,i]}\| \geq 2\sqrt{n_2} \text{ or } \|E_{[\cdot,j]}\| \geq 2\sqrt{n_2} \right]} \\
& \leq C \exp(-C't^2/4n_2).
\end{aligned}$$

Therefore the random vector $[[E_{[\cdot,i]} \odot E_{[\cdot,j]}]_{s,s'} : s < s']$ is $2\sqrt{n_2}$ -concentrated. \square

Theorem 3.18 (Gaussian concentration bound). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a function which is Lipschitz with constant 1. Consider a random vector $X \sim \mathcal{N}(0, I_n)$. For any $s > 0$ we have*

$$\mathbb{P} \left(|f(X) - \mathbb{E}[f(X)]| \geq s \right) \leq 2e^{-Cs^2},$$

for all $s > 0$ and some absolute constant $C > 0$.

Gaussian Chaoses

In Step 2, we want to show that the inner product of two random vectors of the form $\langle \text{Proj}(u \odot v), \text{Proj}(u' \odot v) \rangle$ is small, where u, u' and v, v' are Gaussian vectors. In order to show this, we treat the inner product as a (homogeneous) Gaussian chaos, which is defined to be a homogeneous polynomial over Gaussian random variables¹⁰. Our analysis builds on the results of many works studying the concentration bound of Gaussian chaoses.

For decoupled Gaussian chaoses, we mostly use the following theorem, which is a simple corollary of Lemma 3.35.

Theorem 3.19. *Suppose $a = (a_{i_1, \dots, i_d})_{1 \leq i_1, \dots, i_d \leq n}$ is a d -indexed array, and $\|a\|_F$ denotes its Frobenius norm. Let $(X_i^{(j)})_{1 \leq i \leq n, j=1, \dots, d}$ be independent copies of $X \sim$*

¹⁰In fact, the squared norm of projected random vectors considered previously is a special case of Gaussian chaos, and we treat it separately.

$\mathcal{N}(0, I_n)$. For any fixed $\epsilon > 0$, with probability at least $1 - C \exp(-C' n^{2\epsilon/d})$,

$$\left| \sum_{i_1, \dots, i_d=1}^n a_{i_1, \dots, i_d} X_{i_1}^{(1)} \cdots X_{i_d}^{(d)} \right| \leq \|a\|_F n^\epsilon.$$

Lemma 3.35 (Gaussian chaos concentration (Corollary 1 in [74])). *Suppose $a = (a_{i_1, \dots, i_d})_{1 \leq i_1, \dots, i_d \leq n}$ is a d -indexed array. Consider a decoupled Gaussian chaos $G = \sum_{i_1, \dots, i_d} a_{i_1, \dots, i_d} X_{i_1}^{(1)} \cdots X_{i_d}^{(d)}$, where $X_i^{(k)}$ are independent copies of the standard normal random variable for all $i \in [n], k \in [d]$.*

$$\mathbb{P}(|G| \geq t) \leq C_d \exp\left(-\frac{1}{C_d} \min_{1 \leq k \leq d} \min_{(I_1, \dots, I_k) \in S(k, d)} \left(\frac{t}{\|a\|_{I_1, \dots, I_k}}\right)^{2/k}\right),$$

where $C_d \in (0, \infty)$ depends only on d , and $S(k, d)$ denotes a set of all partitions of $\{1, \dots, d\}$ into k nonempty disjoint sets I_1, \dots, I_k , and the norm $\|\cdot\|_{I_1, \dots, I_k}$ is given by:

$$\|a\|_{I_1, \dots, I_k} := \sup \left\{ \sum_{i_1, \dots, i_d} a_{i_1, \dots, i_d} x_{i_1}^{(1)} \cdots x_{i_k}^{(k)} : \sum_{i_1} (x_{i_1}^{(1)})^2 \leq 1, \dots, \sum_{i_k} (x_{i_k}^{(k)})^2 \leq 1 \right\}.$$

Proof. (of Theorem 3.19) Apply the inequality:

$$\|a\|_{\{1\}, \dots, \{d\}} \leq \|a\|_{I_1, \dots, I_k} \leq \|a\|_{[d]} = \|a\|_F, \quad \forall (I_1, \dots, I_k) \in S(k, d).$$

For a fixed order d and for any $\epsilon > 0$, apply Lemma 3.35 and set $t = n^\epsilon \|a\|_F$. We have that $\mathbb{P}(|G| \geq t) \leq C \exp(-C' n^{2\epsilon/d})$, for some constant C, C' . \square

For coupled Gaussian chaoses, namely when $X^{(j)}$'s are identical copies of the same X , we first cite the following decoupling theorem in [40].

Theorem 3.20. (Decoupling) *Let $(a_{i_1, \dots, i_d})_{1 \leq i_1, \dots, i_d \leq n}$ be a symmetric d -indexed array such that $a_{i_1, \dots, i_d} = 0$ whenever there exists $k \neq l$ such that $i_k = i_l$. Let X_1, \dots, X_n be independent random variables and $(X_i^{(j)})_{1 \leq i \leq n}$ for $j = 1, \dots, d$, be independent*

copies of the sequence $(X_i)_{1 \leq i \leq n}$, then for all $t \geq 0$,

$$\begin{aligned} & L_d^{-1} \Pr \left[\left| \sum_{i_1, \dots, i_d=1}^n a_{i_1, \dots, i_d} X_{i_1}^{(1)} \cdots X_{i_d}^{(d)} \right| \geq L_d t \right] \\ & \leq \Pr \left[\left| \sum_{i_1, \dots, i_d=1}^n a_{i_1, \dots, i_d} X_{i_1} \cdots X_{i_d} \right| \geq L_d t \right] \\ & \leq L_d \Pr \left[\left| \sum_{i_1, \dots, i_d=1}^n a_{i_1, \dots, i_d} X_{i_1}^{(1)} \cdots X_{i_d}^{(d)} \right| \geq L_d^{-1} t \right], \end{aligned}$$

where $L_d \in (0, \infty)$ depends only on d .

Essentially this theorem shows for a symmetric tensor with no “diagonal” terms, i.e., $a_{i_1, \dots, i_d} = 0$ whenever there exists $k \neq l$ such that $i_k = i_l$, there is only a constant factor difference between the coupled and decoupled Gaussian chaos distribution.

In most of our applications, we do have symmetric tensors with no “diagonal” terms. However there is one case where we do have diagonal terms, for which we need the following lemma.

Lemma 3.36. *Let $(a_{i_1, i_2, i_3})_{1 \leq i_1, \dots, i_3 \leq n}$ be a symmetric 3-indexed array and let $\|a\|_F$ denote its Frobenius norm. Let $X \sim \mathcal{N}(0, I_n)$, then for any $\epsilon > 0$, with probability at least $1 - Cn \exp(-C'n^{2\epsilon/3})$,*

$$\left| \sum_{i_1, i_2, i_3=1}^n a_{i_1, i_2, i_3} X_{i_1} X_{i_2} X_{i_3} \right| \leq 4\|a\|_F n^{0.5+\epsilon}.$$

Proof. The sum of the “diagonal” terms is equal to $3 \sum_{i \neq j} a_{i, i, j} X_i^2 X_j + 1/2 \sum_i a_{i, i, i} X_i^3$. Since X_i are independent standard Gaussian random variables, with probability at least $1 - Cn \exp(-C'n^{2\epsilon/3})$ (union bound), $|X_i| \leq n^{\epsilon/3}$ for all $i \in [n]$. Conditioned on

this high probability event, the absolute value of the sum is bounded by:

$$\begin{aligned}
\left| 3 \sum_{i \neq j} a_{i,i,j} X_i^2 X_j + \frac{1}{2} \sum_i a_{i,i,i} X_i^3 \right| &\leq 3 \sum_{i,j=1}^n |a_{i,i,j}| |X_j| X_i^2 \\
&\leq 3 \| (a_{i,i,j})_{1 \leq i,j \leq n} \|_1 n^\epsilon \\
&\leq 3 \sqrt{n} \| (a_{i,i,j})_{1 \leq i,j \leq n} \|_F n^\epsilon \\
&\leq 3 \|a\|_F n^{0.5+\epsilon}.
\end{aligned}$$

By Theorem 3.19, we know that with probability at least $1 - C \exp(-C' n^{2\epsilon/3})$, the absolute value of the sum of the “non-diagonal” terms is bounded by $\|a\|_F n^\epsilon$. Therefore we can conclude the proof by applying the union bound. \square

Chapter 4

Realization Problems of Hidden Markov Models

4.1 Problem Statement

Preliminaries on HMMs An HMM determines the joint probability distribution over sequences of hidden states $\{x_t : t \in \mathbb{Z}\}$ and observations $\{y_t : t \in \mathbb{Z}\}$. For simplicity, we call each output y_t as a “letter” taking value from some discrete alphabet $[d]$, and a sequence of n letters is referred to as a “string”, taking value from the Cartesian product $[d]^n$. We use $[d^N] \equiv \{1, \dots, d^N\}$ to denote the vectorized indices in $[d]^n$.

The joint distribution of $\{x_t, y_t : t \in \mathbb{Z}\}$ from a stationary HMM is parameterized by a pair of matrices: the state transition matrix $Q \in \mathbb{R}_+^{k \times k}$, and the observation matrix $O \in \mathbb{R}_+^{d \times k}$, which satisfy $\mathbf{e}^\top O = \mathbf{e}^\top$ and $\mathbf{e}^\top Q = \mathbf{e}^\top$, where \mathbf{e} is the all ones vector. The hidden state x_t evolves following a Markov process:

$$\mathbb{P}(x_{t+1} = j | x_t = i) = Q_{j,i}.$$

Let π denote the stationary state distribution, i.e., $\pi_i = \mathbb{P}[x_t = i]$ and $Q\pi = \pi$. Without loss of generality, we assume that $\pi_i > 0$ for all $i \in [k]$. We also define the

backward transition matrix $\tilde{Q} \in \mathbb{R}^{k \times k}$:

$$\mathbb{P}(x_{t-1} = j | x_t = i) = \tilde{Q}_{j,i}.$$

Observe that the matrix \tilde{Q} is related to Q as: $\tilde{Q} = \text{Diag}(\pi)Q^\top \text{Diag}(\pi)^{-1}$. Conditioned on the hidden state taking value i , the probability of observing letter j is:

$$\mathbb{P}(y_t = j | x_t = i) = O_{j,i}.$$

We call two HMMs equivalent if the output processes are statistically indistinguishable.

The order of the HMM is defined to be the number of hidden states, denoted by k . We will denote the class of all HMMs with output alphabet size d and order k by $\Theta_{(d,k)}^h$.

Realization problems Hidden Markov Models (HMMs) are widely used for describing discrete random processes, especially in the applications involving temporal pattern recognition such as speech and gesture recognition, part-of-speech tagging and parsing, and bioinformatics. The Markovian property of the hidden state evolution potentially leads to a low complexity representation of the output random process. In this work, we consider the long-standing HMM realization problem: given some partial knowledge about the output process of an unknown HMM, can we generalize it to a full description of the random process?

Consider a discrete random process $\{y_t : t \in \mathbb{Z}\}$, which assumes values in a finite alphabet $[d] \equiv \{1, \dots, d\}$. Assume that y_t is the output process of a stationary HMM of finite order. Let the random vector $\mathbf{y}_1^N = (y_1, \dots, y_N)$ denote an string of length N , which assumes values in the N -ary Cartesian product $[d]^N$. The process y_t is fully characterized by the joint probabilities of strings of any length in the countably infinite table (denoted by $\mathcal{P}^{(\infty)}$):

$$\left\{ \mathbb{P}(y_1 = l_1, \dots, y_N = l_N) : \forall \mathbf{l}_1^N \in [d]^N, \forall N \in \mathbb{Z} \right\}.$$

There are three main concerns in the realization problem:

1. **(Informational complexity)** Suppose that the underlying HMM is of order k , and we are given the joint probabilities of all the length N strings, namely:

$$\mathcal{P}^{(N)} \equiv \left\{ \mathbb{P}(y_1 = l_1, \dots, y_N = l_N) : \forall \mathbf{1}_1^N \in [d]^N \right\},$$

how large does N need to be so that we can compute $\mathcal{P}^{(\infty)}$ based on $\mathcal{P}^{(N)}$?

2. **(Computational complexity)** Can we solve the realization problem with runtime polynomial in the dimensions (alphabet size d and order of the underlying HMM k)?
3. **(Statistical complexity)** When $\mathcal{P}^{(N)}$ is estimated from sample sequences and has some estimation error, are the realization algorithms robust to the input errors?

These are long standing questions, and there are several lines of work within different communities at tempting to address these questions. It has long been known that, in the information theoretic sense, there exist hard cases of HMMs that are not efficiently PAC learnable [66] [89]. However, a more practical question is, can we efficiently solve the realization / learning problem for most HMMs? In this work, we focus on generic analysis and show that, for almost all HMMs, i.e., excluding those whose parameters are in a measure zero set ¹, the realization problems can be efficiently solved with poly time algorithms.

Minimal realization problems The *realization problem* takes as inputs the probabilities of finite length strings for a fixed window size N ($\mathcal{P}^{(N)}$), and finds a finite state model of the minimal order to describe the entire output process ($\mathcal{P}^{(\infty)}$). We aim to find the most succinct description of the process, namely the minimal order realization, where the “order” refers to the number of states of the underlying finite

¹ In our setting, algebraic genericity coincides with the measure theoretic notion of generic. Throughout the discussion, for fixed alphabet size d and order k , we call an HMM in general position if its transition and observation matrix are in general position, which is equivalent to “almost everywhere in the parameter space of $\{Q \in \mathbb{R}_+^{k \times k}, O \in \mathbb{R}_+^{d \times k} : \mathbf{e}^\top Q = \mathbf{e}^\top, \mathbf{e}^\top O = \mathbf{e}^\top\}$ ”.

state model. Without loss of generality, we assume that the process has a minimal realization of order k and examine under what conditions the algorithms can recover an equivalent minimal order realization.

Next, we introduce two classes of finite state models, both of which can realize an HMM output process.

Definition 4.1 (Quasi-HMM realization [123]). *Let θ° be a tuple: $\theta^\circ = (k, u, v \in \mathbb{R}^k, A^{(j)} \in \mathbb{R}^{k \times k} : \forall j \in [d])$. We call θ° a quasi-HMM realization of order k for a stationary process $\{y_t : t \in \mathbb{Z}\}$ if the three conditions hold: $(\forall \mathbf{1}_1^N \in [d]^N, \forall N \in \mathbb{Z})$*

$$\mathbb{P}(\mathbf{y}_1^N = \mathbf{1}_1^N) = u^\top A^{(l_1)} A^{(l_2)} \dots A^{(l_N)} v, \quad (4.1)$$

$$u^\top \left(\sum_{j=1}^d A^{(j)} \right) = u^\top, \quad (4.2)$$

$$\left(\sum_{j=1}^d A^{(j)} \right) v = v. \quad (4.3)$$

Definition 4.2 (Equivalent quasi-HMM realizations). *Two quasi-HMM realizations $\theta^\circ = (k, u, v, A^{(j)} : j \in [d])$ and $\tilde{\theta}^\circ = (k, \tilde{u}, \tilde{v}, \tilde{A}^{(j)} : j \in [d])$ are called equivalent, if there is a full rank matrix $T \in \mathbb{R}^{k \times k}$ such that:*

$$\tilde{u} = T^\top u, \quad \tilde{v} = T^{-1} v, \quad \tilde{A}^{(j)} = T^{-1} A^{(j)} T, \quad \forall j \in [d].$$

Definition 4.3 (HMM realization). *Let θ^h be a tuple: $\theta^h = (k, O \in \mathbb{R}_+^{d \times k}, Q \in \mathbb{R}_+^{k \times k})$. We call θ^h an HMM realization of order k for a stationary random process $\{y_t : t \in \mathbb{Z}\}$, if the matrices Q and O are column stochastic, and the output process of the HMM defined by the transition matrix Q and observation matrix O has the same distribution as y_t .*

HMM realizations are in a subset of the model class of quasi-HMM realizations. Given an HMM realization $\theta^h = (k, O, Q)$, one can construct the following quasi-

HMM realization $\theta^o = (k, u, v, A^{(j)} : j \in [d])$:

$$u = \mathbf{e}, \tag{4.4}$$

$$v = \pi, \tag{4.5}$$

$$A^{(j)} = Q\text{Diag}(O_{[j,1]}), \quad \forall j \in [d]. \tag{4.6}$$

The minimal (quasi-)HMM realization problem is formally stated below: Assume that the random process is the output of an HMM of order k . How large does the window size N need to be, so that given the joint probabilities $\mathcal{P}^{(N)}$ we can efficiently construct a minimal (quasi-)HMM realization for the process?

4.2 Main results

To study the HMM realization problems, we focus on algorithms based on the spectral method. The basic idea is to exploit the recursive structural properties of the underlying finite state model, and write the joint probabilities in $\mathcal{P}^{(N)}$ into a specific form which admits *rank decomposition*, where the rank reveals the minimal order of the realization and the model parameters can be extracted from the factors.

In the first part (Section 4.2.1), we consider the problem of finding the minimal quasi-HMM realization. Quasi-HMMs are associated with different names in different communities, for example finite state regular automata [17, 18], regular quasi realization [123, 89], and operator models [89, 58]. We mostly follow the terminologies in [123]. Algorithm 12 is the well-known algorithm for finding the minimal order quasi-HMM realization (to be rigorously defined later). However, in general the window size N can not be specified a priori and thus the complexity of the algorithm cannot be explicitly determined. In Theorem 4.2, we show that, if the output process is generated by an general position HMM with order k , we only need the window size N in the order of $\mathcal{O}(\log_d(k))$ for pinning down $\mathcal{P}^{(\infty)}$ based on $\mathcal{P}^{(N)}$, where d is the output alphabet size. Moreover, we show that Algorithm 12 has runtime and sample complexity both polynomial in the relevant parameters.

In the second part (Section 4.2.2), we consider the problem of finding the minimal HMM realization, using tensor decomposition methods, which rely on the uniqueness of tensor decomposition to recover the minimal order HMM that is unique up to hidden states permutation. Tensor decomposition based algorithms for learning HMMs are studied in [7, 5, 26]. In these works, the transition matrix is always assumed to be of full rank. Similar to that in the quasi-HMM realization problem, in general the required window size N and also the complexity of the algorithm cannot be determined a priori. In [5], the authors examined the generic identifiability conditions of HMM, and showed that generically it suffices to pick the window size $N = 2n + 1$ for some positive integer n , such that $\binom{n+d-1}{d-1} \geq k$. In the case where d is much smaller than k , n needs to be in the order of $\mathcal{O}(k^{1/d})$. Another bound on the window size N is given in [26], which is in the order of $\mathcal{O}(k/d)$. However, the size of the tensor in the decomposition is exponential in n , thus all these bound lead to runtime exponential in k .

In Section 4.2.2, we propose a two-step realization approach, and analyze the identifiability issue of the two steps. Then, we show that for the processes generated by almost all HMMs, the window size N only needs to be in the order of $\mathcal{O}(\log_d(k))$ for finding the minimal HMM realization. This means that for most HMMs, finding minimal quasi-HMM and minimal HMM realizations are actually of equal difficulty.

4.2.1 Minimal Quasi-HMM Realization

In this section, we address the minimal quasi-HMM realization problem. We first review the widely used algorithm [11, 18]; then we show for HMMs in general position, the window size N only needs to be in the order of $\mathcal{O}(\log_d(k))$ to guarantee the correctness of the algorithm; we also give an example of hard case (degenerate) which needs N to be as large as k ; finally we examine the stability of the algorithm.

Algorithm For notational convenience, we define the bijective mapping $L : [d]^n \rightarrow [d^n]$ which maps the multi-index $\mathbf{l}_1^N = (l_1, \dots, l_n) \in [d]^n$ to the index $L(\mathbf{l}_1^N) = (l_1 - 1)d^{n-1} + (l_2 - 1)d^{n-2} + \dots + l_n \in [d^n]$.

Algorithm 12: Minimal quasi-HMM realization

Input: $H^{(0)}, H^{(j)} \in \mathbb{R}^{d^n \times d^n}$ for all $j \in [d]$

Output: $\tilde{\theta}^o = (k, \tilde{u}, \tilde{v}, \tilde{A}^{(j)} : j \in [d])$

1. Compute the SVD of $H^{(0)}$: $H^{(0)} = U_H D_H V_H'$. Set $U = U_H D_H^{1/2}$, $V = V_H D_H^{1/2}$.
 2. Let \tilde{k} be the rank of $H^{(0)}$, and let $\tilde{u} = U'e$, $\tilde{v} = V'e$.
 3. Let U^\dagger and V^\dagger be the pseudo inverse of U and V . Set $\tilde{A}^{(j)} = U^\dagger H^{(j)} (V^\dagger)'$, $\forall j \in [d]$.
-

Given the length N joint probabilities $\mathcal{P}^{(N)}$, where $N = 2n + 1$ for some positive number n , we form two matrices $H^{(0)}, H^{(j)} \in \mathbb{R}^{d^n \times d^n}$ for all $j \in [d]$ as below:

$$[H^{(0)}]_{L(\mathbf{I}_1^n), L(\mathbf{I}_1^{-n})} = \mathbb{P}(\mathbf{y}_{-1}^{-n} = \mathbf{I}_1^{-n}, \mathbf{y}_0^{n-1} = \mathbf{I}_1^n), \quad (4.7)$$

$$[H^{(j)}]_{L(\mathbf{I}_1^n), L(\mathbf{I}_1^{-n})} = \mathbb{P}(\mathbf{y}_{-1}^{-n} = \mathbf{I}_1^{-n}, y_0 = j, \mathbf{y}_1^n = \mathbf{I}_1^n), \quad (4.8)$$

where $\mathbf{I}_1^n = (l_1, \dots, l_n)$ and $\mathbf{I}_1^{-n} = (l_{-1}, l_{-2}, \dots, l_{-n}) \in [d]^n$ denotes the length n string corresponding to the future and the past n time slots, respectively. Note that the “future” observations and the “past” observations are independent conditioned on the “current” state, which is the Markovian property that Algorithm 12 relies on.

The core idea of Algorithm 12 was discussed in [60], and it has been rediscovered numerous times in the literature in slightly different forms [11, 18]. We summarize the main idea below.

Remark 4.1 (Minimal order). *Let $\theta^o = (k, u, v, A^{(j)} : j \in [d])$ be a minimal quasi-HMM realization of order k for the process considered. Since the joint probabilities can be factorized in terms of the $A^{(j)}$'s as in (4.1), one can factorize $H^{(0)}$ and $H^{(j)}$'s as below:*

$$H^{(0)} = EF^\top, \quad H^{(j)} = EA^{(j)}F^\top,$$

where the matrices $E, F \in \mathbb{R}^{d^n \times k}$ are functions of θ^o . In particular, the $L(\mathbf{I}_1^n)$ -th row

of E and F are given by:

$$E_{[L(\mathbb{1}_1^*),:]} = u^\top (A^{(l_n)} \dots A^{(l_1)}), \quad (4.9)$$

$$F_{[L(\mathbb{1}_1^*),:]} = v^\top (A^{(l_n)} \dots A^{(l_1)})^\top. \quad (4.10)$$

Note that if both E and F have full column rank k , then $H^{(0)}$ has rank k , according to Sylvester's inequality. Any rank factorization leads to an equivalent minimal quasi-HMM realization of order k . The minimal order condition, though not explicitly enforced, is reflected in the rank factorization, as any quasi-HMM realization of lower order results in a matrix $H^{(0)}$ of lower rank, which leads to a contradiction.

The correctness of the algorithm crucially relies on matrix $H^{(0)}$ achieving its maximal rank k , which equals the order of the minimal realization. A necessary condition for the correctness of the algorithm is stated below.

Lemma 4.1 (Correctness of Algorithm 12). *Assume the process has a minimal quasi-realization θ° of order k . Algorithm 12 returns a minimal quasi-HMM realization $\tilde{\theta}^\circ$ that is equivalent to θ° , if the matrices E, F defined in (4.9) and (4.10) have full column rank k .*

Increasing the window size N can potentially boost the rank of $H^{(0)}$, in the hope that the $H^{(0)}$ reaches its maximal rank and Algorithm 12 can correctly find the minimal realization. However, for a given random process, the study of [106] showed that it is undecidable to verify whether it has a *finite order* quasi-HMM realization. Even under our assumption that the process indeed has an order k minimal quasi-HMM realization, it is still not clear how large the size of matrix $H^{(0)}$ ($d^n \times d^n$) needs to be so that it achieves the maximal rank k . In previous works, it was usually implicitly assumed that N is large enough so that $H^{(0)}$ achieves its maximal rank [18]. Yet without a bound on n or N the computational complexity of the algorithm is ambiguous.

Generic Analysis of Information Complexity We desire a small window size N while guaranteeing the full column rank of the matrices E and F defined in (4.9)

and (4.10). The following theorem shows that if the random process is generated by an order k HMM in general position, then we only need window size $N > 4\lceil\log_d(k)\rceil + 1$ to guarantee the correctness of Algorithm 12.

Theorem 4.2 (Window size N for quasi-HMM). *(1) Consider $\Theta_{(d,k)}^h$, the class of all HMMs with output alphabet size d and order k . There exists a measure zero set $\mathcal{E} \in \Theta_{(d,k)}^h$, such that for all the output process generated by HMMs in $\Theta_{(d,k)}^h \setminus \mathcal{E}$, Algorithm 12 returns a minimal quasi-HMM realization, if window size $N = 2n + 1$ for some n such that:*

$$n > 8\lceil\log_d(k)\rceil. \quad (4.11)$$

(2) For any pair of (d, k) , randomly pick an instance from the class $\Theta_{(d,k)}^h$. If for a given window size $N = 2n + 1$, the matrix $H^{(0)}$ achieves its maximal rank k , then for all HMMs in $\Theta_{(d,k)}^h$, excluding a measure zero set, N is sufficiently large for the correctness of Algorithm 12.

Since the elements of matrices E and F are polynomials of the parameters Q and O , in order to show E has full column rank for Q and O in general position, it suffices to construct an instance of HMM for which the matrix E has full column rank. In particular, we fix the transition matrix Q and randomize the observation matrix O and bound the singular values of E in probability. The detailed proof is provided in Appendix 4.3.

For all (d, k) pairs in the set $\{2 \leq d \leq k < 3000\}$, we implemented the test in Theorem 4.2 (2), and found that for all these cases $n = \lceil\log_d(k)\rceil$ is sufficient. We conjecture that in general, $n \geq \log_d(k)$ is enough.

In the worst case [123], the ‘‘Hankel rank’’ of the matrix $H^{(0)}$ with infinite window size can be larger than the rank of any finite size block of the infinite matrix. Instead of the worst case analysis, our generic analysis examines the average cases, and it has the following implications: if the process is generated by some average case HMM of order k , then the Hankel rank equals k ; moreover, the window size n only needs to be

in the order of $O(\log_a(k))$ so that the rank of finite matrix $H^{(0)}$ achieves the Hankel rank.

Existence of Hard Cases We showed that for generic HMM output processes, Algorithm 12 is has polynomial runtime. There exists a long line of hardness results for learning HMMs [66, 89, 114], showing that in the worst case (lie in the measure zero set in the parameter space) learning the distribution of an HMM *can* be computationally hard under cryptographic assumptions.

In Fig. 4-1, we adapt the hardness results to our setting and give an example to lower bound the worst case computational complexity. The state diagram describes the transition and observation probabilities. In the state transition diagram, for stage $t = 1, \dots, T - 1$, the emission state E_t is uniformly distributed over $\{0, 1\}$ and is observed. For stage $t = 2, \dots, T - 1$, the parity state S_t computes $E_{t-1} \oplus S_{t-1}$, except for at one unknown stage s , $S_t = S_{t-1}$. At stage T , with probability η , the correct parity state S_{T-1} is revealed, and with probability $1 - \eta$, the complement is observed. $(T + 1)$ is a reset stage, with probability ρ it stays in the reset stage. Solving the realization problem is equivalent to learning the joint distribution of the process. One can verify that the window size N needs to be at least as large as T , which is proportional to the order of the underlying HMM, and therefore the computation complexity is exponential in the order of the HMM.

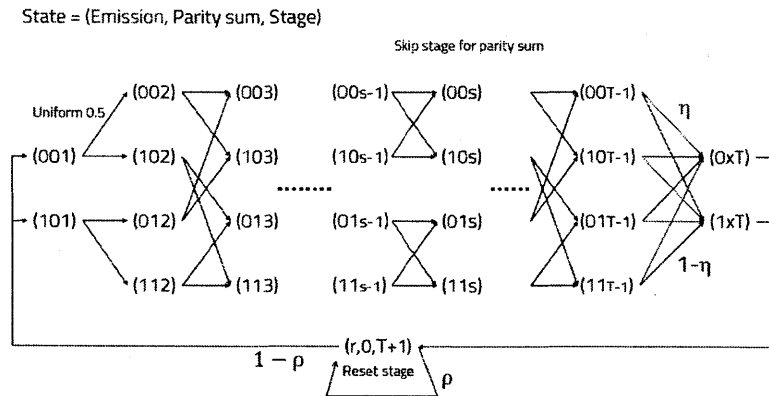


Figure 4-1: Reduction of HMM to noisy parity

We point out that not all HMMs in the measure zero set are information theo-

retically hard to learn. For instance, consider the degenerate HMM in [5] with the transition matrix $Q = I_{k \times k}$ and with general position observation matrix O . Suppose that $d \ll k$, it was shown that the window size N needs to be in the order of $k^{\frac{1}{d}}$ so that matrices E and F attain full column rank. However the distribution of this i.i.d. process is not fundamentally difficult to learn. It remains an open problem to find realization algorithm that can handle more cases.

Stability Analysis In practice, the joint probabilities in $\mathcal{P}^{(N)}$ are estimated based on finite sample sequences of the process. In the next theorem, we show that in order to achieve ϵ -accuracy in the parameters of the minimal quasi-HMM realization, the number of sample sequences we need to estimate $\mathcal{P}^{(N)}$ is polynomial in all relevant parameters, including the order k .

Theorem 4.3. *Given T independent sample sequences of the output process of an HMM of order k and with alphabet size d . Construct $\widehat{H}^{(0)}$ and $\widehat{H}^{(j)}$'s as in (4.7) and (4.8) with the empirical probabilities. Let $N = 2n + 1$, and $n = 2\lceil \log_d(k) \rceil$. Let $\widehat{\theta}^\circ = (k, \widehat{u}, \widehat{v}, \widehat{A}^{(j)} : j \in [d])$ and $\widetilde{\theta}^\circ = (k, \widetilde{u}, \widetilde{v}, \widetilde{A}^{(j)} : j \in [d])$ be the output of Algorithm 12 with the empirical probabilities and the exact probabilities for the input, respectively. Then, in order to achieve ϵ -accuracy in the output with probability at least $1 - \eta$, namely:*

$$\|\widehat{u} - \widetilde{u}\| \leq \epsilon, \|\widehat{v} - \widetilde{v}\| \leq \epsilon, \|\widehat{A}^{(j)} - \widetilde{A}^{(j)}\| \leq \epsilon, \forall j,$$

the number of independent sample sequences we need is given by:

$$T = \frac{Ck^6d^4}{\epsilon^4\sigma_k^8} \log\left(\frac{2k^4d^2}{\eta}\right),$$

where σ_k is the k -th singular value of $H^{(0)}$ and C is some absolute constant.

Since the core of the algorithm is singular value decomposition of the matrix $H^{(0)}$, the stability analysis mostly uses the standard matrix perturbation results. The detailed proof is provided in [59].

Remark 4.4. Note that Theorem 4.2 shows that for window size N large enough ($O(\log_d(k))$), the exact realization problem (no estimation noise) can be solved with poly time algorithm. When empirical probabilities are used, Theorem 4.3 shows that the required number of independent samples is polynomial in k , d , and $1/\sigma_k$. σ_k depends on the HMM that generates the process. In the proof of Theorem 4.2, it is showed that there exist cases for which σ_k is lower bounded by constant, for which case the sample complexity is indeed polynomial; however there also exists hard cases for which σ_k is arbitrarily small. We defer the analysis of sample complexity, which relies on understanding the relation between window size, HMM parameter, and σ_k , to future work.

4.2.2 Minimal HMM Realization Problem

Recall that an HMM can be easily converted to a quasi-HMM of the same order as shown in (4.4)–(4.6), yet given a quasi-HMM realization it is difficult to construct an HMM [11]. In this section, we apply tensor decomposition techniques to study the minimal HMM realization problem and discuss its connection to the previous section. In particular, we show that for processes generated by general position HMMs, the two realization problems have similar computational complexity.

Formulation For a fixed window size $N = 2n + 1$, given the exact joint probabilities in $\mathcal{P}^{(N)}$, similar to the construction of $H^{(0)}$ in (4.7), one can construct a 3rd order tensor $M \in \mathbb{R}^{d^n \times d^n \times d}$ as below:

$$M_{L(\mathbf{I}_1^n), L(\mathbf{I}_{-1}^n), l_0} = \mathbb{P}(\mathbf{y}_{-n}^n = \mathbf{I}_{-n}^n), \quad \forall \mathbf{I}_{-n}^n \in [d]^N. \quad (4.12)$$

Suppose that the process has a minimal HMM realization $\theta^h = (k, Q, O)$ of order k . We can write M as a tensor product:

$$M = A \otimes B \otimes C, \quad (4.13)$$

where the matrices $A, B \in \mathbb{R}^{d^n \times k}$ and $C \in \mathbb{R}^{d \times k}$ correspond to the conditional probabilities:

$$A_{L(\mathbf{l}_1^n), m} = \mathbb{P}\left(\mathbf{y}_1^n = \mathbf{l}_1^n \mid x_0 = m\right), \quad (4.14)$$

$$B_{L(\mathbf{l}_{-1}^{-n}), m} = \mathbb{P}\left(\mathbf{y}_{-1}^{-n} = \mathbf{l}_{-1}^{-n} \mid x_0 = m\right), \quad (4.15)$$

$$C_{l, m} = \mathbb{P}\left(y_0 = l, x_0 = m\right). \quad (4.16)$$

Moreover, observe that A and B are recursive linear functions of the model parameters Q and O as below:

$$A^{(n)} = \mathbb{P}\left(\mathbf{y}_1^n \mid x_0 = m\right) = (O \odot A^{(n-1)})Q, \quad (4.17)$$

$$B^{(n)} = \mathbb{P}\left(\mathbf{y}_{-1}^{-n} \mid x_0 = m\right) = (O \odot B^{(n-1)})\tilde{Q}, \quad (4.18)$$

and $A^{(1)} = OQ$ and $B^{(1)} = O\tilde{Q}$. In particular, for the given window size $N = 2n + 1$, we have:

$$A = A^{(n)}, \quad B = B^{(n)}, \quad C = O \text{Diag}(\pi). \quad (4.19)$$

The basic idea of *recovering* the minimal HMM realization θ^h (up to hidden state relabeling) is to first *recover* the factors A, B and C via tensor decomposition, and then extract the transition and observation probabilities from the factors. The minimal order condition is again reflected in the tensor *rank* factorization, as any HMM realization of lower order results in a tensor M of lower tensor rank, which is a contradiction.

Identifiability The identifiability of the minimal HMM relies on the fact that the tensor rank decomposition indeed recovers the factor A, B, C defined in (4.14)–(4.16). Note that by definition, the column stochastic observation matrix O must have Kruskal rank greater than 2, otherwise there exist two identical columns in O , and the corresponding two hidden states can be merged to give an equivalent HMM

realization of smaller order.

Lemma 4.2 (Uniqueness of tensor decomposition). *Given window size N , if the matrices $A, B \in \mathbb{R}^{d^n \times k}$ defined in (4.17)–(4.19) have full column rank k , then M can be uniquely decomposed into column stochastic matrices A, B, C as in (4.13) (up to common column permutation).*

In parallel with Theorem 4.2, the next theorem shows that the condition above is satisfied for a general position HMM process with sufficiently large window size N .

Theorem 4.5 (Choice of N for HMM realization). *Consider $\Theta_{(d,k)}^h$, the class of all HMMs with output alphabet size d and order k . There exists a measure zero set $\mathcal{E} \in \Theta_{(d,k)}^h$ such that for all output processes generated by HMMs in the set $\Theta_{(d,k)}^h \setminus \mathcal{E}$, the minimal quasi-HMM realization can be computed based on the joint probabilities in $\mathcal{P}^{(N)}$, if window size $N = 2n + 1$ for some n such that:*

$$n > 8 \lceil \log_d(k) \rceil. \quad (4.20)$$

Algorithms The matrices A, B and C , defined in (4.17)–(4.19), are polynomial functions of the parameters Q and O of the minimal HMM realization. The following theorem exploits the recursive structure of these polynomials to recover the parameters Q and O if the factors A, B, C are given.

Theorem 4.6 (Recovering Q and O from A, B, C). *Given the matrix C , one can obtain the observation matrix by:*

$$O_{[:,i]} = C_{[:,i]} / (\mathbf{e}^\top C_{[:,i]}), \quad \forall i \in [k]. \quad (4.21)$$

Given the matrix $A \in \mathbb{R}^{d^n \times k}$, we first scale each of the column similar to (4.21) so that each column is stochastic, and corresponds to the conditional probabilities $\mathbb{P}(\mathbf{y}_1^n | x_0)$ as shown in (4.14). We marginalize the conditional distribution to get $A^{(1)} = \mathbb{P}(y_1 | x_0) \in \mathbb{R}^{d \times k}$ and $A^{(n-1)} = \mathbb{P}(\mathbf{y}_1^{n-1} | x_0) \in \mathbb{R}^{d^{n-1} \times k}$.

(1) If A has full column rank k ([5]):

$$Q = \left(O \odot A^{(n-1)} \right)^\dagger A. \quad (4.22)$$

(2) If C has full column rank k :

$$Q = O^\dagger A^{(1)}. \quad (4.23)$$

where $(X)^\dagger = (X^\top X)^{-1} X^\top$ denotes the pseudo-inverse of a matrix X .

In the proof of Theorem 4.5, we show that for general position HMMs with sufficiently large window size, the matrices A and B achieve full column rank k . When this holds, Algorithm 1 computes the unique tensor decomposition to recover the factors A, B, C . Theorem 4.6 (1) applies to recover Q and O from the factors.

However, if the transition matrix Q of the minimal HMM realization does not have full rank, and no matter how large the window size is, the matrix A never achieves full rank. Note that these HMMs are degenerate cases belonging to the measure zero set in Theorem 4.5, and Algorithm 1 is not applicable for decomposing the tensor M . However, it is still possible to apply Algorithm 2. Note that a necessary condition for it to work is that $d \geq k$ and the observation matrix is of full column rank.

Let $\Theta_{(d,k,r)}^h$ denote the model class of HMMs with output alphabet d and order k , for $d \geq k$ and the transition matrix Q has rank $r < k$. Note that $\Theta_{(d,k,r)}^h$ is a subset of the measure zero set \mathcal{E} in Theorem 4.5. The following theorem shows that if Algorithm 2 runs correctly for a random instance in this subset, then the algorithm works for almost all HMMs in this subset.

Theorem 4.7 (Correctness of Algorithm 2). *Given d, k and r and consider the set $\Theta_{(d,k,r)}^h$. Let A, B, C be defined as in (4.17)–(4.19) for $n = 1$, and let $M = A \otimes B \otimes C$. If Algorithm 13 returns “yes”, then there exists a measure zero set $\mathcal{E} \in \Theta_{(d,k,r)}^h$, such that Algorithm 2 returns the tensor decomposition $M = A \otimes B \otimes C$ for all HMMs in the set $\Theta_{(d,k,r)}^h \setminus \mathcal{E}$. Moreover, if the latter is true, Algorithm 13 returns “yes” with probability 1.*

For this class of degenerate HMMs, Theorem 4.6 (2) applies to recover Q and O .

Note that for both the general position case and this degenerate case, the computation complexity to recover the parameters of the minimal HMM realization are polynomial in both d and k , and this is an immediate result of the log upper bound of the window size.

Algorithm 13: Check Condition

1. Randomly choose an HMM from $\theta^h \in \Theta_{(d,k,r)}^h$.
 2. Construct matrices A, B, C with (Q, O) as defined in (4.17)–(4.19) for $n = 1$, namely $A = OQ$, $B = O\tilde{Q}$, and $C = O\text{Diag}(\pi)$.
 3. Let $M = A \otimes B \otimes C$. Run Algorithm 2 with the input M .
 4. Return “yes” if the algorithm returns A, B, C uniquely up to a common column permutation, and “no” otherwise.
-

4.3 Proofs for Chapter 4

4.3.1 Proofs

Assume that the observed process has a minimal HMM realization θ^h of order k , i.e., $\theta^h \in \Theta_{(d,k)}^h$, and let θ^o denote the equivalent order k quasi-HMM as shown in (4.4)–(4.6). For window size $N = 2n + 1$, define the matrices E and F for θ^o as in (4.9) and (4.10) and note that:

$$\begin{aligned}
 & E_{L(\mathbf{l}_1^n), i} \\
 &= [u^\top (A^{(l_n)} \cdots A^{(l_1)})]_i \\
 &= \mathbf{e}^\top \mathbb{P}(x_n, y_{n-1} = l_n | x_{n-1}) \cdots \mathbb{P}(x_1, y_0 = l_1 | x_0 = i) \\
 &= \mathbb{P}(\mathbf{y}_0^{n-1} = \mathbf{l}_1^n | x_0 = i),
 \end{aligned}$$

and similarly,

$$F_{L(\mathbf{1}_1^n),i} = [A^{(l_n)} \dots A^{(l_1)} \pi]_i = \mathbb{P}\left(\mathbf{y}_{-1}^{-n} = \mathbf{1}_1^n, x_0 = i\right).$$

Lemma 4.1 shows that a sufficient condition for the correctness of Algorithm 12 is that both E and F have full column rank k . In this proof, we show that when Q and O of the HMM $\theta^h \in \Theta_{(d,k)}^h$ are in general position, this rank condition is satisfied if the window size $N = 2n + 1$ satisfies (4.11).

Note that the minors of E and F are polynomials in the elements of Q and O , thus it defines a algebraic set in the parameter space by setting all the minors to zero to make E and F to be rank deficient. By basic algebraic geometry [51], the algebraic set either occupies the entire Zariski closure or is a low-dimensional manifold of Lebesgue measure zero. In particular, the Zariski closure of $\Theta_{(d,k)}^h$, defined to be the smallest algebraic set containing $\Theta_{(d,k)}^h$, is given by $\bar{\Theta}_{(d,k)}^h := \{O \in \mathbb{R}^{d \times k}, Q \in \mathbb{R}^{k \times k} : \mathbf{e}^\top O = \mathbf{e}^\top, \mathbf{e}^\top Q = \mathbf{e}^\top\}$ (note that the element-wise non-negativity constraints can be omitted when considering the Zariski closure). Therefore, it is enough to show that for some specific choice of Q and O in $\bar{\Theta}_{(d,k)}^h$, the matrices E and F achieve full column rank k . Moreover to construct an instance, we can further ignore the stochastic constraints, as scaling does not the independence property of the columns in E and F .

We fix the transition matrix Q to be the state shifting matrix as below:

$$Q_{i-1,i} = 1, \text{ for } 2 \leq i \leq k, \text{ and } Q_{k,1} = 1, \quad (4.24)$$

Note that with this choice of Q , $\pi = \frac{1}{k}\mathbf{e}$, and $\tilde{Q} = Q^\top$. Due to the symmetry of the forward and backward transitions, we can focus on showing that E has full column rank and the same argument applies to F .

We randomize the observation matrix O and let the columns be independent random variables uniformly distributed on the d -dimensional sphere. In order to show that there exists a construction of (Q, O) such that E has full column rank, it suffices to show that E achieves full column rank with positive probability over the randomness of O . We apply Gershgorin's theorem to prove that the columns of E

are incoherent.

Note that for the shifting matrix Q , we have:

$$E_{[:,i]} = O_{[:,i]} \odot \cdots O_{[:,i+n-1]}.$$

Since we have $d \geq 2$ and $n < k$, for notational convenience, we slightly abuse notation to write the j -th column of O as $O_{[:,j]}$, while for $k < j \leq 2k$, it actually refer to the $(j - k)$ -th column of O .

Define matrix $X \in \mathbb{R}^{k \times k}$ to be:

$$X_{i,j} = E_{[:,i]}^\top E_{[:,j]} = \prod_{m=0}^{n-1} (O_{[:,i+m]}^\top O_{[:,j+m]}), \quad \forall i, j \in [k].$$

By the assumption that the columns of O are uniformly distributed on the d -dimensional sphere, we have $X_{i,i} = 1$, for all $i \in [k]$.

Fix some $\beta, \gamma = \beta^2 \in (0, 1)$. Suppose that, for any $i \neq j$,

$$\mathbb{P} \left(|X_{i,j}| < \frac{\beta}{k} \right) > 1 - \frac{\gamma}{k^2}. \quad (4.25)$$

Then apply union bound on j , we have for any i :

$$\begin{aligned} \mathbb{P} \left(\sum_{j \neq i}^k |X_{i,j}| < \beta \right) &\geq \mathbb{P} \left(\forall j \in [k], j \neq i, |X_{i,j}| < \frac{\beta}{k} \right) \\ &> 1 - \frac{\gamma}{k}. \end{aligned}$$

Again apply union bound on i , we have:

$$\mathbb{P} \left(\forall i \in [k], |X_{i,i}| - \sum_{j \neq i} |X_{i,j}| \geq 1 - \beta \right) > 1 - k \frac{\gamma}{k} = 1 - \gamma.$$

Apply Gershgorin's theorem, we have that with probability at least γ , the matrix $X = E^\top E$ is of full rank k , and the smallest singular value is at least $1 - \beta$. There must exist some instance of O such that this statement holds.

Next, we verify the statement in (4.25). Equivalently, we want to show that for $i \neq j$:

$$\begin{aligned}
1 - \frac{\gamma}{k^2} &< \mathbb{P} \left(\prod_{m=0}^{n-1} |O_{[:,i+m]}^\top O_{[:,j+m]}| < \frac{\beta}{k} \right) \\
&= \mathbb{P} \left(\sum_{m=0}^{n-1} \log(|O_{[:,i+m]}^\top O_{[:,j+m]}|) < -\log\left(\frac{k}{\beta}\right) \right) \\
&= \mathbb{P} \left(\sum_{m=0}^{n-1} \log \left(\frac{1}{|O_{[:,i+m]}^\top O_{[:,j+m]}|} \right) > \log\left(\frac{k}{\beta}\right) \right) \\
&= \mathbb{P} \left(\sum_{m=1}^n \log \left(\frac{1}{|v_m|} \right) > \log\left(\frac{k}{\beta}\right) \right)
\end{aligned}$$

where v_m are i.i.d. random variables with the distribution as the projection of a uniform unit-norm vector in \mathbb{R}^d onto the first dimension. The last equality is due to the independence of the columns of O .

Define the indicator random variable s_m for $m \in [n]$:

$$s_m = \mathbf{1} \left[\log\left(\frac{1}{|v_m|}\right) < \frac{1}{c} \log(d) \right] = \mathbf{1} \left[|v_m| > \frac{1}{d^{1/c}} \right],$$

where we pick constant $c = 4$. Assume that $d \geq 2 + (8e)^2$ (as we really only care about the scaling), apply Johnson Lindenstrauss lemma, setting u_1 to be v_m and t to be $1/d^{1/c}$, we have:

$$\mu = \mathbb{P}(s_m = 1) < \frac{4}{\sqrt{d-2}} e^{-\frac{d-2}{2d^{2/c}}} < \frac{1}{2e} e^{-\frac{d-2}{2d^{2/c}}}$$

Note that by definition:

$$\sum_{m=1}^n \log \left(\frac{1}{|v_m|} \right) > \sum_{m=1}^n \frac{1}{c} \log(d)(1 - s_m).$$

Therefore it suffices to show that

$$1 - \frac{\gamma}{k^2} < \mathbb{P} \left(\sum_{m=1}^n \frac{1}{c} \log(d)(1 - s_m) > \log\left(\frac{k}{\beta}\right) \right),$$

or equivalently,

$$\begin{aligned} \frac{\gamma}{k^2} &> \mathbb{P} \left(\sum_{m=1}^n s_m > n - c \frac{\log(k/\beta)}{\log(d)} \right) \\ &= \mathbb{P} \left(\sum_{m=1}^n s_m > \alpha c \frac{\log(k/\beta)}{\log(d)} \right), \end{aligned}$$

where we set $n = (1 + \alpha)c \log_d(k/\beta)$ for some $\alpha > 1$.

Apply the multiplicative Chernoff bound, by setting $X_m = s_m$ for $m = 1, \dots, n$, and set $\delta n \mu = \alpha c \frac{\log(k/\beta)}{\log(d)}$, and $\frac{\varepsilon}{\delta} = \frac{\varepsilon n \mu}{\alpha c \frac{\log(k/\beta)}{\log(d)}} = \frac{1+\alpha}{\alpha} e \mu < e^{-\sqrt{d}/2} < 1$, we have

$$\mathbb{P} \left(\sum_{m=1}^n s_m > \alpha c \frac{\log(k/\beta)}{\log(d)} \right) < \left(\frac{1 + \alpha}{\alpha} e \mu \right)^{\alpha c \frac{\log(k/\beta)}{\log(d)}}.$$

We want to show that the RHS is less than γ/k^2 . Taking log, this is equivalent to:

$$\alpha c \frac{\log(k/\beta)}{\log(d)} \log_d \left(\frac{\alpha}{(1 + \alpha)e\mu} \right) > \frac{\log(k^2/\gamma)}{\log(d)}$$

Recall that we have $\gamma = \beta^2$, $\frac{1+\alpha}{\alpha} e \mu \leq e^{-\frac{d-2}{2d^{2/c}}}$, $c = 4$ the above inequality holds if we pick $\alpha = 4/c = 1$, as

$$\alpha c \log_d \left(\frac{\alpha}{(1 + \alpha)e\mu} \right) \geq 4 \frac{\log(e^{\frac{\sqrt{d}}{2}})}{\log(d)} \geq 2 \frac{\sqrt{d}}{\log(d)} \geq 2.$$

Now we can conclude that (4.25) holds. □

4.3.2 Other proofs

(Proof of Theorem 4.3)

Recall that the output of Algorithm 12 is given by:

$$\begin{aligned}\widehat{A}^{(j)} &= \widehat{D}^{-1/2} \widehat{U}_H^\top \widehat{H}^{(j)} \widehat{V}_H \widehat{D}^{1/2}, \\ \widehat{u} &= \widehat{D}^{-1/2} \widehat{U}_H^\top \mathbf{e}, \quad \widehat{v} = \widehat{D}^{-1/2} \widehat{V}_H^\top \mathbf{e},\end{aligned}$$

where \widehat{U}_H and \widehat{V}_H are the first k left and right singular vectors of $\widehat{H}^{(0)}$, and the diagonal matrix \widehat{D} has the first k singular values of $\widehat{H}^{(0)}$ on its main diagonal. In order to bound the distance between $\widehat{A}^{(j)}$ and $\widetilde{A}^{(j)}$, \widehat{u} and \widetilde{u} , \widehat{v} and \widetilde{v} , we analyze the perturbation bound for each of the factor separately and apply Lemma 1.6 to bound the overall perturbation of the product form.

First, denote $E_j = \widehat{H}^{(j)} - H^{(j)}$ for $j = 0, 1, \dots, d$. For any element in E_j we can bound its norm using Hoeffding's inequality (Lemma 1.8): with probability at least $1 - 2e^{-2T\delta^2}$, the (i_1, i_2) -th element of E_j is bounded by: $\|[E_j]_{i_1, i_2}\| \leq \delta < 1$. Moreover, apply union bound to j and all elements in each E_j , with probability at least $1 - 2k^4 d^3 e^{-2T\delta^2}$, for all $j = 0, 1, \dots, d$, we have

$$\|E_j\|_F \leq \sqrt{k d^n} \delta < k^{1.5} d^{0.5} \delta,$$

where the last inequality is due to $d^n < k^2 d$.

Second, we apply the matrix perturbation bound (Lemma ??) to bound the distance of the singular vectors:

$$\|\widehat{U}_H - U_H\| \leq \frac{\sqrt{2}\|E_0\|_F}{\sigma_k(H^{(0)})}, \quad \|\widehat{V}_H - V_H\| \leq \frac{\sqrt{2}\|E_0\|_F}{\sigma_k(H^{(0)})}.$$

And we can apply Mirsky's theorem (Lemma ??) to bound the distance of the singular values:

$$\|\widehat{D} - D\| \leq \|E_0\|_F.$$

Denote $\Delta_i = \sigma_i(\widehat{H}^{(0)}) - \sigma_i(H^{(0)})$ and let $\sigma_i = \sigma_i(H^{(0)})$. Note that if $\|E_0\| \leq \sigma_k/2$, we

have that for any $i = 1, \dots, k$, $|\Delta_i| \leq \|E_0\| \leq \sigma_i/2$, then

$$\begin{aligned} \left(\frac{1}{\sqrt{\sigma_i}} - \frac{1}{\sqrt{\sigma_i + \Delta_i}}\right)^2 &= \frac{1}{\sigma_i + \Delta_i} (\sqrt{1 + \Delta_i/\sigma_i} - 1)^2 \\ &\leq \frac{2}{\sigma_i} (\Delta_i/\sigma_i + 2 - 2\sqrt{1 + \Delta_i/\sigma_i}) \\ &\leq \frac{2}{\sigma_i} (3|\Delta_i|/\sigma_i) \\ &\leq \frac{6}{\sigma_k^2} |\Delta_i|, \end{aligned}$$

where the first inequality is due to $|\Delta_i| \leq \delta_i/2$, and the second inequality is due to $\sqrt{1 + \Delta_i/\sigma_i} \geq 1 - |\Delta_i/\sigma_i|$. Therefore we have that

$$\|\widehat{D}^{-1/2} - D^{-1/2}\| \leq \frac{\sqrt{6 \sum_{i=1}^k |\Delta_i|}}{\sigma_k} \leq \frac{\sqrt{6\sqrt{k}} \|\widehat{D} - D\|}{\sigma_k}.$$

Finally, we apply Lemma 1.6 to bound the output perturbation. Note that $\|D^{-1/2}\| = 1/\sqrt{\sigma_k}$, $\|U_H\| = 1$, $\|V_H\| = 1$. Moreover note that the probabilities in each row of $H^{(j)}$ sum up to less than 1, therefore by Perron-Frobenius theorem we have $\|H^{(j)}\| \leq 1$. Therefore we have

$$\begin{aligned} &\|\widehat{A}^{(j)} - \widetilde{A}^{(j)}\| \\ &\leq 2^4 \left(\frac{2\sqrt{6k^{1/2}} \|E_0\|_F}{\sigma_k^{1.5}} + \frac{2\sqrt{2} \|E_0\|_F}{\sigma_k^2} + \frac{\|E_j\|}{\sigma_k} \right) \\ &\leq 2^4 \left(\frac{2\sqrt{6} k^{0.75} d^{0.25} \delta^{0.5}}{\sigma_k^{1.5}} + \frac{2\sqrt{2} k d^{0.5} \delta}{\sigma_k^2} + \frac{k d^{0.5} \delta}{\sigma_k} \right) \\ &\leq \frac{144 k d^{0.5}}{\sigma_k^2} \delta^{0.5}, \end{aligned}$$

where the first inequality is due to $\|E_j\| \leq \|E_j\|_F$, and the second inequality is due to $\delta < 1$ and $\sigma_k \leq \sigma_1 \leq 1$.

Similarly we can bound $\|\widehat{u} - \widetilde{u}\|$ and $\|\widehat{v} - \widetilde{v}\|$ by:

$$\|\widehat{u} - \widetilde{u}\| \leq \|\widehat{D}^{-1/2} \widehat{U}_H^\top - D^{-1/2} U_H^\top\| \sqrt{d^n} \leq \frac{4k^{1.5} d}{\sigma_k^{1.5}} \delta^{0.5}.$$

In summary, if we want to achieve ϵ accuracy in the output, we need δ to be no larger than $\epsilon^2 \sigma_k^4 / (144k^3 d^2)$. Set the failure probability to be $\eta = 2k^4 d^3 e^{-2T\delta^2}$, then number of sample sequences needed to estimate the empirical probabilities is given by:

$$T = 2 \frac{144^2 k^6 d^4}{\epsilon^4 \sigma_k^8} \log \left(\frac{2k^4 d^3}{\eta} \right).$$

□

(Proof of Theorem 4.5)

With exactly the same argument and constructional proof as for Theorem 4.2, we can show that for the window size $N = 2n + 1$ satisfies (4.20), the matrices A and B have full column rank. By Lemma 4.2 we have that the tensor decomposition of M is unique. Moreover, by the argument in Theorem 4.6 (1), we have that the model parameters Q, O can be uniquely recovered from the factors A, B, C . Thus in conclusion $\mathcal{P}^{(N)}$ is sufficient for finding the minimal HMM realization.

□

(Proof of Theorem 4.6)

By the uniqueness of tensor decomposition (up to column permutation and scaling) the columns of C are proportional to the columns of O (up to some hidden state permutation), and each column of O must satisfy the normalization constraint: $\mathbf{e}^\top O_{[:,i]} = 1, \forall i \in [k]$. The normalization in (4.21) recovers O from C .

Recall that

$$A = A^{(n)} = (O \odot A^{(n-1)})Q.$$

Since the matrix A has full column rank k , the matrices $Q \in \mathbb{R}^{k \times k}$ and $(O \odot A^{(n-1)}) \in \mathbb{R}^{d^n \times k}$ both have full column rank k , as well as the pseudo-inverse of $(O \odot \tilde{A})$, therefore $Q = (O \odot A^{(n-1)})^\dagger A$.

By definition we have $A^{(1)} = OQ$, thus if O is of full column rank k , we can obtain $Q = O^\dagger A^{(1)}$.

□

(Proof of Theorem 4.7)

Denote the minimal order HMM realization by $\theta^h = (k, Q, O)$, and since $n = 1$, the matrices are given by:

$$A = OQ, \quad B = O\tilde{Q}, \quad C = O\text{Diag}(\pi).$$

Define two linear operators $I_{d^2 \times d^2} : \mathbb{R}^{d^2} \rightarrow \mathbb{R}^{d^2}$ and $P_{d^2 \times d^2} : \mathbb{R}^{d^2} \rightarrow \mathbb{R}^{d^2}$, such that for any matrix $X \in \mathbb{R}^{d \times d}$: $I_{d^2 \times d^2} \text{vec}(X) = \text{vec}(X)$ and $P_{d^2 \times d^2} \text{vec}(X) = \text{vec}(X^\top)$. Moreover, define matrix $R \in \mathbb{R}^{d^2 \times d^2}$ and $G \in \mathbb{R}^{d^4 \times d^2}$ to be:

$$R = I_{d^2 \times d^2} - P_{d^2 \times d^2}, \quad G = R \odot R.$$

Note that the kernel of $(I_{d^2 \times d^2} - P_{d^2 \times d^2})$ is the space of symmetric matrices, thus R is of rank $d^2 - d(d+1)/2 = d(d-1)/2$, and G is of rank $d^2(d-1)^2/4$. Define matrix $G^\perp \in \mathbb{R}^{d^4 \times (d^4 - \frac{d^2(d-1)^2}{4})}$ such that its columns are orthogonal to the columns of G .

According to [41, 42, 63], there are two deterministic conditions for Algorithm 2 to correctly recover the factors A, B, C from the rank k tensor M :

1. Both $A \odot B$ and C have full column rank k .
2. Define $T \in \mathbb{R}^{d^4 \times (m + (k-1)k/2)}$ to be:

$$T = \left[\begin{array}{l} G^\perp_{[:,i]} : 1 \leq i \leq d^4 - \frac{d^2(d-1)^2}{4}, \\ A_{[:,k_1]} \odot A_{[:,k_2]} \odot B_{[:,k_1]} \odot B_{[:,k_2]} : 1 \leq k_1 < k_2 \leq k \end{array} \right].$$

The columns of T are linear independent.

Parameterize the rank r transition matrix by $Q = UV^\top$ for some matrices $U, V \in \mathbb{R}^{k \times r}$. Define the parameter space \mathcal{Q} :

$$\mathcal{Q} = \{Q \in \mathbb{R}^{k \times k} : Q = UV^\top, U, V \in \mathbb{R}^{k \times r}, \mathbf{e}^\top Q = \mathbf{e}^\top\}$$

Note that by construction, the minors of $A \odot B$ and T are nonzero polynomials in

the elements of the parameters U, V and O , in order to show that the two deterministic rank conditions are satisfied for almost all instances in the class $\Theta_{(d,k,r)}^h$, it is enough to construct an instance in the model class that satisfies the two conditions (by the random check in Algorithm 13). Moreover, if it is true, then with probability one, the two conditions are satisfied for a randomly chosen instance in the model class.

□

(Proof of Lemma 4.1)

If both E and F have full column rank k , by Sylvester inequality the rank of the matrix $H^{(0)}$ is also equal to k , the order of minimal quasi-HMM realization. Therefore, for the two matrices U and V obtained in Step 2 in Algorithm 1, there exists some full rank matrix $W \in \mathbb{R}^{k \times k}$ such that:

$$U = EW, \quad V^\top = W^{-1}F^\top.$$

Therefore, Step 3 returns

$$\tilde{A}^{(j)} = W^{-1}E^\dagger EA^{(j)}F^\top (F^\top)^\dagger W = W^{-1}A^{(j)}W.$$

By the normalization constraint in Definition 4.1, we have

$$u^\top W = u^\top \sum_{j=1}^d A^{(j)}W = u^\top W \sum_{j=1}^d \tilde{A}^{(j)}.$$

Moreover, since

$$U = \begin{bmatrix} u^\top(A^{(1)} \dots A^{(1)}) \\ u^\top(A^{(1)} \dots A^{(2)}) \\ \vdots \\ u^\top(A^{(d)} \dots A^{(d)}) \end{bmatrix} W = u^\top W \begin{bmatrix} \tilde{A}^{(1)} \dots \tilde{A}^{(1)} \\ \tilde{A}^{(1)} \dots \tilde{A}^{(2)} \\ \vdots \\ \tilde{A}^{(d)} \dots \tilde{A}^{(d)} \end{bmatrix},$$

in Step 2 we obtain $\tilde{u}^\top = u^\top W$, and similarly, we can argue that $\tilde{v} = W^{-1}v$. Thus we conclude that the output $\tilde{\theta}^o = (k, \tilde{u}, \tilde{v}, \tilde{A}^{(j)} : j \in [d])$ is a valid minimal quasi-HMM

realization of order k , and is equivalent to θ^o up to a linear transformation.

□

Chapter 5

Super-resolution

5.1 Problem Statement

5.1.1 Formulation

We follow the standard mathematical abstraction of this problem (Candes & Fernandez-Granda [31, 30]): consider a d -dimensional signal $x(t)$ modeled as a weighted sum of k Dirac measures in \mathbb{R}^d :

$$x(t) = \sum_{j=1}^k w_j \delta_{\mu^{(j)}}, \quad (5.1)$$

where the point sources, the $\mu^{(j)}$'s, are in \mathbb{R}^d . Assume that the weights w_j are complex valued, whose absolute values are lower and upper bounded by some positive constant. Assume that we are given k , the number of point sources¹.

Define the measurement function $f(s) : \mathbb{R}^d \rightarrow \mathbb{C}$ to be the convolution of the point source $x(t)$ with a low-pass point spread function $e^{i\pi\langle s, t \rangle}$ as below:

$$f(s) = \int_{t \in \mathbb{R}^d} e^{i\pi\langle t, s \rangle} x(dt) = \sum_{j=1}^k w_j e^{i\pi\langle \mu^{(j)}, s \rangle}. \quad (5.2)$$

In the noisy setting, the measurements are corrupted by uniformly bounded pertur-

¹An upper bound of the number of point sources suffices.

bation z :

$$\tilde{f}(s) = f(s) + z(s), \quad |z(s)| \leq \epsilon_z, \forall s. \quad (5.3)$$

Suppose that we are only allowed to measure the signal $x(t)$ by evaluating the measurement function $\tilde{f}(s)$ at any $s \in \mathbb{R}^d$, and we want to recover the parameters of the point source signal, i.e., $\{w_j, \mu^{(j)} : j \in [k]\}$. We follow the standard normalization to assume that:

$$\mu^{(j)} \in [-1, +1]^d, \quad |w_j| \in [0, 1] \quad \forall j \in [k].$$

Let $w_{\min} = \min_j |w_j|$ denote the minimal weight, and let Δ be the minimal separation of the point sources defined as follows:

$$\Delta = \min_{j \neq j'} \|\mu^{(j)} - \mu^{(j')}\|_2, \quad (5.4)$$

where we use the Euclidean distance between the point sources for ease of exposition². These quantities are key parameters in our algorithm and analysis. Intuitively, the recovery problem is harder if the minimal separation Δ is small and the minimal weight w_{\min} is small.

The first question is that, given exact measurements, namely $\epsilon_z = 0$, where and how many measurements should we take so that the original signal $x(t)$ can be exactly recovered.

Definition 5.1 (Exact recovery). *In the exact case, i.e. $\epsilon_z = 0$, we say that an algorithm achieves exact recovery with m measurements of the signal $x(t)$ if, upon input of these m measurements, the algorithm returns the exact set of parameters $\{w_j, \mu^{(j)} : j \in [k]\}$.*

Moreover, we want the algorithm to be measurement noise tolerant, in the sense that in the presence of measurement noise we can still recover good estimates of the point sources.

²Our claims hold without using the “wrap around metric”, as in [31, 30], due to our random sampling. Also, it is possible to extend these results for the ℓ_p -norm case.

Definition 5.2 (Stable recovery). *In the noisy case, i.e., $\epsilon_z \geq 0$, we say that an algorithm achieves stable recovery with m measurements of the signal $x(t)$ if, upon input of these m measurements, the algorithm returns estimates $\{\widehat{w}_j, \widehat{\mu}^{(j)} : j \in [k]\}$ such that*

$$\min_{\pi} \max \{ \|\widehat{\mu}^{(j)} - \mu^{(\pi(j))}\|_2 : j \in [k] \} \leq \text{poly}(d, k) \epsilon_z,$$

where the min is over permutations π on $[k]$ and $\text{poly}(d, k)$ is a polynomial function in d and k .

By definition, if an algorithm achieves stable recovery with m measurements, it also achieves exact recovery with these m measurements.

The terminology of “super-resolution” is appropriate due to the following remarkable result (in the noiseless case) of Donoho [43]: suppose we want to accurately recover the point sources to an error of γ , where $\gamma \ll \Delta$. Naively, we may expect to require measurements whose frequency depends inversely on the desired accuracy γ . Donoho [43] showed that it suffices to obtain a finite number of measurements, whose frequencies are bounded by $O(1/\Delta)$, in order to achieve *exact* recovery; thus resolving the point sources far more accurately than that which is naively implied by using frequencies of $O(1/\Delta)$. Furthermore, the work of Candes & Fernandez-Granda [31, 30] showed that stable recovery, in the univariate case ($d = 1$), is achievable with a cutoff frequency of $O(1/\Delta)$ using a convex program and a number of measurements whose size is polynomial in the relevant quantities.

5.1.2 Related Work

We are interested in stable recovery procedures with the following desirable statistical and computational properties: we seek to use coarse (low frequency) measurements; we hope to take a (quantifiably) small number of measurements; we desire our algorithm run quickly. Informally, our main result is as follows:

Theorem 5.1 (Informal statement of Theorem 5.3). *For a fixed probability of error, the proposed algorithm achieves stable recovery with a number of measurements and*

	$d = 1$			$d \geq 1$		
	cutoff freq	measurements	runtime	cutoff freq	measurements	runtime
SDP	$\frac{1}{\Delta}$	$k \log(k) \log(\frac{1}{\Delta})$	$poly(\frac{1}{\Delta}, k)$	$\frac{C_d}{\Delta_\infty}$	$(\frac{1}{\Delta_\infty})^d$	$poly((\frac{1}{\Delta_\infty})^d, k)$
MP	$\frac{1}{\Delta}$	$\frac{1}{\Delta}$	$(\frac{1}{\Delta})^3$	-	-	-
Ours	$\frac{1}{\Delta}$	$(k \log(k))^2$	$(k \log(k))^2$	$\frac{\log(kd)}{\Delta}$	$(k \log(k) + d)^2$	$(k \log(k) + d)^2$

Table 5.1: See Section 5.1.2 for description. See Lemma 5.1 for details about the cutoff frequency. Here, we are implicitly using $O(\cdot)$ notation.

with computational runtime that are both on the order of $O((k \log(k) + d)^2)$. Furthermore, the algorithm makes measurements which are bounded in frequency by $O(1/\Delta)$ (ignoring log factors).

Notably, our algorithm and analysis directly deal with the multivariate case, with the univariate case as a special case. Importantly, the number of measurements and the computational runtime do *not* depend on the minimal separation of the point sources. This may be important even in certain low dimensional imaging applications where taking physical measurements are costly (indeed, super-resolution is important in settings where Δ is small). Furthermore, our technical contribution of how to decompose a certain tensor constructed with Fourier measurements may be of broader interest to related questions in statistics, signal processing, and machine learning.

Table 5.1 summarizes the comparisons between our algorithm and the existing results. The multi-dimensional cutoff frequency we refer to in the table is the maximal coordinate-wise entry of any measurement frequency s (i.e. $\|s\|_\infty$). “SDP” refers to the semidefinite programming (SDP) based algorithms of Candes & Fernandez-Granda [30, 31]; in the univariate case, the number of measurements can be reduced by the method in Tang et. al. [112] (this is reflected in the table). “MP” refers to the matrix pencil type of methods, studied in [80] and [84] for the univariate case. Here, we are defining the infinity norm separation as $\Delta_\infty = \min_{j \neq j'} \|\mu^{(j)} - \mu^{(j')}\|_\infty$, which is understood as the wrap around distance on the unit circle. $C_d \geq 1$ is a problem dependent constant (discussed below).

Observe the following differences between our algorithm and prior work:

- 1) Our minimal separation is measured under the ℓ_2 -norm instead of the infinity norm, as in the SDP based algorithm. Note that Δ_∞ depends on the coordinate system; in the worst case, it can underestimate the separation by a $1/\sqrt{d}$ factor, namely $\Delta_\infty \sim \Delta/\sqrt{d}$.
- 2) The computation complexity and number of measurements are polynomial in dimension d and the number of point sources k , and surprisingly do not depend on the minimal separation of the point sources! Intuitively, when the minimal separation between the point sources is small, the problem should be harder, this is only reflected in the sampling range and the cutoff frequency of the measurements in our algorithm.
- 3) Furthermore, one could project the multivariate signal to the coordinates and solve multiple univariate problems (such as in [96, 91], which provided only exact recovery results). Naive random projections would lead to a cutoff frequency of $O(\sqrt{d}/\Delta)$.

SDP approaches: The work in [30, 31, 46] formulates the recovery problem as a total-variation minimization problem; they then show the dual problem can be formulated as an SDP. They focused on the analysis of $d = 1$ and only explicitly extend the proofs for $d = 2$. For $d \geq 1$, Ingham-type theorems (see [102, 71]) suggest that $C_d = O(\sqrt{d})$.

The number of measurements can be reduced by the method in [112] for the $d = 1$ case, which is noted in the table. Their method uses sampling “off the grid”; technically, their sampling scheme is actually sampling random points from the grid, though with far fewer measurements.

Matrix pencil approaches: The matrix pencil method, MUSIC and Prony’s method are essentially the same underlying idea, executed in different ways. The original Prony’s method directly attempts to find roots of a high degree polynomial, where the root stability has few guarantees. Other methods aim to robustify the algorithm.

Recently, for the univariate matrix pencil method, Liao & Fannjiang [80] and Moitra [84] provide a stability analysis of the MUSIC algorithm. Moitra [84] studied the optimal relationship between the cutoff frequency and Δ , showing that if the cutoff frequency is less than $1/\Delta$, then stable recovery is not possible with matrix pencil method (with high probability).

5.2 Main Results

5.2.1 Warm-up

1-D case: revisiting the matrix pencil method Let us first review the matrix pencil method for the univariate case, which stability was recently rigorously analyzed in Liao & Fannjiang [80] and Moitra [84].

A square matrix H is called a *Hankel* matrix if its skew-diagonals are constants, namely $H_{i,j} = H_{i-1,j+1}$. For some positive constants $m \in \mathbb{Z}$, sample to get the measurements $f(s)$ evaluated at the sampling set $\mathcal{S}_3 = \{0, 1, \dots, 2m\}$, and construct two Hankel matrices $H_0, H_1 \in \mathbb{C}^{m \times m}$:

$$H_0 = \begin{bmatrix} f(0) & f(1) & \dots & f(m-1) \\ f(1) & f(2) & \dots & f(m) \\ \vdots & & & \vdots \\ f(m-1) & f(m) & \dots & f(2m-1) \end{bmatrix}, \quad H_1 = \begin{bmatrix} f(1) & f(2) & \dots & f(m) \\ f(2) & f(3) & \dots & f(m+1) \\ \vdots & & & \vdots \\ f(m) & f(m+1) & \dots & f(2m) \end{bmatrix}. \quad (5.5)$$

Define $D_w \in \mathbb{C}_{diag}^{k \times k}$ to be the diagonal matrix with the weights on the main diagonal: $[D_w]_{j,j} = w_j$. Define $D_\mu \in \mathbb{C}_{diag}^{k \times k}$ to be $[D_\mu]_{j,j} = e^{i\pi\mu(j)}$.

A matrix V is called a *Vandermonde matrix* if each column is a geometric pro-

gression. defined the Vandermonde matrix $V_m \in \mathbb{C}^{m \times k}$ as below:

$$V_m = \begin{bmatrix} 1 & \dots & 1 \\ (e^{i\pi\mu^{(1)}})^1 & \dots & (e^{i\pi\mu^{(k)}})^1 \\ \vdots & & \vdots \\ (e^{i\pi\mu^{(1)}})^{m-1} & \dots & (e^{i\pi\mu^{(k)}})^{m-1} \end{bmatrix}. \quad (5.6)$$

The two Hankel matrices H_0 and H_1 admit the following simultaneous diagonalization:

$$H_0 = V_m D_w V_m^\top, \quad H_1 = V_m D_w D_\mu V_m^\top. \quad (5.7)$$

As long as V_m is of full rank, this simultaneous diagonalization can be computed by solving the generalized eigenvalue problem, and the parameters of the point source can thus be obtained from the factor V_m and D_w .

The univariate matrix pencil method only needs $m \geq k$ to achieve exact recovery. In the noisy case, the stability of generalized eigenvalue problem depends on the condition number of the Vandermonde matrix V_m and the minimal weight w_{min} .

Since all the nodes ($e^{i\pi\mu^{(j)}}$'s) of this Vandermonde matrix lie on the unit circle in the complex plane, it is straightforward to see that asymptotically $\lim_{m \rightarrow \infty} \text{cond}_2(V_m) = 1$. Furthermore, for $m > 1/\Delta$, [80, 84] showed that $\text{cond}_2(V_m)$ is upper bounded by a constant that does not depend on k and m . This bound on condition number is also implicitly discussed in [96].

Another way to view the matrix pencil method is that it corresponds to the low rank 3rd order tensor decomposition (see for example [8]). This view will help us generalize matrix pencil method to higher dimension d in a direct way, without projecting the signal on each coordinate and apply the univariate algorithm multiple times. For $m \geq k$, construct a 3rd order tensor $F \in \mathbb{C}^{m \times m \times 2}$ with elements of H_0 and H_1 defined in (5.5) as:

$$F_{i,i',j} = [H_{j-1}]_{i,i'}, \quad \forall j \in [2], i, i' \in [m].$$

Note that the two slices along the 3rd dimension of F are H_0 and H_1 . Namely $F(I, I, e_1) = H_0$, and $F(I, I, e_2) = H_1$. Recall the matrix decomposition of H_0 and H_1 in (5.7). Since $m \geq k$ and the $\mu^{(j)}$'s are distinct, we know that F has the *unique* rank k tensor decomposition:

$$F = V_m \otimes V_m \otimes (V_2 D_w).$$

Given the tensor F , the basic idea of the well-known Jennrich's algorithm ([55, 77]) for finding the unique low rank tensor decomposition is to consider two random projections $v_1, v_2 \in \mathbb{R}^m$, and then with high probability the two matrices $F(I, I, v_1)$ and $F(I, I, v_2)$ admit simultaneous diagonalization. Therefore, the matrix pencil method is indeed a special case of Jennrich's algorithm by setting $v_1 = e_1$ and $v_2 = e_2$

The multivariate case: a toy example One could naively extend the matrix pencil method to higher dimensions by using taking measurements from a hyper-grid, which is of size exponential in the dimension d . We now examine a toy problem which suggests that the high dimensional case may not be inherently more difficult than the univariate case.

The key ideas is that an appropriately sampled set can significantly reduce the number of measurements (as compared to using all the grid points). Tang et al [112] made a similar observation for the univariate case. They used a small random subset of measurements (actually still from the grid points) and showed that this contains enough information to recover all the measurement on the grid; the full measurements were then used for stably recovering the point sources.

Consider the case where the dimension $d \geq k$. Assume that w_j 's are real valued, and for all $j \in [k]$ and $n \in [d]$, the parameters $\mu_n^{(j)}$ are i.i.d. and uniformly distributed over $[-1, +1]$. This essentially corresponds to the standard (L_2) incoherence conditions (for the $\mu^{(j)}$'s).³ The following simple algorithm achieves stability with

³ This setting is different from the 2-norm separation condition. To see the difference, note that the toy algorithm does not work for constant shift $\mu^{(1)} = \mu^{(2)} + \Delta$. This issue is resolved in the general algorithm, when the condition is stated in terms of 2-norm separation.

polynomial complexity.

First, take d^3 number of measurements by evaluating $f(s)$ in the set $\mathcal{S}_3 = \{s = e_{n_1} + e_{n_2} + e_{n_3} : [n_1, n_2, n_3] \in [d] \times [d] \times [d]\}$, noting that \mathcal{S}_3 contains only a subset of d^3 points from the grid of $[3]^d$. Then, construct a 3rd order tensor $F \in \mathbb{C}^{d \times d \times d}$ with the measurements in the following way:

$$F_{n_1, n_2, n_3} = f(s)|_{s=e_{n_1}+e_{n_2}+e_{n_3}}, \quad \forall n_1, n_2, n_3 \in [d].$$

Note that we have the measurement

$$f(e_1 + e_2 + e_3) = \sum_{j=1}^k w_j e^{i\pi(\mu_1^{(j)} + \mu_2^{(j)} + \mu_3^{(j)})} = \sum_{j=1}^k w_j e^{i\pi\mu_1^{(j)}} e^{i\pi\mu_2^{(j)}} e^{i\pi\mu_3^{(j)}}.$$

It is straightforward to verify that F has a rank- k tensor factorization $F = V_d \otimes V_d \otimes (V_d D_w)$, where the factor $V_d \in \mathbb{R}^{d \times k}$ is given by:

$$V_d = \begin{bmatrix} e^{i\pi\mu_1^{(1)}} & \dots & e^{i\pi\mu_1^{(k)}} \\ e^{i\pi\mu_2^{(1)}} & \dots & e^{i\pi\mu_2^{(k)}} \\ \vdots & \dots & \vdots \\ e^{i\pi\mu_d^{(1)}} & \dots & e^{i\pi\mu_d^{(k)}} \end{bmatrix}. \quad (5.8)$$

Under the distribution assumption of the point sources, the entries $e^{i\pi\mu_n^{(j)}}$ are i.i.d. and uniformly distributed over the unit circle on the complex plane. Therefore almost surely the factor V_d has full column rank, and thus the tensor decomposition is unique. Moreover here w_j 's are real and each element of V_S has unit norm, we have a rescaling constraint with the tensor decomposition, with which we can uniquely obtain the factor V_S and the weights in D_w . By taking element-wise log of V_S we can read off the parameters of the point sources from V_S directly. Moreover, with high probability, we have that $\text{cond}_2(V_d)$ concentrates around 1, thus the simple algorithm achieves stable recovery.

Algorithm 14: General algorithm

Input: R, m , noisy measurement function $\tilde{f}(\cdot)$.

Output: Estimates $\{\hat{w}_j, \hat{\mu}^{(j)} : j \in [k]\}$.

1. Take measurements:

Let $\mathcal{S} = \{s^{(1)}, \dots, s^{(m)}\}$ be m i.i.d. samples from the Gaussian distribution $\mathcal{N}(0, R^2 I_{d \times d})$. Set $s^{(m+n)} = e_n$ for all $n \in [d]$ and $s^{(m+n+1)} = 0$. Denote $m' = m + d + 1$.

Take another random samples v from the unit sphere, and set $v^{(1)} = v$ and $v^{(2)} = 2v$. Construct a tensor $\tilde{F} \in \mathbb{C}^{m' \times m' \times 3}$: $\tilde{F}_{n_1, n_2, n_3} = \tilde{f}(s)|_{s=s^{(n_1)}+s^{(n_2)}+v^{(n_3)}}$.

2. Tensor Decomposition: Set $(\hat{V}_{S'}, \hat{D}_w) = \text{TensorDecomp}(\tilde{F})$.

For $j = 1, \dots, k$, set $[\hat{V}_{S'}]_j = [\hat{V}_{S'}]_j / [\hat{V}_{S'}]_{m', j}$

3. Read of estimates: For $j = 1, \dots, k$, set $\hat{\mu}^{(j)} = \text{Real}(\log([\hat{V}_{S'}]_{[m+1:m+d, j]})) / (i\pi)$.

4. Set $\hat{W} = \arg \min_{W \in \mathbb{C}^k} \|\hat{F} - \hat{V}_{S'} \otimes \hat{V}_{S'} \otimes \hat{V}_d D_w\|_F$.

5.2.2 Our Algorithm

We briefly describe the steps of Algorithm 14 below:

(Take measurements) Given positive numbers m and R , randomly draw a sampling set $\mathcal{S} = \{s^{(1)}, \dots, s^{(m)}\}$ of m i.i.d. samples of the Gaussian distribution $\mathcal{N}(0, R^2 I_{d \times d})$. Form the set $\mathcal{S}' = \mathcal{S} \cup \{s^{(m+1)} = e_1, \dots, s^{(m+d)} = e_d, s^{(m+d+1)} = 0\} \subset \mathbb{R}^d$. Denote $m' = m + d + 1$. Take another independent random sample v from the unit sphere, and define $v^{(1)} = v$, $v^{(2)} = 2v$. Construct the 3rd order tensor $\tilde{F} \in \mathbb{C}^{m' \times m' \times 3}$ with noise corrupted measurements $\tilde{f}(s)$ evaluated at the points in $\mathcal{S}' \oplus \mathcal{S}' \oplus \{v^{(1)}, v^{(2)}\}$, arranged in the following way:

$$\tilde{F}_{n_1, n_2, n_3} = \tilde{f}(s)|_{s=s^{(n_1)}+s^{(n_2)}+v^{(n_3)}}, \forall n_1, n_2 \in [m'], n_3 \in [2]. \quad (5.9)$$

(Tensor decomposition) Define the *characteristic matrix* V_S to be:

$$V_S = \begin{bmatrix} e^{i\pi\langle\mu^{(1)},s^{(1)}\rangle} & \dots & e^{i\pi\langle\mu^{(k)},s^{(1)}\rangle} \\ e^{i\pi\langle\mu^{(1)},s^{(2)}\rangle} & \dots & e^{i\pi\langle\mu^{(k)},s^{(2)}\rangle} \\ \vdots & \dots & \vdots \\ e^{i\pi\langle\mu^{(1)},s^{(m)}\rangle} & \dots & e^{i\pi\langle\mu^{(k)},s^{(m)}\rangle} \end{bmatrix}. \quad (5.10)$$

and define matrix $V' \in \mathbb{C}^{m' \times k}$ to be

$$V_{S'} = \begin{bmatrix} V_S \\ V_d \\ 1, \dots, 1 \end{bmatrix}, \quad (5.11)$$

where $V_d \in \mathbb{C}^{d \times k}$ is defined in (5.8). Define

$$V_2 = \begin{bmatrix} e^{i\pi\langle\mu^{(1)},v^{(1)}\rangle} & \dots & e^{i\pi\langle\mu^{(k)},v^{(1)}\rangle} \\ e^{i\pi\langle\mu^{(1)},v^{(2)}\rangle} & \dots & e^{i\pi\langle\mu^{(k)},v^{(2)}\rangle} \\ 1 & \dots & 1 \end{bmatrix}.$$

Note that in the exact case ($\epsilon_z = 0$) the tensor F constructed in (5.9) admits a rank- k decomposition:

$$F = V_{S'} \otimes V_{S'} \otimes (V_2 D_w), \quad (5.12)$$

Assume that $V_{S'}$ has full column rank, then this tensor decomposition is unique up to column permutation and rescaling with very high probability over the randomness of the random unit vector v . Since each element of $V_{S'}$ has unit norm, and we know that the last row of $V_{S'}$ and the last row of V_2 are all ones, there exists a proper scaling so that we can uniquely recover w_j 's and columns of $V_{S'}$ up to common permutation.

Here we adopt Jennrich's algorithm (see Algorithm 15) for tensor decomposition. Other algorithms, for example tensor power method ([8]) and recursive projection

Algorithm 15: TensorDecomp

Input: Tensor $\tilde{F} \in \mathbb{C}^{m \times m \times 3}$, rank k .

output: Factor $\hat{V} \in \mathbb{C}^{m \times k}$.

1. Compute the truncated SVD of $\tilde{F}(I, I, e_1) = \hat{P}\hat{\Lambda}\hat{P}^\top$ with the k leading singular values.
 2. Set $\hat{E} = \tilde{F}(\hat{P}, \hat{P}, I)$. Set $\hat{E}_1 = \hat{E}(I, I, e_1)$ and $\hat{E}_2 = \hat{E}(I, I, e_2)$.
 3. Let the columns of \hat{U} be the eigenvectors of $\hat{E}_1\hat{E}_2^{-1}$ corresponding to the k eigenvalues with the largest absolute value.
 4. Set $\hat{V} = \sqrt{m}\hat{P}\hat{U}$.
-

([122]), which are possibly more stable than Jennrich's algorithm, can also be applied here.

(Read off estimates) Let $\log(V_d)$ denote the element-wise logarithm of V_d . The estimates of the point sources are given by:

$$[\mu^{(1)}, \mu^{(2)}, \dots, \mu^{(k)}] = \frac{\log(V_d)}{i\pi}.$$

Remark 5.2. *In the toy example, the simple algorithm corresponds to using the sampling set $\mathcal{S}' = \{e_1, \dots, e_d\}$. The conventional univariate matrix pencil method corresponds to using the sampling set $\mathcal{S}' = \{0, 1, \dots, m\}$ and the set of measurements $\mathcal{S}' \oplus \mathcal{S}' \oplus \mathcal{S}'$ corresponds to the grid $[m]^3$.*

5.2.3 Performance Guarantees

In this section, we discuss how to pick the two parameters m and R and prove that the proposed algorithm indeed achieves stable recovery in the presence of measurement noise.

Theorem 5.3 (Stable recovery). *There exists a universal constant C such that the following holds.*

Fix $\epsilon_x, \delta_s, \delta_v \in (0, \frac{1}{2})$;

pick m such that $m \geq \max \left\{ \frac{k}{\epsilon_x} \sqrt{8 \log \frac{k}{\delta_s}}, d \right\}$;

for $d = 1$, pick $R \geq \frac{\sqrt{2 \log(1+2/\epsilon_x)}}{\pi \Delta}$; for $d \geq 2$, pick $R \geq \frac{\sqrt{2 \log(k/\epsilon_x)}}{\pi \Delta}$.

Assume the bounded measurement noise model as in (5.3) and that $\epsilon_z \leq \frac{\Delta \delta_v w_{min}^2}{100 \sqrt{dk^5}} \left(\frac{1-2\epsilon_x}{1+2\epsilon_x} \right)^{2.5}$.

With probability at least $(1 - \delta_s)$ over the random sampling of \mathcal{S} , and with probability at least $(1 - \delta_v)$ over the random projections in Algorithm 15, the proposed Algorithm 14 returns an estimation of the point source signal $\hat{x}(t) = \sum_{j=1}^k \hat{w}_j \hat{\delta}_{\mu^{(j)}}$ with accuracy:

$$\min_{\pi} \max \left\{ \|\hat{\mu}^{(\pi(j))} - \mu^{(\pi(j))}\|_2 : j \in [k] \right\} \leq C \frac{\sqrt{dk^5} w_{max}}{\Delta \delta_v w_{min}^2} \left(\frac{1+2\epsilon_x}{1-2\epsilon_x} \right)^{2.5} \epsilon_z,$$

where the min is over permutations π on $[k]$. Moreover, the proposed algorithm has time complexity in the order of $O((m')^3)$.

The next lemma shows that essentially, with overwhelming probability, all the frequencies taken concentrate within the hyper-cube with cutoff frequency R' on each coordinate, where R' is comparable to R ,

Lemma 5.1 (The cutoff frequency). *For $d > 1$, with high probability, all of the $2(m')^2$ sampling frequencies in $\mathcal{S}' \oplus \mathcal{S}' \oplus \{v^{(1)}, v^{(2)}\}$ satisfy that $\|s^{(j_1)} + s^{(j_2)} + v^{(j_3)}\|_{\infty} \leq R'$, $\forall j_1, j_2 \in [m], j_3 \in [2]$, where the per-coordinate cutoff frequency is given by $R' = O(R\sqrt{\log md})$.*

For $d = 1$ case, the cutoff frequency R' can be made to be in the order of $R' = O(1/\Delta)$.

Remark 5.4 (Failure probability). *Overall, the failure probability consists of two pieces: δ_v for random projection of v , and δ_s for random sampling to ensure the bounded condition number of V_S . This may be boosed to arbitrarily high probability through repetition.*

5.2.4 Key Lemmas

Stability of tensor decomposition: In this paragraph, we give a brief description

and the stability guarantee of the well-known Jennrich's algorithm ([55, 77]) for low rank 3rd order tensor decomposition. We only state it for the symmetric tensors as appeared in the proposed algorithm.

Consider a tensor $F = V \otimes V \otimes (V_2 D_w) \in \mathbb{C}^{m \times m \times 3}$ where the factor V has full column rank k . Then the decomposition is unique up to column permutation and rescaling, and Algorithm 15 finds the factors efficiently. Moreover, the eigen-decomposition is stable if the factor V is well-conditioned and the eigenvalues of $F_a F_b^\dagger$ are well separated.

Lemma 5.2 (Stability of Jennrich's algorithm). *Consider the 3rd order tensor $F = V \otimes V \otimes (V_2 D_w) \in \mathbb{C}^{m \times m \times 3}$ of rank $k \leq m$, constructed as in Step 1 in Algorithm 1.*

Given a tensor \tilde{F} that is element-wise close to F , namely for all $n_1, n_2, n_3 \in [m]$, $|\tilde{F}_{n_1, n_2, n_3} - F_{n_1, n_2, n_3}| \leq \epsilon_z$, and assume that the noise is small $\epsilon_z \leq \frac{\Delta \delta_v w_{\min}^2}{100 \sqrt{dk} w_{\max} \text{cond}_2(V)^5}$. Use \tilde{F} as the input to Algorithm 15. With probability at least $(1 - \delta_v)$ over the random projections $v^{(1)}$ and $v^{(2)}$, we can bound the distance between columns of the output \hat{V} and that of V by:

$$\min_{\pi} \max_j \left\{ \|\hat{V}_j - V_{\pi(j)}\|_2 : j \in [k] \right\} \leq C \frac{\sqrt{dk}^2 w_{\max}}{\Delta \delta_v w_{\min}^2} \text{cond}_2(V)^5 \epsilon_z, \quad (5.13)$$

where C is a universal constant.

Condition number of $V_{S'}$: The following lemma is helpful:

Lemma 5.3. *Let $V_{S'} \in \mathbb{C}^{(m+d+1) \times k}$ be the factor as defined in (5.11). Recall that $V_{S'} = [V_S; V_d; 1]$, where V_d is defined in (5.8), and V_S is the characteristic matrix defined in (5.10).*

We can bound the condition number of $V_{S'}$ by

$$\text{cond}_2(V_{S'}) \leq \sqrt{1 + \sqrt{k} \text{cond}_2(V_S)}. \quad (5.14)$$

Condition number of the characteristic matrix V_S : Therefore, the stability analysis of the proposed algorithm boils down to understanding the relation between

the random sampling set \mathcal{S} and the condition number of the characteristic matrix V_S . This is analyzed in Lemma 5.5 (main technical lemma).

Lemma 5.4. *For any fixed number $\epsilon_x \in (0, 1/2)$. Consider a Gaussian vector s with distribution $\mathcal{N}(0, R^2 I_{d \times d})$, where $R \geq \frac{\sqrt{2 \log(k/\epsilon_x)}}{\pi \Delta}$ for $d \geq 2$, and $R \geq \frac{\sqrt{2 \log(1+2/\epsilon_x)}}{\pi \Delta}$ for $d = 1$. Define the Hermitian random matrix $X_s \in \mathbb{C}_{herm}^{k \times k}$ to be*

$$X_s = \begin{bmatrix} e^{-i\pi \langle \mu^{(1)}, s \rangle} \\ e^{-i\pi \langle \mu^{(2)}, s \rangle} \\ \vdots \\ e^{-i\pi \langle \mu^{(k)}, s \rangle} \end{bmatrix} \begin{bmatrix} e^{i\pi \langle \mu^{(1)}, s \rangle}, e^{i\pi \langle \mu^{(2)}, s \rangle}, \dots, e^{i\pi \langle \mu^{(k)}, s \rangle} \end{bmatrix}. \quad (5.15)$$

We can bound the spectrum of $\mathbb{E}_s[X_s]$ by:

$$(1 - \epsilon_x) I_{k \times k} \preceq \mathbb{E}_s[X_s] \preceq (1 + \epsilon_x) I_{k \times k}. \quad (5.16)$$

Lemma 5.5 (Main technical lemma). *In the same setting of Lemma 5.4, Let $\mathcal{S} = \{s^{(1)}, \dots, s^{(m)}\}$ be m independent samples of the Gaussian vector s . For $m \geq \frac{k}{\epsilon_x} \sqrt{8 \log \frac{k}{\delta_s}}$, with probability at least $1 - \delta_s$ over the random sampling, the condition number of the factor V_S is bounded by:*

$$\text{cond}_2(V_S) \leq \sqrt{\frac{1 + 2\epsilon_x}{1 - 2\epsilon_x}}. \quad (5.17)$$

5.3 Discussions

5.3.1 Numerical results

We empirically demonstrate the performance of the proposed super-resolution algorithm in this section.

First, we look at a simple instance with dimension $d = 2$ and the minimal separation $\Delta = 0.05$. Our perturbation analysis of the stability result limits to small noise, i.e. ϵ_z is inverse polynomially small in the dimensions, and the number of measure-

ments m needs to be polynomially large in the dimensions. However, we believe these are only the artifact of the crude analysis, instead of being intrinsic to the approach. In the following numerical example, we examine a typical instance of 8 randomly generated 2-D point sources. The minimal separation Δ is set to be 0.01, and the weights are uniformly distributed in $[0.1, 1.1]$. The measurement noise level ϵ_z is set to be 0.1, and we take only 2178 noisy measurements ($\ll 1/\Delta^2$). Figure 5-1 shows the recovery result. The xy plane shows the coordinates of the point sources: true point sources (cyan), the two closest points (blue), and the estimated points (red); the z axis shows the corresponding mixing weights.

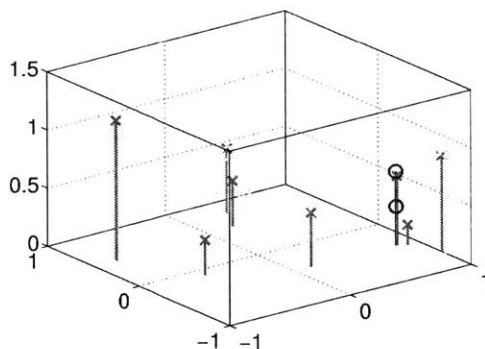


Figure 5-1: Simulation result for 2-D super-resolution

Next, we examine the phase transition properties implied by the main theorem.

Figure 5-2 shows the dependency between the cutoff frequency and the minimal separation. For each fixed pair of the minimal separation and the cutoff frequency (Δ, R) , we randomly generate $k = 8$ point sources in 4-dimensional space while maintaining the same minimal separation. The weights are uniformly distributed in $[0.1, 1.1]$. The recovery is considered successful if the error $\sum_{j \in [k]} \sqrt{\|\widehat{\mu}^{(j)} - \mu^{(j)}\|_2^2} \leq 0.1$ (on average it tolerates around 4% error per coordinate per point source). This process is repeated 50 times and the rate of success was recorded. Figure 5-2 plots the success rate in gray-scale, where 0 is black and 1 is white.

We observe that there is a sharp phase transition characterized by a linear relation between the cutoff frequency and the inverse of minimal separation, which is implied by Theorem 5.3.

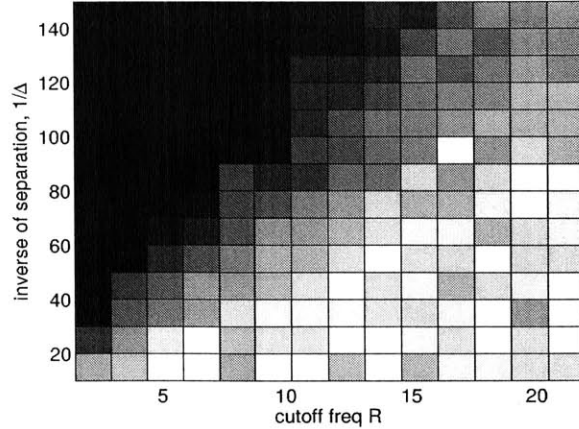


Figure 5-2: Cutoff frequency versus the required minimal separation

In a similar setup, we examine the success rate while varying the minimal separation Δ and the number of measurement m .

Fix dimension $d = 4$, number of point sources $k = 8$, and the measurement noise level $\epsilon_z = 0.03$. We vary the minimal separation such that Δ ranges from 0.01 to 0.2, and we use the corresponding cutoff frequency $R = \frac{0.26}{\Delta}$. We also vary the number of measurements m from 4 to 64. For each pair of (Δ, m) we randomly generate k point sources and run the proposed algorithm to recover the point sources. The recovery is considered successful if the error $\sum_{j \in [k]} \sqrt{\|\hat{\mu}^{(j)} - \mu^{(j)}\|_2^2} \leq 0.1$. This process is repeated 50 times and the rate of success was recorded.

In Figure 5-3, we observe that there is a threshold of m below which the number of measurements is too small to achieve stable recovery; when m is above the threshold, the success rate increases with the number of measurements as the algorithm becomes more stable. However, note that given the appropriately chosen cutoff frequency R , the number of measurements required does not depend on the minimal separation, and thus the computation complexity does not depend on the minimal separation neither.

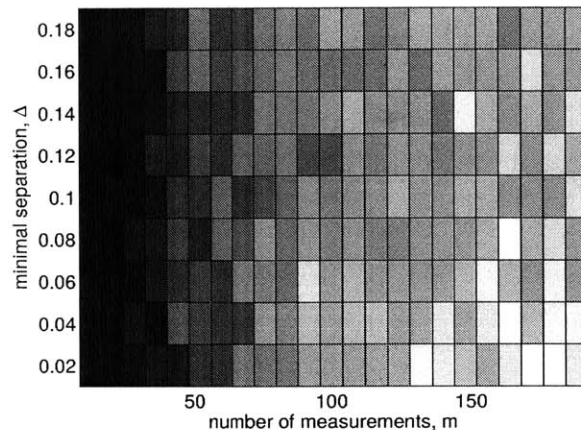


Figure 5-3: Number of measurements versus the required minimal separation

5.3.2 Connection with learning GMMs

One reason we are interested in the scaling of the algorithm with respect to the dimension d is that it naturally leads to an algorithm for learning Gaussian mixture models (GMMs).

Recall the problem of learning GMMs: given a number of N i.i.d. samples coming from a random one out of k Gaussian distributions in d dimensional space, the learning problem asks to estimate the means and the covariance matrices of these Gaussian components, as well as the mixing weights. We denote the parameters by $\{(w_j, \mu^{(j)}, \Sigma^{(j)})\}_{i \in [k]}$ where the mean vectors $\mu^{(j)} \in [-1, +1]^d$, the covariance matrices $\Sigma^{(j)} \in \mathbb{R}^{d \times d}$ and the mixing weights $w_j \in \mathbb{R}_+$.

In this brief discussion, we only consider the case where the components are spherical Gaussians with common covariance matrices, namely $\Sigma^{(j)} = \sigma^2 I_{d \times d}$ for all j . Moreover, we define the separation Δ_G by:

$$\Delta_G = \frac{\min_{j \neq j'} \|\mu^{(j)} - \mu^{(j')}\|_2}{\sigma},$$

and we will focus on the well-separated case where Δ_G is sufficiently large. This class of well-separated GMMs is often used in data clustering.

By the law of large numbers, for large d , the probability mass of a d -dimensional

Gaussian distribution tightly concentrates within a thin shell with a $\sqrt{d}\sigma$ distance from the mean vector. This concentration of distance leads to a line of works of provably learning GMMs in the well-separated case, started by the seminal work of Dasgupta[37] (spherical and identical Σ , $\Delta_G \geq \Omega(d^{1/2})$, complexity $poly(d, k)$) and followed by works of Dasgupta & Schulman [39] (spherical and identical Σ , $d \gg \log(k)$, $\Delta_G \geq \Omega(d^{1/4})$, complexity $poly(d, k)$), Arora & Kannan [103] (general and identical Σ , $\Delta_G \geq \Omega(d^{1/4})$ complexity $O(k^d)$).

Instead of relying on the concentration of distance and use distance based clustering to learn the GMM, we observe that in the well-separated case the characteristic function of the GMM has nice properties, and one can exploit the concentration of the characteristic function to learn the parameters. Note that we do not impose any other assumption on the dimensions k and d .

Next, we sketch the basic idea of applying the proposed super-resolution algorithm to learn well-separated GMMs, guaranteeing that N the required number of samples from the GMM, as well as the computation complexity both are in the order of $poly(d, k)$. Since σ is a bounded scalar parameter, we can simply apply grid-search to find the best match. In the following we assume that the σ is given and focus on learning the mean vectors and the mixing weights.

Evaluate the characteristic function of a d dimensional Gaussian mixture X , with identical and spherical covariance matrix $\Sigma = \sigma^2 I_{d \times d}$, at $s \in \mathbb{R}^d$:

$$\phi_X(s) = \mathbb{E}[e^{i\langle x, s \rangle}] = \sum_{j \in [k]} w_j e^{-\frac{1}{2}\sigma^2 \|s\|_2^2 + i\langle \mu^{(j)}, s \rangle}.$$

Also we let $\hat{\phi}_X(s)$ denote the empirical characteristic function evaluated at s based on N i.i.d. samples $\{x_1, \dots, x_N\}$ drawn from this GMM:

$$\hat{\phi}_X(s) = \frac{1}{N} \sum_{l \in [N]} e^{i\langle x_l, s \rangle}.$$

Note that $|e^{i\langle x_l, s \rangle}| = 1$ for all samples, thus we can apply Bernstein concentration inequality to the characteristic function and argue that $|\hat{\phi}_X(s) - \phi_X(s)| \leq O(\frac{1}{\sqrt{N}})$ for

all s .

In order to apply the proposed super-resolution algorithm, define

$$f(s) = e^{\frac{1}{2}\sigma^2\pi^2\|s\|_2^2}\phi_X(\pi s) = \sum_{j \in [k]} w_j e^{i\pi\langle \mu^{(j)}, s \rangle}, \quad \text{and} \quad \tilde{f}(s) = e^{\frac{1}{2}\pi^2\sigma^2\|s\|_2^2}\widehat{\phi}_X(s).$$

In the context of learning GMM, taking measurements of $\tilde{f}(s)$ corresponding to evaluating the empirical characteristic function at different s , for $\|s\|_\infty \leq R$, where R is the cutoff frequency. Note that this implies $\|s\|_2^2 \leq dR^2$. Therefore, we have that with high probability the noise level ϵ_z can be bounded by

$$\epsilon_z = \max_{\|s\|_\infty \leq R} |f(s) - \tilde{f}(s)| = O\left(\frac{e^{\sigma^2 d R^2}}{\sqrt{N}}\right).$$

In order to achieve stable recovery of the mean vector $\mu^{(j)}$'s using the proposed algorithm, on one hand, we need the cutoff frequency $R = \Omega(1/\sigma\Delta_G)$; on the other hand, we need the noise level $\epsilon_z = o(1)$. It suffices to require $\sigma^2 d R^2 = o(1)$, namely having large enough separation $\Delta_G \geq \Omega(d^{1/2})$. In summary, when the separation condition is satisfied, to achieve target accuracy in estimating the parameters, we need the noise level ϵ_z to be upper bounded by some inverse polynomial in the dimensions, and this is equivalent to requiring the number of samples from the GMM to be lower bounded by $\text{poly}(k, d)$.

Although this algorithm does not outperform the scaling result in Dasgupta[37], it still sheds light on a different approach of learning GMMs. We leave it as future work to apply super-resolution algorithms to learn more general cases of GMMs or even learning mixtures of log-concave densities.

5.3.3 Open problems

In a recent work, Chen & Chi [35] showed that via structured matrix completion, the sample complexity for stable recovery can be reduced to $O(k \log^4 d)$. However, the computation complexity is still in the order of $O(k^d)$ as the Hankel matrix is

of dimension $O(k^d)$ and a semidefinite program is used to complete the matrix. It remains an open problem to reduce the sample complexity of our algorithm from $O(k^2)$ to the information theoretical bound $O(k)$, while retaining the polynomial scaling of the computation complexity.

Recently, Schiebinger et al [104] studied the problem of learning a mixture of shifted and re-scaled point spread functions $f(s) = \sum_j w_j \varphi(s, \mu^{(j)})$. This model has the Gaussian mixture as a special case, with the point spread function being Gaussian point spread $\varphi(s, \mu^{(j)}) = e^{-(s-\mu^{(j)})^T \Sigma_j^{-1} (s-\mu^{(j)})}$. We have discussed the connection between super-resolution and learning GMM. Another interesting open problem is to generalize the proposed algorithm to learn mixture of broader classes of nonlinear functions.

5.4 Proofs for Chapter 5

Proof. (of Theorem 5.3) The algorithm is correct if the tensor decomposition in Step 2 is unique, and achieves stable recovery if the tensor decomposition is stable. By the stability Lemma of tensor decomposition (Lemma 5.2), this is guaranteed if we can bound the condition number of $V_{S'}$. It follows from Lemma 5.3 that the condition number of $V_{S'}$ is at most $\sqrt{1 + \sqrt{k}}$ times of $\text{cond}_2(V_S)$. By the main technical lemma (Lemma 5.5) we know that with the random sampling set \mathcal{S} of size m , the condition number $\text{cond}_2(V_S)$ is upper bounded by a constant. Thus we can bound the distance between $V_{S'}$ and the estimation $\widehat{V}_{S'}$ according to (5.13).

Since we adopt Jennrich's algorithm for the low rank tensor decomposition, the overall computation complexity is roughly the complexity of SVD of a matrix of size $m' \times m'$, namely in the order of $O((m')^3)$. \square

Proof. For $d > 1$ case, with straightforward union bound over the $m' = O(k^2)$ samples each of which has d coordinates, one can show that the cutoff frequency is in the order of $R\sqrt{\log(kd)}$, where R is in the order of $\frac{\sqrt{\log(k)}}{\Delta}$ as shown in Theorem 5.3.

For $d = 1$ case, we bound the cutoff frequency with slightly more careful analysis. Instead of Gaussian random samples, consider uniform samples from the interval

$[-R', R']$. We can modify the proof of Lemma 5.4 and show that if $R' \geq 1/(\Delta(1+\epsilon_x))$:

$$\begin{aligned} \sum_{j' \neq j} |Y_{j,j'}| &= \sum_{j' \neq j} \frac{1}{2R'} \int_{-R', R'} e^{i\pi(\mu^{j'} - \mu^{(j)})s} = \sum_{j' \neq j} \frac{\sin(\pi|\mu^{(j')} - \mu^{(j)}|R')}{\pi|\mu^{(j')} - \mu^{(j)}|R'} \\ &\leq \sum_{l=1}^k \frac{\sin(l\pi\Delta R')}{(l\pi\Delta R')} \leq \frac{\sin(\pi\Delta R')/(\pi\Delta R')}{1 - \sin(\pi\Delta R')/(\pi\Delta R')} \leq \epsilon_x \end{aligned}$$

where the second last inequality uses the inequality that $\frac{\sin(a+b)}{a+b} \leq \frac{\sin(a)}{a} \frac{\sin(b)}{b}$. \square

Proof. (of Lemma 5.2) The proof is mostly based on the arguments in [89, 9], we still show the clean arguments here for our case.

We first introduce some notations for the exact case. Define $D_1 = \text{diag}([V_2]_{1,:}, D_w)$ and $D_2 = \text{diag}([V_2]_{2,:}, D_w)$. Recall that the symmetric matrix $F_1 = F(I, I, e_1) = VD_1V^\top$. Consider its SVD $F_1 = P\Lambda P^\top$. Denote $U = P^\top V \in \mathbb{C}^{k \times k}$. Define the whitened rank- k tensor

$$E = F(P, P, I) = (P^\top V) \otimes (P^\top V) \otimes (V_2 D_w) = U \otimes U \otimes (V_2 D_w) \in \mathbb{C}^{k \times k \times 3}.$$

Denote the two slices of the tensor E by $E_1 = E(I, I, e_1) = UD_1U^\top$ and $E_2 = E(I, I, e_2) = UD_2U^\top$. Define $M = E_1E_2^{-1}$, and its eigen decomposition is given by $M = UDU^{-1}$, where $D = D_1D_2^{-1}$. Note that in the exact case, D is given by:

$$D = \text{diag}(e^{i\pi\langle \mu^{(j)}, v^{(1)} - v^{(2)} \rangle} : j \in [k])$$

Note that $|D_{j,j}| = 1$ for all j . Define the minimal separation of the diagonal entries in D to be:

$$\text{sep}(D) = \min\{\min_{j \neq j'} |D_{j,j} - D_{j',j'}|\},$$

1. We first apply perturbation bounds to show that the noise in \tilde{F} propagates the estimates \hat{P} and \hat{E} in a mild way when the condition number of V is bounded by a constant.

Proof. Apply Wedin's matrix perturbation bound, we have:

$$\|\widehat{P} - P\|_2 \leq \frac{\|\widetilde{F}_1 - F_1\|_2}{\sigma_{\min}(F_1)} \leq \frac{\epsilon_z \sqrt{m}}{w_{\min} \sigma_{\min}(V)^2}$$

And then for the two slices of $\widehat{E} = \widetilde{F}(\widehat{P}, \widehat{P}, I)$, namely $\widehat{E}_i = E_i + Z_i$ for $i = 1, 2$, we can bound the distance between estimates and the exact case, namely $Z_i = \widehat{P}^\top \widetilde{F}_i \widehat{P} - P^\top F_i P$, by:

$$\|Z_i\| \leq 8\|F_i\|\|P\|\|\widehat{P} - P\| + 4\|P\|^2\|\widetilde{F}_i - F_i\| \leq 16 \frac{w_{\max}}{w_{\min}} \text{cond}_2(V)^2 \epsilon_z \sqrt{m}$$

□

2. Then, recall that $M = E_1 E_2^{-1} = U D U^{-1}$. Note that

$$\widehat{M} = (E_1 + Z_1)(E_2 + Z_2)^{-1} = E_1 E_2^{-1} (I - Z_2 (I + E_2^{-1} Z_2)^{-1} E_2^{-1}) + Z_1 E_2^{-1}.$$

Let H and G denote the perturbation matrices:

$$H = -Z_2 (I + E_2^{-1} Z_2)^{-1} E_2^{-1}, \quad G = Z_1 E_2^{-1}.$$

In the following claim, we show that given $\widehat{M} = \widehat{E}_1 \widehat{E}_2^{-1} = M(I + H) + G$ for some small perturbation matrix H and G , if the perturbation $\|H\|$ and $\|G\|$ are small enough and that $\text{sep}(D)$ is large enough, the eigen decomposition $\widehat{M} = \widehat{U} \widehat{D} \widehat{U}^{-1}$ is close to that of M .

Claim 5.1. *If $\|MH + G\| \leq \frac{\text{sep}(D)}{2\sqrt{k} \text{cond}_2(U)}$, then the eigenvalues of \widehat{M} are distinct and we can bound the columns of \widehat{U} and U by:*

$$\min_{\pi} \max_j \|\widehat{U}_j - U_{\pi(j)}\|_2 \leq 3 \frac{\sigma_{\max}(H) \sigma_{\max}(D) + \sigma_{\max}(G)}{\sigma_{\min}(U) \text{sep}(D)} \|\widehat{U}_j\|_2 \|V_j\|_2.$$

Proof. Let λ_j and U_j for $j \in [k]$ denote the eigenvalue and corresponding eigenvectors

of M . If $\|MH + G\| \leq \frac{sep(D)}{2\sqrt{k}cond_2(U)}$, we can bound

$$\|\widehat{M} - M\| = \|U^{-1}(M + (MH + G))U - D\| = \|U^{-1}(MH + G)U\| \leq sep(D)/2\sqrt{k},$$

thus apply Gershgorin's disk theorem, we have $|\widehat{\lambda}_j - \lambda_j| \leq \|[U^{-1}(MH + G)U]_j\|_1 \leq \sqrt{k}\|[U^{-1}(MH + G)U]_j\|_2 \leq sep(D)/2$. Therefore, the eigenvalues are distinct and we have

$$|\widehat{\lambda}_j - \lambda_{j'}| \geq |\lambda_j - \lambda_{j'}| - |\widehat{\lambda}_j - \lambda_j| \geq \frac{1}{2}|\lambda_j - \lambda_{j'}| \geq \frac{1}{2}sep(D). \quad (5.18)$$

Note that $\{U_{j'}\}$ and $\{\widehat{U}_j\}$ define two sets of basis vectors, thus we can write $\widehat{U}_j = \sum_{j'} c_{j'} U_{j'}$ (with the correct permutation for columns of \widehat{U}_j and U_j) for some coefficients $\sum_{j'} c_{j'}^2 = 1$. Apply first order Taylor expansion of eigenvector definition we have:

$$\widehat{\lambda}_j \widehat{U}_j = \widehat{M} \widehat{U}_j = (M + (MH + G)) \sum_{j'} c_{j'} U_{j'} = \sum_{j'} \lambda_{j'} c_{j'} U_{j'} + (MH + G) \widehat{U}_j.$$

Since we also have $\widehat{\lambda}_j \widehat{U}_j = \sum_{j'} \widehat{\lambda}_j c_{j'} U_{j'}$, we can write $\sum_{j'} (\widehat{\lambda}_j - \lambda_{j'}) c_{j'} U_{j'} = (MH + G) \widehat{U}_j$, and we can solve for the coefficients $c_{j'}$'s from the linear system as $[(\widehat{\lambda}_j - \lambda_{j'}) c_{j'} : j' \in [k]] = U^{-1}(MH + G) \widehat{U}_j$. Finally plug in the inequality in (5.18) we have that for any j :

$$\begin{aligned} \|\widehat{U}_j - U_j\|_2^2 &= \sum_{j' \neq j} c_{j'}^2 \|U_{j'}\|_2^2 + (c_j - 1)^2 \|U_j\|_2^2 \\ &\leq 2 \sum_{j' \neq j} c_{j'}^2 \|V_{j'}\|_2^2 \\ &\leq 8 \frac{\|U^{-1}(MH + G) \widehat{U}_j\|_2^2}{sep(D)^2} \\ &\leq 8 \frac{(\sigma_{max}(D) \sigma_{max}(H) + \sigma_{max}(G))^2}{\sigma_{min}(U)^2 sep(D)^2} \|\widehat{U}_j\|_2^2 \|V_j\|_2^2 \end{aligned}$$

□

3. Note that in the above bound for $\|\widehat{U}_j - U_j\|$, we can bound the perturbation matrices H and G by:

$$\begin{aligned}\sigma_{\max}(H) &\leq \frac{\|Z_2\|}{(1 - \sigma_{\max}(E_2^{-1}Z_2))\sigma_{\min}(E_2)} \leq \frac{\|Z_2\|}{\sigma_{\min}(E_2) - \|Z_2\|} \leq \frac{\|Z_2\|}{\sigma_{\min}(U)^2\sigma_{\min}(D_2) - \|Z_2\|}, \\ \sigma_{\max}(G) &\leq \frac{\sigma_{\max}(Z_1)}{\sigma_{\min}(E_2)} \leq \frac{\|Z_2\|}{\sigma_{\min}(U)^2\sigma_{\min}(D_2)},\end{aligned}$$

Note that $\sigma_{\min}(D_2) \geq w_{\min}$ and $\sigma_{\max}(D) = 1$ by definition. In the following claim, we apply anti-concentration bound to show that with high probability $\text{sep}(D)$ is large.

Claim 5.2. *For any $\delta_v \in (0, 1)$, with probability at least $1 - \delta_v$, we can bound $\text{sep}(D)$ by:*

$$\text{sep}(D) \geq \frac{\Delta\delta_v}{\sqrt{dk^2}}.$$

Proof. Denote $v = v^{(1)} - v^{(2)}$, and note that $\|v\| \leq \sqrt{2}$. In the regime we concern, for any pair $j \neq j'$, we have $|e^{i\pi\langle \mu^{(j)}, v \rangle} - e^{i\pi\langle \mu^{(j')}, v \rangle}| \leq |\langle \mu^{(j)} - \mu^{(j')}, v \rangle|$. Apply Lemma 5.8, we have that for $\delta \in (0, 1)$,

$$\mathbb{P}\left(|\langle \mu^{(j)} - \mu^{(j')}, v \rangle| \leq \|\mu^{(j)} - \mu^{(j')}\| \frac{\delta}{\sqrt{d}}\right) \leq \delta.$$

Take a union bound over all pairs of $j \neq j'$, we have that

$$\mathbb{P}\left(\text{for some } j \neq j', |\langle \mu^{(j)} - \mu^{(j')}, v \rangle| \leq \|\mu^{(j)} - \mu^{(j')}\| \frac{\delta}{\sqrt{dk^2}}\right) \leq k^2 \frac{\delta}{k^2} = \delta.$$

Recall that $\Delta = \min_{j \neq j'} \|\mu^{(j)} - \mu^{(j')}\|$. □

4. Recall that $U = P^\top V$. Note that since P has orthonormal columns, we have $\sigma_{\min}(U) = \sigma_{\min}(V)$ and $\|U_i\| \leq \|V_i\| = \sqrt{m}$.

Finally we apply perturbation bound to the estimates $\widehat{V}_i = \widehat{P}\widehat{U}_i$ and conclude

with the above inequalities:

$$\begin{aligned}
\|\widehat{V}_i - V_i\| &\leq 2(\|\widehat{P} - P\| \|U_i\| + \|P\| \|\widehat{U}_i - U_i\|) \\
&\leq 2 \left(\frac{\epsilon_z \sqrt{m}}{w_{\min} \sigma_{\min}(V)^2} + 3 \frac{\sigma_{\max}(H) \sigma_{\max}(D) + \sigma_{\max}(G)}{\sigma_{\min}(U) \text{sep}(D)} \|V_i\| \right) \|V_i\| \\
&\leq 2 \left(\frac{\epsilon_z \sqrt{m}}{w_{\min} \sigma_{\min}(V)^2} + 6 \frac{\|Z_2\| \|V_i\|}{(\sigma_{\min}(V)^2 \sigma_{\min}(D_2) - \|Z_2\|) \sigma_{\min}(V) \text{sep}(D)} \right) \|V_i\| \\
&\leq C \left(\frac{\sqrt{d} k^2 m w_{\max} \text{cond}_2(V)^2}{\Delta \delta_v w_{\min}^2 \sigma_{\min}(V)^3} \right) \|V_i\| \epsilon_z,
\end{aligned}$$

for some universal constant C . Note that the last inequality used the assumption that ϵ_z is small enough. \square

Proof. (of Lemma 5.3) By definition, there exist some constants λ and λ' such that $\text{cond}_2(V_S) = \lambda'/\lambda$, and for all $w \in \mathcal{P}_{1,2}^k$, we have $\lambda \leq \|V_S w\| \leq \lambda'$. Note that each element of the factor $V_{S'}$ lies on the unit circle in the complex plane, then we have:

$$\lambda^2 \leq \|V_S w\|_2^2 \leq \|V_{S'} w\|_2^2 \leq (\lambda')^2 + \sqrt{kd}.$$

We can bound the condition number of $V_{S'}$ by:

$$\text{cond}_2(V_{S'}) \leq \sqrt{\frac{(\lambda')^2 + \sqrt{kd}}{\lambda^2}} = \sqrt{1 + \frac{\sqrt{kd}}{(\lambda')^2}} \text{cond}_2(V_S) \leq \sqrt{1 + \sqrt{k} \text{cond}_2(V_S)},$$

where the last inequality is because that $\max_w \|V_S w\|_2^2 \geq \|V_S e_1\|_2^2 = d$, we have $(\lambda')^2 \geq d$. \square

Proof. (of Lemma 5.4) Denote $Y = \mathbb{E}_s[X_s]$. Note that $Y_{j,j} = 1$ for all diagonal entries. For $d = 1$ case, the point sources all lie on the interval $[-1, 1]$, we can bound

the summation of the off diagonal entries in the matrix Y by:

$$\begin{aligned}
\sum_{j' \neq j} |Y_{j,j'}| &= \mathbb{E}_s [e^{i\pi \langle \mu^{(j')} - \mu^{(j)}, s \rangle}] \\
&= \sum_{j' \neq j} e^{-\frac{1}{2}\pi^2 \|\mu^{(j)} - \mu^{(j')}\|_2^2 R^2} \\
&\leq 2(e^{-\frac{1}{2}(\pi\Delta R)^2} + e^{-\frac{1}{2}(\pi(2\Delta)R)^2} + \dots + e^{-\frac{1}{2}(\pi(k/2)\Delta R)^2}) \\
&\leq 2e^{-\frac{1}{2}(\pi\Delta R)^2} / (1 - e^{-\frac{1}{2}(\pi\Delta R)^2}) \\
&\leq \epsilon_x.
\end{aligned}$$

For $d \geq 2$ case, we simply bound each off-diagonal entries by:

$$Y_{j,j'} = e^{-\frac{1}{2}\pi^2 \|\mu^{(j)} - \mu^{(j')}\|_2^2 R^2} \leq e^{-\frac{1}{2}\pi^2 \Delta^2 R^2} \leq \epsilon_x / k.$$

Apply Lemma 5.7 (Gershgorin's Disk Theorem) and we know that all the eigenvalues of Y are bounded by $1 \pm \epsilon_x$. \square

Proof. (of Lemma 5.5) Let $\{X^{(1)}, \dots, X^{(m)}\}$ denote the i.i.d. samples of the random matrix X_s defined in (5.15), with s evaluated at the i.i.d. random samples in \mathcal{S} . Note that we have

$$\|V_S w\|_2^2 = w^\top V_S^* V_S w = w^\top \left(\frac{1}{m} \sum_{i=1}^m X^{(i)} \right) w.$$

By definition of condition number, to show that $\text{cond}_2(V_S) \leq \sqrt{\frac{1+2\epsilon_x}{1-2\epsilon_x}}$, it suffices to show that

$$(1 - 2\epsilon_x)I_{k \times k} \preceq \left(\frac{1}{m} \sum_{i=1}^m X^{(i)} \right) \preceq (1 + 2\epsilon_x)I_{k \times k}.$$

By Lemma 5.4, the spectrum of $\mathbb{E}_s[X_s]$ lies in $(1 - \epsilon_x, 1 + \epsilon_x)$. Here we only need to show that the spectrum of the sample mean $(\frac{1}{m} \sum_{i=1}^m X^{(i)})$ is close to the spectrum of the expectation $\mathbb{E}_s[X_s]$. Since each element of the random matrix $X_s \in \mathbb{C}^{k \times k}$ lies on the unit circle in the complex plane, we have $X_s^2 \preceq k^2 I$ almost surely. Therefore we can apply Lemma 5.6 (Matrix Hoeffding) to show that for $m > \frac{k}{\epsilon_x} \sqrt{8 \log \frac{k}{\delta_s}}$, with

probability at least $1 - \delta_s$, it holds that $\left\| \frac{1}{m} \sum_{i=1}^m X^{(i)} - \mathbb{E}_s[X_s] \right\|_2 \leq \epsilon_x$. \square

Auxiliary lemmas

Lemma 5.6 (Matrix Hoeffding). *Consider a set $\{X^{(1)}, \dots, X^{(m)}\}$ of independent, random, Hermitian matrices of dimension $k \times k$, with identical distribution X . Assume that $\mathbb{E}[X]$ is finite, and $X^2 \preceq \sigma^2 I$ for some positive constant σ almost surely, then, for all $\epsilon \geq 0$,*

$$\Pr \left(\left\| \frac{1}{m} \sum_{i=1}^m X^{(i)} - \mathbb{E}[X] \right\|_2 \geq \epsilon \right) \leq k e^{-\frac{m^2 \epsilon^2}{8\sigma^2}}.$$

Lemma 5.7 (Gershgorin's Disk Theorem). *The eigenvalues of a matrix $Y \in \mathbb{C}^{k \times k}$ are all contained in the following union of disks in the complex plane: $\cup_{j=1}^k \mathcal{D}(Y_{j,j}, R_j)$, where disk $\mathcal{D}(a, b) = \{x \in \mathbb{C}^k : \|x - a\| \leq b\}$ and $R_j = \sum_{j' \neq j} |Y_{j,j'}|$.*

Lemma 5.8 (Vector Random Projection). *Let $a \in \mathbb{R}^m$ be a random vector distributed uniformly over $\mathcal{P}_{1,2}^m$, and fix a vector $v \in \mathbb{C}^m$. For $\delta \in (0, 1)$, we have:*

$$\Pr \left(|\langle a, v \rangle| \leq \frac{\|v\|_2}{\sqrt{em}} \delta \right) \leq \delta$$

Proof. This follows the argument of Lemma 2.2 from Dasgupta & Gupta [38]. Extension to complex number is straightforward as we can bound the real part and the imaginary part separately. \square

Bibliography

- [1] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014. 2.1.2
- [2] E. Abbe and C. Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015. 2.1.2
- [3] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive closeness testing. In *Conference on Learning Theory (COLT)*, 2011. 2.1.2
- [4] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive classification and closeness testing. In *Conference on Learning Theory (COLT)*, 2012. 2.1.2
- [5] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, pages 3099–3132, 2009. 4.2, 4.2.1, (1)
- [6] N. Alon and A. Naor. Approximating the cut-norm via grothendieck’s inequality. *SIAM Journal on Computing*, 35(4):787–803, 2006. 1.1
- [7] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014. 2.1.2, 2.2.2, 4, 3.2, 3.3.2, 3.4.3, 3.15, 3.4.3, 3.4.3, 4.2

- [8] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014. 5.2.1,
- [9] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. *arXiv preprint arXiv:1203.0683*, 2012. 5.4
- [10] A. Anandkumar, Y. kai Liu, D. J. Hsu, D. P. Foster, and S. M. Kakade. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 25*. 2012. 2.1.2
- [11] B. D. Anderson. The realization problem for hidden markov models. *Mathematics of Control, Signals and Systems*, 12(1):80–120, 1999. 4.2.1, 4.2.1, 4.2.2
- [12] J. Anderson, M. Belkin, N. Goyal, L. Rademacher, and J. Voss. The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures. *arXiv preprint arXiv:1311.2891*, 2013. 3.1.2
- [13] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization—provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 145–162. ACM, 2012. 2.1.2
- [14] S. Arora, R. Ge, and A. Moitra. Learning topic models—going beyond svd. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE, 2012. 2.1.2
- [15] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *arXiv preprint arXiv:1502.03520*, 2015. 2.1.2
- [16] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *CoRR*, abs/1502.03520, 2015. 2.1.2

- [17] R. Bailly. Quadratic weighted automata: Spectral algorithm and likelihood maximization. *Journal of Machine Learning Research*, 20:147–162, 2011. 4.2
- [18] B. Balle, X. Carreras, F. M. Luque, and A. Quattoni. Spectral learning of weighted automata. *Machine Learning*, pages 1–31, 2013. 4.2, 4.2.1, 4.2.1, 4.2.1
- [19] B. Barak, J. A. Kelner, and D. Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. *arXiv preprint arXiv:1407.1543*, 2014. 1.1
- [20] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *Journal of the ACM (JACM)*, 60(1), 2013. 2.1.2
- [21] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Symposium on Theory of Computing (STOC)*, pages 381–390, 2004. 2.1.2
- [22] M. Belkin and K. Sinha. Learning gaussian mixtures with arbitrary separation. *arXiv preprint arXiv:0907.1054*, 2009. 3.1.2
- [23] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010. 2.1.2, 3.1.2, 5
- [24] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 594–603. ACM, 2014. 2.1.2
- [25] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the 46th ACM symposium on Theory of computing*, 2014. 3.1.2, 3.1.2
- [26] A. Bhaskara, M. Charikar, and A. Vijayaraghavan. Uniqueness of tensor decompositions with applications to polynomial identifiability. *arXiv preprint arXiv:1304.8087*, 2013. 4.2

- [27] B. Bhattacharya and G. Valiant. Testing closeness with unequal sized samples. In *Neural Information Processing Systems (NIPS) (to appear)*, 2015. 2.1.2
- [28] L. Birge. Estimating a density under order restrictions: Nonasymptotic minimax risk. *Annals of Statistics*, 15(3):995–1012, 1987. 2.1.2
- [29] S. C. Brubaker and S. S. Vempala. Isotropic pca and affine-invariant clustering. In *Building Bridges*, pages 241–281. Springer, 2008. 3.1.2
- [30] E. J. Candès and C. Fernandez-Granda. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013. 5.1.1, 2, 5.1.1, 5.1.2, 5.1.2
- [31] E. J. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014. 5.1.1, 2, 5.1.1, 5.1.2, 5.1.2
- [32] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009. 1.1
- [33] S.-O. Chan, I. Diakonikolas, R. A. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing, STOC '14*, pages 604–613, New York, NY, USA, 2014. ACM. 1
- [34] J. T. Chang. Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Mathematical Biosciences*, 137:51–73, 1996. 2.1.2
- [35] Y. Chen and Y. Chi. Robust spectral compressed sensing via structured matrix completion. *Information Theory, IEEE Transactions on*, 60(10):6576–6601, 2014. 5.3.3
- [36] P. Chin, A. Rao, and V. Vu. Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery. *arXiv preprint arXiv:1501.05021*, 2015. 2.1.2, 2.2.1, 2.3.1

- [37] S. Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999. 2.1.2, 3.1.2, 5.3.2
- [38] S. Dasgupta and A. Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random structures and algorithms*, 22(1):60–65, 2003. 5.4
- [39] S. Dasgupta and L. J. Schulman. A two-round variant of em for gaussian mixtures. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 152–159. Morgan Kaufmann Publishers Inc., 2000. 3.1.2, 5.3.2
- [40] V. H. de la Peña and S. J. Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate u-statistics. *The Annals of Probability*, pages 806–816, 1995. 3.4.7
- [41] L. De Lathauwer. A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 28(3):642–666, 2006. 4.3.2
- [42] L. De Lathauwer, J. Castaing, and J. Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *Signal Processing, IEEE Transactions on*, 55(6):2965–2973, 2007. 1.3.2, 4.3.2
- [43] D. L. Donoho. Superresolution via sparsity constraints. *SIAM Journal on Mathematical Analysis*, 23(5):1309–1331, 1992. 5.1.1
- [44] U. Feige and E. Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005. 2.1.2, 2.3.1
- [45] J. Feldman, R. A. Servedio, and R. O’Donnell. Pac learning axis-aligned mixtures of gaussians with no separation assumption. In *Learning Theory*, pages 20–34. Springer, 2006. 1

- [46] C. Fernandez-Granda. *A Convex-programming Framework for Super-resolution*. PhD thesis, Stanford University, 2014. 5.1.2
- [47] J. Friedman, J. Kahn, and E. Szemerédi. On the second eigenvalue of random regular graphs. In *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, pages 587–598. ACM, 1989. 2.1.2, 2.3.1
- [48] R. Ge, Q. Huang, and S. M. Kakade. Learning mixtures of gaussians in high dimensions. In *Proceedings of the Symposium on Theory of Computing, STOC 2015*, 2015. 2.1.2
- [49] P. W. Glynn. Upper bounds on poisson tail probabilities. *Operations research letters*, 6(1):9–14, 1987. 2.24
- [50] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. In *Technical Report TR00-020, Electronic Colloquium on Computational Complexity*, 2000. 2.1.2, 2.2.3
- [51] P. Griffiths and J. Harris. *Principles of algebraic geometry*. John Wiley & Sons, 2014. 4.3.1
- [52] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006. 2.1.2
- [53] M. Hardt, R. Meka, P. Raghavendra, and B. Weitz. Computational limits for matrix completion. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, 2014. 3.3.2
- [54] M. Hardt, R. Meka, P. Raghavendra, and B. Weitz. Computational limits for matrix completion. In *Proceedings of The 27th Conference on Learning Theory*, pages 703–725, 2014. 3.3.2
- [55] R. A. Harshman. *Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis*. University of California at Los Angeles Los Angeles, 1970. 5.2.1, 5.2.4

- [56] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983. 2.1.2
- [57] D. Hsu and S. 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. ACM, 2013. 2.1.2, 3.1.2
- [58] D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012. 2.1.2, 4.2
- [59] Q. Huang, R. Ge, S. Kakade, and M. Dahleh. Minimal realization problems for hidden markov models. *arXiv preprint arXiv:1411.3698*, 2015. 4.2.1
- [60] H. Ito, S.-I. Amari, and K. Kobayashi. Identifiability of hidden markov information sources and their minimum degrees of freedom. *Information Theory, IEEE Transactions on*, 38(2):324–333, 1992. 4.2.1
- [61] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013. 3.3.2
- [62] P. Jain and S. Oh. Learning mixtures of discrete product distributions using spectral decompositions. In *Proceedings of The 27th Conference on Learning Theory*, pages 824–856, 2014. 6
- [63] T. Jiang and N. D. Sidiropoulos. Kruskal’s permutation lemma and the identification of candecomp/parafac and bilinear models with constant modulus constraints. *Signal Processing, IEEE Transactions on*, 52(9):2625–2636, 2004. 1.3.2, 4.3.2
- [64] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 553–562. ACM, 2010. 2.1.2, 3.1.2

- [65] A. T. Kalai, A. Samorodnitsky, and S.-H. Teng. Learning and smoothed analysis. In *Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on*, pages 395–404. IEEE, 2009. 3.1.2
- [66] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. 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. 4.1, 4.2.1
- [67] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010. 1.1
- [68] R. H. Keshavan, S. Oh, and A. Montanari. Matrix completion from a few entries. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pages 324–328. IEEE, 2009. 2.1.2
- [69] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008. 1.1
- [70] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009. 1.1, 1.3.2
- [71] V. Komornik and P. Loreti. *Fourier series in control theory*. Springer Science & Business Media, 2005. 5.1.2
- [72] J. B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977. 1.1
- [73] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013. 2.1.2, 2.2.3

- [74] R. Latała et al. Estimates of moments and tails of gaussian chaoses. *The Annals of Probability*, 34(6):2315–2331, 2006. 3.4.2, 3.35
- [75] C. M. Le, E. Levina, and R. Vershynin. Sparse random graphs: regularization and concentration of the laplacian. *arXiv preprint arXiv:1502.03049*, 2015. 2.1.2, 2.3.1, 2.7.6
- [76] C. M. Le and R. Vershynin. Concentration and regularization of random graphs. *arXiv preprint arXiv:1506.00669*, 2015. 2.3.1, 1, 2, 2.4.2, 2.7.1, 2.18, 2.7.1
- [77] S. Leurgans, R. Ross, and R. Abel. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993. 1.3.2, 1, 5.2.1, 5.2.4
- [78] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27*. 2014. 2.1.2
- [79] P. Liang. *Semi-supervised learning for natural language*. PhD thesis, Massachusetts Institute of Technology, 2005. 1.1
- [80] W. Liao and A. Fannjiang. Music for single-snapshot spectral estimation: Stability and super-resolution. *Applied and Computational Harmonic Analysis*, 2014. 5.1.2, 5.1.2, 5.2.1, 5.2.1
- [81] L. Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014. 2.1.2
- [82] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004. 3.1.2, 3.2
- [83] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2.1.2
- [84] A. Moitra. The threshold for super-resolution via extremal functions. *arXiv preprint arXiv:1408.1681*, 2014. 5.1.2, 5.1.2, 5.2.1, 5.2.1

- [85] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010. 2.1.2, 3.1.2, 3.2, 5
- [86] T. K. Moon. The expectation-maximization algorithm. *Signal processing magazine, IEEE*, 13(6):47–60, 1996. 1.1
- [87] E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012. 2.1.2
- [88] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 2014. 2.1.2, 2.2.3
- [89] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden markov models. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 366–375. ACM, 2005. 4.1, 4.2, 4.2.1, 5.4
- [90] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden Markov models. *Annals of Applied Probability*, 16(2):583–614, 2006. 2.1.2
- [91] S. Nandi, D. Kundu, and R. K. Srivastava. Noise space decomposition method for two-dimensional sinusoidal model. *Computational Statistics & Data Analysis*, 58:147–161, 2013. 3)
- [92] S. on Chan, I. Diakonikolas, G. Valiant, and P. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1193–1203, 2014. 2.1.2
- [93] L. Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, 2004. 2.1.2
- [94] K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, pages 71–110, 1894. 3.1.2

- [95] H. Permuter, J. Francos, and H. Jermyn. Gaussian mixture models of texture and colour for image database retrieval. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 3, pages III–569. IEEE, 2003. 3.1.2
- [96] D. Potts and M. Tasche. Parameter estimation for nonincreasing exponential sums by prony-like methods. *Linear Algebra and its Applications*, 439(4):1024–1039, 2013. 3), 5.2.1
- [97] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009. 2.1.2
- [98] K. Ravindran and V. Santosh. *Spectral Algorithms*. Now Publishers Inc, 2009. 1.1
- [99] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010. 1.1, 3.3.2
- [100] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, 3(1):72–83, 1995. 3.1.2
- [101] M. Rudelson and R. Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009. 3.3.2, 3.4.1, 3.4.7, 3.30
- [102] D. L. Russell. Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions. *Siam Review*, 20(4):639–739, 1978. 5.1.2
- [103] A. Sanjeev and R. Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001. 3.1.2, 5.3.2

- [104] G. Schiebinger, E. Robeva, and B. Recht. Superresolution without separation. *arXiv preprint arXiv:1506.03144*, 2015. 5.3.3
- [105] N. D. Sidiropoulos and R. Bro. On the uniqueness of multilinear decomposition of n-way arrays. *Journal of chemometrics*, 14(3):229–239, 2000. 1.1
- [106] E. D. Sontag. On some questions of rationality and decidability. *Journal of Computer and System Sciences*, 11(3):375–381, 1975. 4.2.1
- [107] D. A. Spielman and S.-H. Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004. 3.1.2
- [108] G. Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM review*, 19(4):634–662, 1977. 1.3.3, 1.5
- [109] G. W. Stewart and J.-g. Sun. *Matrix perturbation theory*. Academic press, 1990. 1.3, 1.4
- [110] K. Stratos, M. Collins, and D. Hsu. Model-based word embeddings from decompositions of count matrices. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, 2015. 2.1.2
- [111] K. Stratos, M. C. Do-Kyum Kim, and D. Hsu. A spectral algorithm for learning class-based n-gram models of natural language. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2014. 2.1.2
- [112] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013. 5.1.2, 5.1.2, 5.2.1

- [113] T. Tao and V. Vu. On random ± 1 matrices: singularity and determinant. *Random Structures & Algorithms*, 28(1):1–23, 2006. 3.4.2
- [114] S. A. Terwijn. On the learnability of hidden markov models. In *Grammatical Inference: Algorithms and Applications*, pages 261–268. Springer, 2002. 4.2.1
- [115] D. M. Titterton, A. F. Smith, U. E. Makov, et al. *Statistical analysis of finite mixture distributions*, volume 7. Wiley New York, 1985. 3.1.2
- [116] G. Valiant and P. Valiant. Estimating the unseen: an $n/\log n$ -sample estimator for entropy and support size, shown optimal via new clts. In *Symposium on Theory of Computing (STOC)*, 2011. 2.1.2
- [117] G. Valiant and P. Valiant. The power of linear estimators. In *Symposium on Foundations of Computer Science (FOCS)*, 2011. 2.1.2, 2.2.3
- [118] G. Valiant and P. Valiant. Estimating the unseen: improved estimators for entropy and other properties. In *Neural Information Processing Systems (NIPS)*, 2013. 2.1.2, 2.2.3
- [119] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 51–60, 2014. 2.1.2
- [120] G. J. Valiant. *Algorithmic approaches to statistical questions*. PhD thesis, University of California, Berkeley, 2012. 2
- [121] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004. 2.1.2, 3.1.2
- [122] S. S. Vempala and Y. F. Xiao. Max vs min: Independent component analysis with nearly linear sample complexity. *arXiv preprint arXiv:1412.2954*, 2014.
- [123] M. Vidyasagar. The complete realization problem for hidden markov models: a survey and some new results. *Mathematics of Control, Signals, and Systems*, 23(1-3):1–65, 2011. 4.1, 4.2, 4.2.1

- [124] V. Vu and K. Wang. Random weighted projections, random quadratic forms and random eigenvectors. *arXiv preprint arXiv:1306.3099*, 2013. 3.4.2, 3.4.7, 3.33
- [125] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):pp. 1–25, 1982. 1.1
- [126] F. Yang, S. Balakrishnan, and M. J. Wainwright. Statistical and computational guarantees for the baum-welch algorithm. *arXiv preprint arXiv:1512.08269*, 2015. 1.1
- [127] A. Y. Zhang and H. H. Zhou. Minimax rates of community detection in stochastic block model. 2.2.3, 2.6
- [128] D. Zoran and Y. Weiss. Natural images, gaussian mixtures and dead leaves. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1736–1744. Curran Associates, Inc., 2012. 3.1.2