Efficiently learning structured distributions from untrusted batches

Efficiently Learning Structured Distributions from Untrusted Batches

Sitan Chen*
EECS, MIT
USA
sitanc@mit.edu

Jerry Li
Microsoft Research
USA
jerrl@microsoft.com

Ankur Moitra†
Department of Mathematics, MIT
USA
moitra@mit.edu

ABSTRACT

We study the problem, introduced by Qiao and Valiant, of learning from untrusted batches. Here, we assume m users, all of whom have samples from some underlying distribution $\mathbf{p}$ over $1, \ldots, n$. Each user sends a batch of k i.i.d. samples from this distribution; however an ϵ-fraction of users are untrustworthy and can send adversarially chosen responses. The goal of the algorithm is to learn $\mathbf{p}$ in total variation distance. When $k = 1$ this is the standard robust univariate density estimation setting and it is well-understood that $\Omega(\epsilon)$ error is unavoidable. Suprisingly, Qiao and Valiant gave an estimator which improves upon this rate when k is large. Unfortunately, their algorithms run in time which is exponential in either n or k.

We first give a sequence of polynomial time algorithms whose estimation error approaches the information-theoretically optimal bound for this problem. Our approach is based on recent algorithms derived from the sum-of-squares hierarchy, in the context of high-dimensional robust estimation. We show that algorithms for learning from untrusted batches can also be cast in this framework, but by working with a more complicated set of test functions.

It turns out that this abstraction is quite powerful, and can be generalized to incorporate additional problem specific constraints. Our second and main result is to show that this technology can be leveraged to build in prior knowledge about the shape of the distribution. Crucially, this allows us to reduce the sample complexity of learning from untrusted batches to polylogarithmic in n for most natural classes of distributions, which is important in many applications. To do so, we demonstrate that these sum-of-squares algorithms for robust mean estimation can be made to handle complex combinatorial constraints (e.g. those arising from VC theory), which may be of independent technical interest.

*This work was supported in part by a Paul and Daisy Soros Fellowship, NSF CAREER Award CCF-1453261, and NSF Large CCF-1565235. This work was done in part while S.C. was an intern at Microsoft Research AI.
†This work was supported in part by a Microsoft Trustworthy AI Grant, NSF CAREER Award CCF-1453261, NSF Large CCF-1565235, a David and Lucile Packard Fellowship, an Alfred P. Sloan Fellowship and an ONR Young Investigator Award.

STOC '20, June 22–26, 2020, Chicago, IL, USA

ACM ISBN 978-1-4503-6979-4/20/06…$15.00
https://doi.org/10.1145/3357713.3384337

CCS CONCEPTS

• **Theory of computation** → **Unsupervised learning and clustering**; *Semidefinite programming*; • **Mathematics of computing** → *Probability and statistics.*

KEYWORDS

Robust statistics, sum-of-squares, VC complexity, federated learning

ACM Reference Format:
Sitan Chen, Jerry Li, and Ankur Moitra. 2020. Efficiently Learning Structured Distributions from Untrusted Batches. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC '20), June 22–26, 2020, Chicago, IL, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3357713.3384337

1 INTRODUCTION

Qiao and Valiant [66] introduced the following basic problem in robust distribution learning that they called *learning with untrusted batches*. We are given N batches, each consisting of k samples from a discrete domain of size n. Each uncorrupted batch has the property that its samples were drawn i.i.d. from some distribution p_i that is η-close in total variation distance[1] to a distribution p common to all the batches. Moreover a $1 - \epsilon$ fraction of batches are uncorrupted. The remaining ϵ fraction of the batches are arbitrarily corrupted. In fact, an adversary is allowed to choose the contents of the corrupted batches after observing all of the uncorrupted batches.

The basic question is: *How well can we estimate p in total variation distance?* The key features of this problem are designed to model some of the main challenges in federated learning. In particular, we get batches of data from different users, but no batch is large enough by itself to learn an accurate model. In fact, the batches are generated from heterogenous sources because the ideal model for one user is often different than the ideal model for another. Additionally some of the batches are arbitrarily corrupted by an adversary who wishes to game our learning algorithm. In many applications, a non-trivial fraction of the data is supplied by malicious users. The meta question is: *Can we leverage information across the batches to learn an accurate model?*

In fact, the setup of learning with untrusted batches seems to model many other scenarios of interest. Our main focus will be settings where we have some additional structure or prior knowledge about the distributions we would like to learn. For example, suppose we want to estimate the demand curve across heterogenous groups. In particular, let $q_1 < q_2 < \cdots < q_n$ be a collection of increasing prices. Then set $p_{i,j}$ to be the probability that a random individual

[1]The total variation distance between distributions p, q over a domain D is defined to be $\max_{S \subseteq D} p(S) - q(S)$

from group i would buy the product when offered a price q_j but not at the price q_{j+1}. We may not have enough data from each group to accurately estimate p_i. Nevertheless we can hope to leverage data across the groups to estimate an aggregate curve p that is a good approximation to each p_i. Interestingly, the goal of being robust to an ϵ-fraction of corruptedions now takes on a different meaning in this setting: We are asking whether we can estimate p from data collected across the various groups in such a way that no ϵ-fraction of the groups can bias our estimates too much.

Qiao and Valiant [66] showed that it is possible to estimate p within $O(\eta + \epsilon/\sqrt{k})$ in total variation distance, from untrusted batches. Moreover they showed that this is the best possible up to constant factors. The somewhat surprising aspect of their bound is that it improves with larger k. This is a consequence of the "tensorization" property of the total variation distance which roughly says that the total variation distance between two distributions grows by at least a $\Omega(\sqrt{k})$ factor when we take k repetitions.

However, Qiao and Valiant [66] were only able to give an exponential time algorithm. Their approach was to estimate p by estimating the total probability it assigns to every subset of the domain. Each of these subproblems is again a problem of learning with untrusted batches, but one on a discrete domain with just two elements. Qiao and Valiant [66] gave another algorithm, but one that requires $\eta = 0$ – i.e. each of the uncorrupted batches must be generated from the same underlying distribution. Their second algorithm was based on low-rank tensor approximation. They wrote down an order k tensor whose entries represent the probability of seeing any particular k tuple of samples as a batch, and showed that some slice of this tensor is an accurate estimate of p. This algorithm also has the drawback that in order to estimate the entries of the tensor, you need n^k samples. In most applications, it would be infeasible to have so much data that you see essentially every possible batch. Their work left open the problem of getting efficient algorithms for learning with untrusted batches.

1.1 Our Results

In this work, we use the sum-of-squares (SoS) hierarchy to design new algorithms for learning from untrusted batches. An important feature of our approach is that it is easy to incorporate additional prior information about the shape of the distribution, and get even better running time and sample complexity. But first, as a warm up, we will study the original learning with untrusted batches problem. We give a sequence of polynomial time algorithms whose estimation error approaches the information-theoretically optimal bound:

Theorem 1.1 (See Theorem 3.1 for formal statement). *For any integer $t \geq 4$, there is a polynomial time algorithm to estimate p to within $O(\eta + \epsilon^{1-1/t}/\sqrt{k/t})$ in total variation from N ϵ-corrupted batches, each of size k, where $N = poly(n)$.*

This result improves over the 2^n time algorithm of Qiao and Valiant [66]. Note that the other algorithm of Qiao and Valiant [66] runs in time n^k but only works in the special case where $\eta = 0$ – i.e. all the uncorrupted batches come from the same underlying distribution. Moreover, in the above result, if we set $t = \log 1/\epsilon$ then we get within a polylogarithmic factor of the optimal estimation error, but at the expense of running in quasipolynomial time:

Corollary 1.2. *There is an $n^{O(\log 1/\epsilon)}$-time algorithm to estimate p to within $O(\eta + \epsilon\sqrt{\log 1/\epsilon}/\sqrt{k})$ in total variation from N ϵ-corrupted batches, each of size k, where $N = n^{O(\log 1/\epsilon)}$.*

Finally, we come to what we believe to be our main contribution. In many applications, getting samples is expensive and we might only be able to afford a number of samples that is sublinear in the size of the domain. In such cases, it is important to utilize additional information such as prior knowledge about the shape of the distribution. Indeed, this is the case in the example we discussed earlier, where we often know that the distribution p satisfies the monotone hazard rate condition. It is known that such distributions can be well-approximated by piecewise polynomials [2, 15, 16].

In fact, the idea of imposing structure on the underlying distribution has a long and storied history in statistics and machine learning where it leads to better estimation rates and algorithms that use fewer samples [11, 44, 83]. We ask: *Can prior information about the shape of a distribution be leveraged to get better algorithms for learning from untrusted batches?* Our main result is:

Theorem 1.3 (See Theorem 4.1 for formal statement). *For any integer $t \geq 4$, if p is approximated by an s-part piecewise polynomial function with degree at most d, there is a polynomial time algorithm to estimate p to within $O(\eta + \epsilon^{1-1/t}/\sqrt{k/t})$ in total variation from N ϵ-corrupted batches, each of size k, where $N = poly(\log n, s, d)$.*

While learning a piecewise polynomial distribution may not seem natural in applications, previous work of [2, 15, 16] has demonstrated that this can be combined with results from approximation theory [76] to achieve strong density estimation results for a large class of distribution families like log-concave distributions, Gaussians, monotone distributions, monotone hazard rate distributions, binomials, Poisson distributions, and mixtures thereof [2].

Next, we describe our techniques at a high level. The main takeaway is that the SoS hierarchy gives a seamless way to incorporate prior structural information into the estimation problem, leading to much better sample complexity guarantees.

1.2 Our Techniques

Recently, there has been a flurry of progress in high-dimensional robust estimation [17, 29, 30, 58]. While the techniques seem to be quite diverse – some relying on iterative filtering algorithms to remove outliers, and others on sum-of-squares proofs of identifiability – at their heart, they are about re-weighting the empirical distribution on the observed samples in such a way that it has bounded moments along any one-dimensional projection [34, 45, 57].

Our main observation is that algorithms for learning from untrusted batches can also be derived from this framework, but by working with a different family of test functions. When we consider moments of a one-dimensional projection, we are looking at test functions that are unit vectors (or tensor powers of them) in the ℓ_2-norm. In comparison, the exponential time algorithm of Qiao and Valiant [66] tries all ways of partitioning the domain into two sets. We can equivalently think about it as choosing a test vector (or tensor power of one) that has unit ℓ_∞-norm. In this way, we study the families of distributions for which we can find a sum-of-squares certificate that they have bounded moments with respect to

unit ℓ_∞ test functions. We show that the multinomial distribution has this property, and using the proofs-to-algorithms methodology [45, 57], this gives our improved algorithm for the general problem of learning with untrusted batches.

The beauty of this common abstraction is that it flexibly lets us build in other problem-specific constraints, like shape constraints on p. Here, classical results from VC theory [27, 79] say that it suffices to learn the distribution in a weaker norm (see Definition 4.2) than total variation. From our perspective, the change is that, in this case, instead of allowing all unit ℓ_∞ test functions, we only have to consider those which come from tensor powers of a vector that has a *bounded number of sign changes*. Encoding this constraint in the sum-of-squares hierarchy is quite non-trivial, as it is not clear how to encode this combinatorial constraint within the algebraic language of the sum-of-squares proof system. To get around this, we demonstrate that we can relax this combinatorial constraint into a linear algebraic one, namely, sparsity in the *Haar wavelet basis*. We then exploit properties of the Haar wavelet basis to encode this into our relaxation. *The main open question of our work is to push this philosophy further, and explore what other sorts of provably robust algorithms can be built out of different choices of test functions.*

1.3 Related Work

The problem of learning from untrusted batches was introduced by [66] and is motivated by problems in reliable distributed learning such as *federated learning* [56, 63]. In TCS, the problem of learning from batches has been considered in a number of settings [60, 75], but these results cannot tolerate noise in the data.

More generally, univariate density estimation, and specifically, density estimation of structured distributions, has a vast literature that we cannot hope to fully survey here. See [9] for a survey of classical results in the area. Many natural structural assumptions have been considered in statistics and learning theory, such as monotonicity [10, 41, 42, 50], monotone hazard rate [15, 20, 46], unimodality [39, 69, 82], convexity and concavity [43, 55], log-concavity [6, 37, 81], k-modality [7, 13, 40], smoothness [12, 35, 36, 53], and mixtures of structured distributions [3, 5, 19, 21–26, 31–33, 38, 51, 64, 70, 80]. The reader is referred to [28, 65] for a more extensive review of this vast literature. Recently it has been demonstrated that the classical piecewise polynomial (or spline) methods, see e.g. [73, 74, 84, 85], can be adapted to obtain general estimators for almost all these problems with nearly-optimal sample complexity and runtime [1, 2, 14–16]. While these estimators are typically tolerant of worst-case noise, it is unclear how to adapt them to the batch setting, to obtain improved statistical rates.

Finally, our work is also related to a recent line of work on robust statistics [17, 29, 30, 45, 57, 58], a classical problem dating back to the 60s and 70s [4, 47, 77, 78]. See [62, 71] for a more comprehensive survey of this line of work. We remark that the majority of this work focuses on estimation in ℓ_2-norm or Frobenius norm, with two notable exceptions: [8] uses learning in a sparsity-inducing norm to improve the sample complexity for sparse mean estimation, and [72] gives an information-theoretic characterization of when mean estimation in general norms is possible, but they do not give efficient algorithms. Our techniques are most closely related to the sum-of-squares algorithms of [45, 57], and this general technique has also found application in other robust learning problems such as robust regression [54] and list-decodable regression [52, 68].

Lastly, we mention the concurrent and independent work [48] which only considered the unstructured case and obtained a polynomial time, non-SoS algorithm getting the same error guarantee as our Theorem 1.1. Since then, follow-up work [18, 49] of the authors of the present work as well as the authors of [48] has further improved the runtime in the shape-constrained case to polynomial.

1.4 Organization

In Section 2, we provide a high-level overview of our techniques and give a sum-of-squares proof that multinomial distributions have bounded moments. In Section 3 we prove Theorem 1.1. In Section 4, we prove Theorem 1.3. The technical heart of this work is Section 5, where we fill in the details on how to efficiently encode key constraints from our SoS relaxations using matrix SoS.

2 HIGH-LEVEL ARGUMENT

In this section we give an overview of how we prove Theorems 1.1 and 1.3. The ideas required for the latter are a strict subset of those for the former, so we first describe the aspects common to both proofs before elaborating in Section 2.4 and 2.5 on techniques specific to Theorem 1.3, which we view as the main contribution of this work. As these latter sections are somewhat technical, readers new to the use of sum-of-squares for robust mean estimation may feel free to skip them on first reading, as the other sections will be sufficient for understanding the proof of Theorem 1.1.

2.1 Robust Mean Estimation

We first recast learning from untrusted batches as a generalization of robustly estimating the mean of a multinomial in L_1.

To the i-th batch of k samples $Z^i = (Z^i_1, \ldots, Z^i_k)$ from $[n]$ we may associate the vector of frequencies $Y_i \in \triangle^n$ (where $\triangle^n \subset \mathbb{R}^n$ is the probability simplex) whose j-th entry is $\frac{1}{k}\sum_{\nu=1}^{k} \mathbb{1}[Z^i_\nu = j]$ for every $j \in [n]$. If $Z^1, \ldots, Z^N$ are independent batches of k i.i.d. draws from $\mathbf{p}_1, \ldots, \mathbf{p}_N$ respectively, then $Y_1, \ldots, Y_N$ are independent draws from $\mathrm{Mul}_k(\mathbf{p}_1), \ldots, \mathrm{Mul}_k(\mathbf{p}_N)$ respectively, where $\mathrm{Mul}_k(\mathbf{p}_i)$ is defined to be the normalized multinomial distribution given by k draws from $\mathbf{p}_i$. We can think of the learning algorithm as taking in vectors $X_1, \ldots, X_N \in \triangle^n$, such that a $(1-\epsilon)N$-sized subset of them, indexed by $S_g \subset [N]$, are independent draws from $\mathrm{Mul}_k(\mathbf{p}_j)$ for $j \in S_g$, and the remaining points are arbitrary vectors in $\triangle^n$. The goal of the learning algorithm is to learn $\mathbf{p}$ in L_1 distance. Note that when $\mathbf{p}_i = \mathbf{p}$ for all $i \in S_g$, this is precisely the problem of robustly estimating the mean $\mathbf{p}$ of a (normalized) multinomial distribution.

For simplicity, we will assume that $\delta = 0$ for the rest of this subsection, i.e. that $\mathbf{p}_1 = \cdots = \mathbf{p}_N$.[2]

2.2 Searching for a Moment-Bounded Subset

A recurring theme in the robust learning literature [29, 34, 45, 57, 58] is that one can detect corruptions in the data by looking for anomalies in the empirical moments. In our setting, one useful

[2] Indeed, one appealing feature of our techniques is the ease with which one can extend the techniques we describe below to handle the case of nonzero δ.

feature of multinomial distributions $\mathrm{Mul}_k(\mathbf{p})$ is that their moments up to degree k satisfy sub-Gaussian-type bounds.

Theorem 2.1 ([59]). *For a (normalized) binomial random variable $Z \sim \frac{1}{k} \cdot Bin(k,p)$, $\mathbb{E}[(Z-p)^t]^{1/t} \lesssim \sqrt{t/k}$ for any even $t \le k$.*

Multinomial distributions inherit these same properties:

Lemma 2.2. *For any discrete distribution $\mathbf{p}$ and any $v \in \{\pm 1\}^n$, if $X \sim Mul_k(\mathbf{p})$, then $\mathbb{E}[\langle X - \mathbf{p}, v\rangle^t]^{1/t} \lesssim \sqrt{t/k}$ for any even $t \le k$.*

At a high level, our algorithms will search for a $(1-\epsilon)N$-sized subset S of the samples whose empirical moments satisfy these bounds, namely

$$\frac{1}{|S|}\sum_{i\in S}\langle X_i - \hat{\mathbf{p}}, v\rangle^t \le (8t/k)^{t/2} \quad \forall v \in \{\pm 1\}^n, \tag{1}$$

where $\hat{\mathbf{p}} = \frac{1}{|S|}\sum_{i\in S} X_i$ is the empirical mean of S. This search problem can be reformulated as solving some system $\mathcal{P}$ of polynomial equalities and inequalities (see Section 3 for a formal specification). So if we could solve this system and argue that the empirical mean of *any* subset $S \subset [N]$ which satisfies the system is $O(\epsilon/\sqrt{k})$-close in L_1 to $\mathbf{p}$, then we'd be done.

There are two complications to this approach:

(A) Solving polynomial systems is NP-hard in general.
(B) (1) is a collection of exponentially many constraints.

By now it is well-understood how to circumvent issues like (A): use the sum-of-squares (SoS) hierarchy to relax the problem of searching for a single solution to $\mathcal{P}$, or even a *distribution* over solutions, to that of searching for a *pseudodistribution* over solutions. Roughly speaking, a pseudodistribution satisfying $\mathcal{P}$ is a linear functional indistinguishable from a distribution when evaluated on low-degree polynomials arising from polynomials in $\mathcal{P}$.

The key point then is that if one can write down a "simple" proof that any solution to $\mathcal{P}$ has empirical mean close to $\mathbf{p}$, i.e. a proof using only low-degree polynomials arising from the polynomials in $\mathcal{P}$,[3] then the following learning algorithm will succeed:

(1) Solve an SDP to find a pseudodistribution $\tilde{\mathbb{E}}$ satisfying $\mathcal{P}$.
(2) Extract from $\tilde{\mathbb{E}}$ an estimate for $\mathbf{p}$.[4]

We remark that this methodology of extracting SoS algorithms from simple proofs of identifiability has been used extensively in many recent works, see [67] for a comprehensive overview.

2.3 Quantifying over $\{\pm 1\}^n$ via Matrix SoS

We now show how to address issue (B) above. The key is to design a smaller system of polynomial constraints which imply each of the exponentially many constraints in (1) under the SoS proof system, that is to say, we should be able to derive all of the constraints in (1) from the constraints in the smaller system, using only "low-degree" steps like Cauchy-Schwarz and Holder's. We remark that although the trick we will describe for doing this has appeared previously in the literature under the name of "matrix SoS proofs" [45], we believe a complete but informal treatment of this technique will help the reader better appreciate the subtleties in how we extend this approach to obtain Theorem 4.1.

To describe the trick, we first abstract out the problem-specific details of the programs we consider. Say we wish to encode the following exponentially large program with a smaller one.

Program $\mathcal{Q}$. *The variables consist of $\{Z_{\alpha,\beta}\}$ for all multisets $\alpha,\beta \subseteq [n]$ of size $t/2$, as well as some other variables $x_1,\dots,x_M$. The constraints include $\{p_1(x,Z) \ge 0,\dots,p_m(x,Z) \ge 0, q_1(x,Z) = 0,\dots,q_m(x,Z) = 0\}$ as well as the constraint*

$$\langle Z, v^{\otimes t/2}(v^{\otimes t/2})\rangle \le 1 \quad \forall v \in \{\pm 1\}^n. \tag{2}$$

Suppose we know that $\mathcal{Q}$ has a satisfying assignment (Z^*, x^*) to its variables— in the systems we will actually work with, the existence of a satisfying assignment will be immediate, e.g. the set of all uncorrupted points is a satisfying assignment to the program sketched in Section 2.2.

Remark 2.3. *While the meaning of Z will be irrelevant to the proceeding discussion, the reader might find it helpful to think of Z^*, up to scaling, as the matrix $\mathbf{Z}[S_g]$ defined by:*

$$\mathbf{Z}[S] \triangleq \frac{1}{|S|}\sum_{i\in S}\left[(X_i-\mathbf{p}_i)^{\otimes t/2}\right]^\top\left[(X_i-\mathbf{p}_i)^{\otimes t/2}\right] - \frac{1}{|S|}\sum_{i\in S}\mathop{\mathbb{E}}_{X\sim\mathcal{D}_i}\left[(X-\mathbf{p}_i)^{\otimes t/2}\right]^\top\left[(X-\mathbf{p}_i)^{\otimes t/2}\right]. \tag{3}$$

The reason is that via the identity

$$\langle \mathbf{V}[S], v^{\otimes t/2}(v^{\otimes t/2})^\top = \frac{1}{|S|}\sum_{i\in S}\left(\langle X_i-\mathbf{p}_i, v\rangle^t - \mathop{\mathbb{E}}_{X\sim\mathcal{D}_i}\langle X-\mathbf{p}_i, v\rangle^t\right),$$

$\mathbf{Z}[S_g]$ gives a succinct way of describing the deviation of the empirical moments of the subset S from the true moments.

Returning to the task at hand, we would like to write down an auxiliary program $\widehat{\mathcal{Q}}$ which satisfies three criteria, namely that $\widehat{\mathcal{Q}}$

(a) has polynomially many variables and constraints
(b) implies $\mathcal{Q}$ under the SoS proof system, and
(c) is satisfiable.

In this case, we would be done: we could simply solve an SDP to find $\tilde{\mathbb{E}}$ satisfying $\widehat{\mathcal{Q}}$ and round it. Because of (*c*) we know our SDP solver will return something, because of (*a*) we know it will do so in polynomial time, and because of (*b*) $\tilde{\mathbb{E}}$ enjoys all the same properties that a pseudodistribution satisfying $\mathcal{Q}$ would.

To see how to design such an auxiliary program $\widehat{\mathcal{Q}}$, let us suppose further that the satisfying assignment (Z^*, x^*) for $\mathcal{Q}$ satisfies the property that (2) holds *as a polynomial inequality in v*. Specifically, if we had formal variables $v_1,\dots,v_n$, suppose that one knew the existence of a proof, starting with just the polynomial equations $\{v_1^2 = 1,\dots,v_n^2 = 1\}$ cutting out the Boolean hypercube, that the inequality $\langle Z^*, v^{\otimes t/2}(v^{\otimes t/2})^\top\rangle \le 1$ held, where we now view this inequality as a polynomial equation solely in the variables $v_1,\dots,v_n$, with coefficients specified by the fixed choice of Z^*.

Showing this last assumption holds in the settings we consider will be nontrivial, but assuming for now that it does, the final idea needed to write down $\widehat{\mathcal{Q}}$ is the following. Instead of searching for Z^* satisfying the exponentially large collection of constraints (2), we can search for Z^* for which the abovementioned SoS proof of

[3] Practically speaking, for a proof to be "simple" in the above sense effectively means that the steps in the proof involve nothing more than applications of Cauchy-Schwarz and Holder's inequalities and avoid use of concentration and union bounds.
[4] We are glossing over this second step, but it turns out that a naive rounding scheme suffices (see Section 3.4).

$\langle Z^*, v^{\otimes t/2}(v^{\otimes t/2})^\top\rangle \le 1$ exists. The key point is that this search problem can be encoded in a much smaller polynomial system.

In particular, as will be evident once we give formal definitions of SoS proofs, the existence of such an SoS proof is equivalent to satisfiability of some new polynomial constraints in Z^* and some auxiliary variables corresponding to the steps of the SoS proof. To form $\hat{Q}$, we will introduce these auxiliary variables and replace constraint (2) with these new polynomial constraints. The reason this general approach is called "matrix SoS" is that these new variables will be matrix-valued, and these new constraints will be inequalities between matrix-valued polynomials. The full details of this approach are provided in the full version.

2.4 VC Meets Sum-of-Squares

Next, we describe the ideas that go into proving Theorem 4.1. The first is that when $\mathbf{p}$ is (η, s)-piecewise degree-d, to learn $\mathbf{p}$ in total variation distance, it is enough to learn $\mathbf{p}$ in a much weaker norm which we will denote by $\|\cdot\|_{\mathcal{A}_K}$, where K is a parameter that depends on s and d. This insight was the workhorse behind state-of-the-art density estimation algorithms for various structured univariate distribution classes [2, 28, 61]. In our setting, the main point is that if we have an estimate $\tilde{\mathbf{p}}$ for $\mathbf{p}$ for which $\|\tilde{\mathbf{p}} - \mathbf{p}\|_{\mathcal{A}_K} \le \zeta$, then by a result of [2], we can refine $\tilde{\mathbf{p}}$ to get an estimate $\mathbf{p}^*$ for which $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{p}^*) \le O(\zeta + \eta)$ efficiently. We review the details for this in a self-contained manner in Section 4.1.

The algorithm of [2] will form an important part of the boilerplate for our learning algorithm, but the key difficulty will be to actually find $\tilde{\mathbf{p}}$ which is close to $\mathbf{p}$ in this weaker norm. We defer definitions to Section 4.1, but informally, $\|\mathbf{p} - \tilde{\mathbf{p}}\|_{\mathcal{A}_K}$ is small if and only if $\langle \mathbf{p} - \tilde{\mathbf{p}}, v\rangle$ is small for all $v \in \mathcal{V}_K^n \subset \{\pm 1\}^n$, where $\mathcal{V}_K^n$ is the set of all $v \in \{\pm 1\}^n$ with at most K sign changes when read as a vector from left to right (for example, $(1, 1, -1, -1, 1, 1, 1) \in \mathcal{V}_2^7$).

The natural approach to do this would be to search for a $(1-\epsilon)N$-sized subset S of the samples whose empirical moments satisfy (1), except with v quantified over $\mathcal{V}_K^n$ instead of $\{\pm 1\}^n$. Roughly, the sample complexity savings would then come from the fact that the empirical moments will concentrate in fewer samples because the set of directions we need to union bound over is much smaller.

If $K = O(1)$, we could simply write down all poly(n) constraints to quantify over $\mathcal{V}_k^n$. For typical applications of piecewise polynomial approximations though, K depends logarithmically on n, so our main challenge is to get runtime that does not depend exponentially on K. Again, we will use matrix SoS. Next, we discuss some subtleties that arise in doing so.

2.5 Quantifying over $\mathcal{V}_K^n$

As in Section 2.3, we will abstract out the problem-specific details and focus on finding an encoding for the following program:

Program Q'. *This program is identical to (Q), except constraint (2) is constrained over $\mathcal{V}_K^n$ instead of $\{\pm 1\}^n$.*

The primary stumbling block is that, unlike $\{\pm 1\}^n$, $\mathcal{V}_K^n$ is not cut out by a small number of polynomial relations. Indeed, conventional wisdom says that the sum-of-squares hierarchy is ill-suited to capturing combinatorial constraints like the ones defining $\mathcal{V}_K^n$.

The first observation is that there is an alternative orthonormal basis, the *Haar wavelet basis*, under which we can express any $v \in \mathcal{V}_K^n$ as a vector with a small number $s = \tilde{O}(K)$ of nonzero entries. One issue with this is that L_0 sparsity cannot be captured by a small number of polynomial constraints, but we could try relaxing this to L_1 sparsity and attempt to derive an SoS proof of the desired moment bound out of the L_1 constraint.

Specifically, one could try to argue that any pseudodistribution $\tilde{\mathbb{E}}$ over the formal variables $v_1, \ldots, v_n, \mathbf{W}_1, \ldots, \mathbf{W}_n$ satisfying

(a) $v_i^2 = 1$ for all $i \in [n]$.
(b) $-\mathbf{W}_i \le (Hv)_i \le \mathbf{W}_i$ for all $i \in [n]$.
(c) $\sum_i \mathbf{W}_i \le s$.

must satisfy

$$\left\langle Z^*, \tilde{\mathbb{E}}\left[v^{\otimes t/2}(v^{\otimes t/2})^\top\right]\right\rangle \le 1,$$

where Z^* is fixed to a satisfying assignment to Q. One can show that the set of matrices of the form $\tilde{\mathbb{E}}\left[v^{\otimes t/2}(v^{\otimes t/2})^\top\right]$ for $\tilde{\mathbb{E}}$ satisfying the inequalities above is contained in the convex set $\mathcal{K}$ of all matrices whose Haar transforms are $L_{1,1}$-norm bounded[5] by s^t and Frobenius norm bounded by $n^{t/2}$.

At this point we instantiate all of this in the setting of this paper. Thinking of Z^*, up to scaling, as $\mathbf{Z}[S_g]$ as defined in (3), we need to ensure that its inner product with any matrix from $\mathcal{K}$ is at most one. The matrix $\mathbf{Z}[S_g]$ depends on the uncorrupted samples N, so now we are merely tasked with proving some large deviation bound (here "proof" is in the literal, non-SoS sense).

We expect this to hold with high probability for N sublinear in n because the covering number of $\mathcal{K}$ should be much smaller than that of the set of all matrices with Frobenius norm bounded by n^t. As covering number bounds can be quite subtle, we opt instead for a shelling argument. Specifically, we can show that any element M with bounded $L_{1,1}$ and Frobenius norms can be written as a sum of s^t-sparse matrices whose Frobenius norms sum to at most $\|M\|_F$ (see Lemma 5.7 and its consequences in Section 5.2 and the appendix of the full version), reducing the task of building a net over $\mathcal{K}$ to building a net $\mathcal{N}$ over s^t-sparse matrices of Frobenius norm bounded by $n^{t/2}$.

The final and perhaps most important subtlety that arises is that as stated, this argument cannot achieve sublinear sample complexity because *the inverse Haar transform of an s^t-sparse matrix with Frobenius norm $n^{t/2}$ may have large max-norm*, which would preclude the sorts of univariate concentration bounds one would hope to apply on each direction in $\mathcal{N}$. More concretely, the issue is that ultimately, the net $\mathcal{N}$ over s^t-sparse matrices of bounded Frobenius norm corresponds to a net $\mathcal{N}'$ over $\mathcal{K}$ given by the inverse Haar transform of all elements of $\mathcal{N}$. And we would need to show that for any given $M \in \mathcal{N}'$, $\langle \mathbf{Z}[S_g], M\rangle$ is at most one with high probability. But if we have no control over the scaling of the max-norm of these M's, this is evidently impossible.

The workaround for this subtlety requires modifying the three inequalities used above, as well as the definition of $\mathcal{K}$, by incorporating properties of the Haar wavelet basis beyond just the fact that vectors from $\mathcal{V}_K^n$ are sparse in this basis. Roughly speaking, the key is to exploit the inherent multi-scale nature of the Haar wavelet basis.

[5] The $L_{1,1}$ norm of a matrix is defined to be the sum of the absolute values of its entries.

This is best understood with an example. Instead of matrices, we will work with vectors (the reader can think of this as the "$t = 1$" case). In the following example, we will first try to convey 1) that there exist sparse vectors with L_2 norm $\sqrt{n}$ but whose inverse Haar transforms are as large as $\sqrt{n/2}$ in L_∞ norm. To reiterate, this is an issue because any $w \in \mathbb{R}^n$ which is a Haar transform of some vector $v \in \{\pm 1\}^n$ with few sign changes is sparse and has L_2 norm $\sqrt{n}$, yet the inverse Haar transform of w, i.e. v itself, has L_∞ norm 1. In other words, simply relaxing the set of $v \in \{\pm 1\}^n$ to the set of all vectors whose Haar transforms are sparse introduces problematic new vectors with substantially different properties than the vectors v. We will then 2) sketch how we circumvent this crucial subtlety.

Example 2.4. Let $n = 2^m$. The Haar wavelet basis for $\mathbb{R}^n$ contains the vector $\psi_\ell \triangleq (1/\sqrt{2}, -1/\sqrt{2}, 0, 0, \ldots, 0)$. Say this is the ℓ-th vector in the basis. Then the vector w which has ℓ-th entry equal to $\sqrt{n}$ and all other entries 0 is clearly sparse and has L_2 norm $\sqrt{n}$. But its inverse Haar transform is $(\sqrt{n/2}, -\sqrt{n/2}, 0, 0, \ldots, 0)$, which has largest entry $\sqrt{n/2}$, whereas any $v \in \{\pm 1\}^n$ has largest entry 1.

One reason this example is not so bad is that if we express any $v \in \{\pm 1\}^n$ as a linear combination of Haar wavelets, the coefficient for the ℓ-th Haar wavelet is $\langle v, \psi_\ell \rangle \le \sqrt{2}$. That is, the Haar transform of any such v has ℓ-th entry at most $\sqrt{2}$. So if we added to the collection of constraints defining $\mathcal{K}$ this additional constraint, we would already get rid of some problematic vectors like w.

More generally, problematic vectors like w in Example 2.4 exist at every "level" of the Haar wavelet basis, and it will be necessary to handle each of these levels appropriately. We defer the details to Lemma 5.3 and its consequences in Sections 5.2 and the appendix of the full version.

2.6 Certifiably Bounded Distributions

Recall from Section 2.3 that a prerequisite for the matrix SoS approach is that the exponentially large program from Section 2.2 must have a satisfying assignment for which there exists an SoS proof of (1) using the axioms $\{v_i^2 = 1 \; \forall i \in [n]\}$. A necessary condition for this to hold is for there to be an SoS proof from these axioms that the true moments of $\mathbf{p}$ itself satisfy these same bounds. We emphasize that these bounds should be regarded as polynomial inequalities solely in the variables $v_1, \ldots, v_n$. Formally:

Definition 2.5. Distribution $\mathcal{D}$ over $\mathbb{R}^d$ with mean μ is (t, ∞)-explicitly bounded with variance proxy σ if for all even $2 \le s \le t$:

$$\{v_i^2 = 1 \; \forall i \in [n]\} \vdash_s \mathop{\mathbb{E}}_{Y \sim \mathcal{D}} [\langle Y - \mu, v \rangle^s] \le (\sigma s)^{s/2} \tag{4}$$

We remark that while a consequence of [59] is that the moments of any multinomial satisfy these bounds, their proof uses exponentials and is thus not SoS. Here we give an SoS proof, at the cost of less desirable constants. To our knowledge, this SoS proof is new.

Lemma 2.6. *Let $\mathcal{D} = Mul_k(\mathbf{p})$ for any $\mathbf{p} \in \triangle^n$. Then $\mathcal{D}$ is (k, ∞)-explicitly bounded with variance proxy $8/k$.*

Proof. It is enough to show (4) for v for which $\|v\|_\infty = 1$. By definition $\mu = \mathbb{E}_{Y \sim \mathcal{D}}[Y]$, so we may symmetrize:

$$\vdash_s \mathop{\mathbb{E}}_{Y} [\langle Y - \mu, v \rangle^s] = \mathop{\mathbb{E}}_{Y} [\langle Y - \mathop{\mathbb{E}}_{Y'} [Y'], v \rangle^s] \le \mathop{\mathbb{E}}_{Y, Y'} [\langle Y - Y', v \rangle^s],$$

where the inequality follows from SoS Cauchy-Schwarz. But note that the random variable $\langle Y, v \rangle$ is the average of k independent copies of the random variable which takes on value v_i with probability p_i for every $i \in [n]$. So define Z to be the symmetric random variable which takes on value $(v_i - v_{i'})$ with probability $p_i p_{i'}'$ for every $(i, i') \in [n] \times [n]$. Then for $Z_1, \ldots, Z_k$ independent copies of Z,

$$\langle Y - Y', v \rangle \overset{d}{=} \frac{1}{k} \sum Z_i$$

We conclude that for any $1 \le s \le k$,

$$\begin{aligned} \vdash_s k^s & \mathop{\mathbb{E}}_{Y \sim \mathcal{D}} [\langle Y - \mu, v \rangle^s] \\ & \le \mathbb{E}[(Z_1 + \cdots + Z_k)^s] = \sum \binom{s}{\beta_1, \ldots, \beta_k} \mathbb{E}[Z_\beta] \\ & = \sum_{\beta : |\beta| = s, \beta_i \text{ even } \forall i} \binom{s}{\beta_1, \ldots, \beta_k} \mathbb{E}[Z_\beta] \le (2sk)^{s/2} \cdot \max_\beta \mathbb{E}[Z_\beta]. \end{aligned}$$

where the sum in the second expression ranges over all monomials β of total degree s. The third step follows from the fact that $\mathbb{E}[Z_\beta] = \prod_{i=1}^k \mathbb{E}[Z_i^{\beta_i}]$ by independence, and $\mathbb{E}[Z_i^d] = 0$ for any odd d because Z is symmetric. For the fourth step, note that by balls-and-bins, there are $\binom{s/2+k-1}{s/2} \le \left(\frac{3ek}{s}\right)^{s/2}$ choices of β, and $\binom{s}{\beta_1, \ldots, \beta_k} \le s! \le s^{s+1/2} e^{-s+1}$, and we may crudely bound the product of these quantities as $(3ek/s)^{s/2} \cdot s^{s+1/2} e^{-s+1} \le (2sk)^{s/2}$. By independence, $\mathbb{E}[Z_\beta] = \prod_{i=1}^k \mathbb{E}[Z_i^{\beta_i}]$. Finally, note that for every even $2 \le d \le s$, there is a degree-s SoS proof that $\mathbb{E}[Z^d] \le 2^d$. This implies $\mathbb{E}[Z_\beta] \le 2^s$ and concludes the proof. □

3 EFFICIENTLY LEARNING FROM UNTRUSTED BATCHES

In this section we prove our result on the general problem of learning from untrusted batches.

Theorem 3.1. *Let $t \ge 4$ be any integer. There is an algorithm that draws an ϵ-corrupted set of N δ-diverse batches of size k from $\mathbf{p}$ for $N \ge \delta^{-2} \epsilon^{-2} n^{O(t)} \cdot k^t / t^{t-1}$, runs in time $\delta^{-2t} \epsilon^{-2t} n^{O(t^2)} \cdot k^{t^2} / t^{t(t-1)}$, and with probability $1 - 1/poly(n)$ outputs a distribution $\hat{\mathbf{p}}$ for which $d_{TV}(\mathbf{p}, \hat{\mathbf{p}}) \le O(\delta + \epsilon^{1-1/t} \sqrt{t/k})$.*

We will describe our polynomial system and algorithm, list deterministic conditions under which our algorithm will succeed, give an SoS proof of identifiability, and conclude the proof of Theorem 3.1 by analyzing the rounding step of our algorithm. We will defer technical details for how to encode some of the constraints of our polynomial system to Section 5.

3.1 An SoS Relaxation

Let t be a power of two. For $\mathbf{p} \in \triangle^n$, let $\mathcal{D} = \mathrm{Mul}_k(\mathbf{p})$. Let $Y_1, \ldots, Y_N \in \triangle^n$ be the set of i.i.d. samples from $\mathcal{D}_1, \ldots, \mathcal{D}_N$ respectively, where for each $i \in [N]$ we have $\mathcal{D}_i = \mathrm{Mul}_k(\mathbf{p}_i)$ for some $\mathbf{p}_i \in \triangle^n$ satisfying $d_{\mathrm{TV}}(\mathbf{p}_i, \mathbf{p}) \le \delta$. Let $\{X_i\}_{i \in [N]} \in \triangle^n$ be those samples after an ϵ-fraction have been corrupted. Let $S_g \subset [N]$ (resp. $S_b \subset [N]$) denote the subset of uncorrupted (resp. corrupted) points.

Program $\mathcal{P}$. *The variables are $\{w_i\}_{i \in [N]}$, $\{\hat{\mathbf{p}}_i\}_{i \in [N]}$, and $\hat{\mathbf{p}}$, and the constraints are*

(1) $w_i^2 = w_i$ *for all* $i \in [N]$.
(2) $\sum w_i = (1-\epsilon)N$.
(3) *For every* $v \in \{\pm 1\}^n$ *and every* $i \in [N]$, $\langle \hat{\mathbf{p}}_i - \hat{\mathbf{p}}, v\rangle \le 5\delta$.
(4) $\sum_{i\in[N]} w_i X_i = \hat{\mathbf{p}} \sum_{i\in[N]} w_i$.
(5) *For every* $v \in \{\pm 1\}^n$

$$\sum_{i\in[N]} w_i \langle X_i - \hat{\mathbf{p}}_i, v\rangle^t \le (8t/k)^{t/2} \cdot \sum_{i\in[N]} w_i \tag{5}$$

(6) $\hat{\mathbf{p}}_i \ge 0$ *for all* $i \in [n]$ *and* $\sum_i \hat{\mathbf{p}}_i = 1$.

Note that constraints (3) and (5) are quantified over all $v \in \{\pm 1\}^n$, so as stated, Program $\mathcal{P}$ is a system of exponentially many polynomial constraints. In Section 5, we will explain how to encode these constraints as a small system of polynomial constraints.

Lemma 3.2. *There is a system* $\hat{\mathcal{P}}$ *of degree-*$O(t)$ *polynomial equations and inequalities in the variables* $\{w_i\}$, $\{\hat{\mathbf{p}}_i\}$, $\hat{\mathbf{p}}$, *and* $n^{O(t)}$ *other variables, whose coefficients depend on* $\epsilon, t, X_1, \ldots, X_n$ *such that*

(1) *(Satisfiability) With probability at least* $1 - 1/poly(n)$, $\hat{\mathcal{P}}$ *has a solution in which* $\hat{\mathbf{p}} = \mathbf{p}$ *and for each* $i \in [N]$, $\hat{\mathbf{p}}_i = \mathbf{p}_i$ *and* w_i *is the indicator for whether* X_i *is an uncorrupted point.*
(2) *(Encodes Moment Bounds)* $\hat{\mathcal{P}} \vdash_{O(t)} \mathcal{P}$.
(3) *(Solvability) If* $\hat{\mathcal{P}}$ *is satisfied, then for every integer* $C > 0$, *there is an* $n^{O(Ct)}$*-time algorithm which outputs a degree-*Ct *pseudodistribution which satisfies* $\hat{\mathcal{P}}$ *up to additive error* 2^{-n}.

We defer the proof of this to Section 5. Lemma 3.2 suggests the following algorithm:

Algorithm 1. LearnFromUntrusted
Input: Corruption parameter ϵ, diversity parameter δ, support size n, batch size k, samples $\{X_i\}_{i\in[N]}$, degree t
Output: Estimate $\hat{\mathbf{p}}$
(1) Run SDP solver to find a pseudodistribution $\tilde{\mathbb{E}}$ of degree $O(t)$ satisfying the constraints of Program $\hat{\mathcal{P}}$.
(2) Return $\tilde{\mathbb{E}}[\hat{\mathbf{p}}]$.

3.2 Deterministic Conditions

We will condition on the following deterministic conditions:

(I) The "Satisfiability" condition of Lemma 3.2 holds.
(II) The mean of the uncorrupted points concentrates:

$$\left\| \frac{1}{N} \sum_{i\in S_g} (X_i - \mathbf{p}_i) \right\|_1 \le O(\delta + \epsilon^{1-1/t}\sqrt{t/k})$$

(III) There is a degree-t SoS proof from $\{v_i^2 = 1 \ \forall i \in [n]\}$ that the empirical t-th moments concentrate:

$$\frac{1}{N} \sum_{i\in[N]} \left(\langle Y_i - \mathbf{p}_i, v\rangle^t - \mathop{\mathbb{E}}_{Y_i\sim\mathcal{D}_i} \langle Y_i - \mathbf{p}_i, v\rangle^t \right) \le (8t/k)^{t/2}$$

Lemma 3.3. *(I), (II), (III) all hold with probability* $1 - 1/poly(n)$.

The proof of this follows straightforwardly from Hoeffding's, and we defer it to the full version.

3.3 Identifiability

The key step is to give an SoS proof of identifiability, i.e. to demonstrate in the SoS proof system that the constraints of Program $\mathcal{P}$ imply $\hat{\mathbf{p}}$ is sufficiently close to $\mathbf{p}$:

Lemma 3.4. *Suppose Conditions (I)-(III) hold. Then for any* $v \in \{\pm 1\}^n$, *we have that*

$$\mathcal{P} \vdash_{O(t)} \langle \hat{\mathbf{p}} - \mathbf{p}, v\rangle^t \le O\left(\delta^t + \epsilon^{t-1}(t/k)^{t/2}\right).$$

First note that for any $i \in [N]$,

$$\begin{aligned}\sum_{i\in[N]} w_i\langle \hat{\mathbf{p}} - \mathbf{p}, v\rangle &= \sum_{i\in[N]} w_i\langle \hat{\mathbf{p}} - \mathbf{p}_i, v\rangle + \sum_{i\in[N]} w_i\langle \mathbf{p}_i - \mathbf{p}, v\rangle \\ &\le \sum_{i\in[N]} w_i\langle \hat{\mathbf{p}} - \mathbf{p}_i, v\rangle + 2N\delta, \end{aligned} \tag{6}$$

where the inequality follows from the assumption that $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{p}_i) \le \delta$. We bound the former term in (6):

$$\begin{aligned}&\sum_{i\in[N]} w_i\langle \hat{\mathbf{p}} - \mathbf{p}_i, v\rangle = \sum_{i\in[N]} w_i\langle X_i - \mathbf{p}_i, v\rangle \\ &= \sum_{i\in S_g} \langle X_i - \mathbf{p}_i, v\rangle + \sum_{i\in S_g} (w_i - 1)\langle X_i - \mathbf{p}_i, v\rangle + \sum_{i\in S_b} w_i\langle X_i - \mathbf{p}_i, v\rangle \\ &= \sum_{i\in S_g} \langle X_i - \mathbf{p}_i, v\rangle + \sum_{i\in S_g} (w_i - 1)\langle X_i - \mathbf{p}_i, v\rangle \\ &+ \sum_{i\in S_b} w_i\langle X_i - \hat{\mathbf{p}}_i, v\rangle + \sum_{i\in S_b} w_i\langle \hat{\mathbf{p}} - \mathbf{p}_i, v\rangle + \sum_{i\in S_b} w_i\langle \hat{\mathbf{p}}_i - \hat{\mathbf{p}}, v\rangle \\ &\le \sum_{i\in S_g} \langle X_i - \mathbf{p}_i, v\rangle + \sum_{i\in S_g} (w_i - 1)\langle X_i - \mathbf{p}_i, v\rangle \\ &+ \sum_{i\in S_b} w_i\langle X_i - \hat{\mathbf{p}}_i, v\rangle + \sum_{i\in S_b} w_i\langle \hat{\mathbf{p}} - \mathbf{p}_i, v\rangle + 5N\epsilon\delta \end{aligned}$$

where the inequality is from Constraint 3 of $\mathcal{P}$. Rearranging, taking t-th powers on both sides of (6), and invoking the above with inequality $\vdash_t (a+b+c+d+e)^t \le \exp(t)(a^t + b^t + c^t + d^t + e^t)$ gives

$$\begin{aligned}\left(\sum_{i\in S_g} w_i\right)^t \langle \hat{\mathbf{p}} - \mathbf{p}, v\rangle^t \le e^{O(t)} \Bigg[&(N(2+5\epsilon)\delta)^t + \underbrace{\left(\sum_{i\in S_g} \langle X_i - \mathbf{p}_i, v\rangle\right)^t}_{\text{Lemma 3.5}} \\ &+ \underbrace{\left(\sum_{i\in S_g} (w_i - 1)\langle X_i - \mathbf{p}_i, v\rangle\right)^t}_{\text{Lemma 3.7}} + \underbrace{\left(\sum_{i\in S_b} w_i\langle X_i - \hat{\mathbf{p}}_i, v\rangle\right)^t}_{\text{Lemma 3.8}} \Bigg], \end{aligned} \tag{7}$$

which we bound using Lemmas 3.5, 3.7, and 3.8 below. Intuitively, the term for Lemma 3.5 corresponds to sampling error from uncorrupted samples from $\mathcal{D}$, the term for Lemma 3.7 corresponds to the possible failure of the subset selected by w_i to capture some small fraction of the uncorrupted samples, and the term for Lemma 3.8 corresponds to the error contributed by the corruptions.

Lemma 3.5. *Suppose Conditions (I)-(III) hold. Then for any* $v \in \{\pm 1\}^n$, $\mathcal{P} \vdash_{O(t)} \left(\sum_{i\in S_g}\langle X_i - \mathbf{p}_i, v\rangle\right)^t \le O(N)^t(\delta^t + \epsilon^{t-1}\cdot(t/k)^{t/2})$.

Proof. By SoS Holder's

$$\mathcal{P} \vdash_{O(t)} \left(\sum_{i\in S_g}\langle X_i - \mathbf{p}_i, v\rangle\right)^t \le \left\|\sum_{i\in S_g}(X_i - \mathbf{p}_i)\right\|_1^t,$$

so the claim follows by (II) and (scalar) Holder's. □

For Lemma 3.7, we will use the following helper lemma, which follows immediately from Lemma 3.3 and Lemma 2.6.

Lemma 3.6. *Suppose Condition (III) holds. Then for any* $v \in \{\pm 1\}^n$, $\mathcal{P} \vdash_{O(t)} \sum_{i\in[N]}\langle Y_i - \mathbf{p}_i, v\rangle^t \le 2N(8t/k)^{t/2}$.

Lemma 3.7. *Suppose Conditions (I)-(III) hold. Then for any* $v \in \{\pm 1\}^n$, $\mathcal{P} \vdash_{O(t)} \left(\sum_{i\in S_g}(w_i - 1)\langle X_i - \mathbf{p}_i, v\rangle\right)^t \le 2\epsilon^{t-1}\cdot N^t\cdot(8t/k)^{t/2}$.

Proof. By SoS Holder's, we have that

$$\mathcal{P} \vdash_{O(t)} \left(\sum_{i\in S_g}(w_i - 1)\langle X_i - \mathbf{p}_i, v\rangle\right)^t = \left(\sum_{i\in S_g}(1 - w_i)\langle X_i - \mathbf{p}_i, v\rangle\right)^t$$

$$\le \left(\sum_{i\in S_g}(1 - w_i)\right)^{t-1}\left(\sum_{i\in S_g}\langle X_i - \mathbf{p}_i, v\rangle^t\right) \le (\epsilon N)^{t-1}\sum_{i\in[N]}\langle Y_i - \mathbf{p}_i, v\rangle^t$$

where the third step follows from the fact that $\vdash_2 \sum_{i\in S_g}(1 - w_i) \le \sum_{i\in[N]}(1 - w_i) = \epsilon N$. The lemma follows from Lemma 3.6. □

Lemma 3.8. *Suppose Conditions (I)-(III) hold. Then for any* $v \in \{\pm 1\}^n$, $\mathcal{P} \vdash_{O(t)} \left(\sum_{i\in S_b} w_i\langle X_i - \hat{\mathbf{p}}_i, v\rangle\right)^t \le 2\epsilon^{t-1}N^t(8t/k)^{t/2}$.

Proof. We have that

$$\mathcal{P} \vdash_{O(t)} \left(\sum_{i\in S_b} w_i\langle X_i - \hat{\mathbf{p}}_i, v\rangle\right)^t = \left(\sum_{i\in S_b} w_i^2\langle X_i - \hat{\mathbf{p}}_i, v\rangle\right)^t \le$$

$$\left(\sum_{i\in S_b} w_i\right)^{t-1}\left(\sum_{i\in[N]} w_i\langle X_i - \hat{\mathbf{p}}_i, v\rangle^t\right) \le |S_b|^{t-1}\cdot 2(8t/k)^{t/2}\sum_{i\in[N]} w_i$$

where the second step follows by SoS Holder's and even-ness of t, and the third follows from (5). □

Lemma 3.4 now follows from (7) and Lemmas 3.5, 3.7, and 3.8.

3.4 Rounding

We are now ready to complete the proof of Theorem 3.1 by specifying how to round a pseudodistribution satisfying $\hat{\mathcal{P}}$.

Lemma 3.9. *Let* $\tilde{\mathbb{E}}$ *be a degree-*$O(t)$ *pseudodistribution satisfying* $\hat{\mathcal{P}}$. *Then* $\tilde{\mathbb{E}}[\hat{\mathbf{p}}] \in \triangle^n$ *and* $d_{TV}(\tilde{\mathbb{E}}[\hat{\mathbf{p}}], \mathbf{p}) \le O(\delta + \epsilon^{1-1/t}\sqrt{t/k})$.

Proof. The fact that $\tilde{\mathbb{E}}[\hat{\mathbf{p}}]$ follows from the fact that $\tilde{\mathbb{E}}$ satisfies Constraints 6 of $\hat{\mathcal{P}}$. For the second part of the lemma, note that by the dual characterization of L_1 distance, it suffices to show that for any $v \in \{\pm 1\}^n$, $\langle\tilde{\mathbb{E}}[\hat{\mathbf{p}}] - \mathbf{p}, v\rangle \le O\left(\delta + \epsilon^{1-1/t}\sqrt{t/k}\right)$. By Lemma 3.2, $\hat{\mathcal{P}} \vdash_{O(t)} \mathcal{P}$. We get that

$$\langle\tilde{\mathbb{E}}[\hat{\mathbf{p}}] - \mathbf{p}, v\rangle^t \le \tilde{\mathbb{E}}\left[\langle\hat{\mathbf{p}} - \mathbf{p}, v\rangle^t\right] \le O\left(\delta^t + \epsilon^{t-1}(t/k)^{t/2}\right),$$

where the first step follows by scalar Holder's, the second follows by Lemma 3.4. We conclude by using that $(a + b)^{1/t} \le a^{1/t} + b^{1/t}$ for positive scalars a, b. □

This and Lemma 3.2 complete the proof of Theorem 3.1.

4 IMPROVED SAMPLE COMPLEXITY UNDER SHAPE CONSTRAINTS

In this section we prove the following, which says that the algorithmic framework of the preceding sections can be leveraged to learn *shape-constrained distributions* from untrusted batches with sample complexity *sublinear* in the domain size n.

Theorem 4.1. *For any integer* $t \ge 4$ *and* $\eta > 0$, *if* $\mathbf{p}$ *is* (η, s)*-piecewise degree-*d, *there is an algorithm that draws an* ϵ*-corrupted set of* $N = \delta^{-2}\epsilon^{-2}(sd\log n)^{O(t)}k^t/t^{t-1}$ δ*-diverse batches of size* k *from* $\mathbf{p}$, *runs in time* $\delta^{-t}\epsilon^{-t}(sdn)^{O(t)}k^{t^2}/t^{t(t-1)}$, *and with probability* $1 - 1/poly(n)$ *outputs a distribution* $\hat{\mathbf{p}}$ *for which* $d_{TV}(p, \hat{\mathbf{p}}) \le O(\eta + \delta + \epsilon^{1-1/t}\sqrt{t/k})$.

Importantly, by combining this result with known approximation theoretic results, we are able to obtain sample complexities that are either independent of the domain size or depend at most polylogarithmically on it, for a large class of natrual distributions, such as monotone distributions, monotone hazard rate distributions, log-concave distributions, discrete Guassians, Poisson Binomial distributions, and mixtures thereof, see e.g. [2] for more details. After giving the basic ingredients from VC complexity for how to learn shape-constrained distributions in sublinear sample complexity in a classical sense, we describe and analyze the polynomial system Program $\mathcal{P}'$, deferring technical details for how to encode some of the constraints of this program to Section 5 and the full version.

4.1 $\mathcal{A}_K$ Norms and VC Complexity

Definition 4.2 ($\mathcal{A}_K$ *norms, see [27]).* For positive integers $K \le n$, define $\mathcal{A}_K$ to be the set of all unions of at most K disjoint intervals over $[n]$, where an interval is any subset of the form $\{a, a+1, \ldots, b-1, b\}$. The $\mathcal{A}_K$ distance between distributions p, q over $[n]$ is given by $\|p - q\|_{\mathcal{A}_K} = \max_{S\in\mathcal{A}_K}|p(S) - q(S)|$. Equivalently, say that $v \in \{\pm 1\}^n$ has $2K$ *sign changes* if there are exactly $2K$ indices $i \in [n-1]$ for which $v_{i+1} \ne v_i$. Then if $\mathcal{V}^n_{2K}$ denotes the set of all such v, we have $\|p - q\|_{\mathcal{A}_K} = \frac{1}{2}\max_{v\in\mathcal{V}^n_{2K}}\langle p - q, v\rangle$. Note that $\|\cdot\|_{\mathcal{A}_1} \le \|\cdot\|_{\mathcal{A}_2} \le \cdots \le \|\cdot\|_{\mathcal{A}_{n/2}} = \|\cdot\|_{\mathrm{TV}}$.

Definition 4.3. We say that a distribution over $[n]$ is (η, s)*-piecewise degree-*d if there is a partition of $[n]$ into t disjoint intervals $\{[a_i, b_i]\}$, together with univariate degree-d polynomials $r_1, \ldots, r_t$ and a distribution $\mathbf{q}$ on $[n]$, such that $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \le \eta$ and such that for all $i \in [t]$, $\mathbf{q}(x) = r_i(x)$ for all $x \in [n]$ in $[a_i, b_i]$.

We defer the proof of the following to the full version.

Lemma 4.4. *Let $K = s(d+1)$. If $\mathbf{p}$ is (η, s)-piecewise degree-d and $\|\mathbf{p} - \hat{\mathbf{p}}\|_{\mathcal{A}_K} \le \zeta$, then there is an algorithm which, given the vector $\hat{\mathbf{p}}$, outputs a distribution $\mathbf{p}^*$ for which $d_{TV}(\mathbf{p}, \mathbf{p}^*) \le 2\zeta + 4\eta$ in time poly$(s, d, 1/\epsilon)$.*

4.2 Another SoS Relaxation

Henceforth, let $K = s(d+1)$ and $\ell = 2s(d+1)$. To prove Theorem 4.1, by Lemma 4.4 it suffices to learn $\mathbf{p}$ in $\mathcal{A}$ distance, that is, we wish to produce a hypothesis $\hat{\mathbf{p}}$ for which $\frac{1}{2}\max_{v \in V_\ell^n} \langle \mathbf{p} - \hat{\mathbf{p}}, v\rangle$ is small.

Program $\mathcal{P}'$. *The variables are $\{w_i\}_{i\in[N]}$, $\{\hat{\mathbf{p}}_i\}_{i\in[N]}$, and $\hat{\mathbf{p}}$, and the constraints are*

1. *$w_i^2 = w_i$ for all $i \in [N]$.*
2. *$\sum w_i = (1-\epsilon)N$.*
3. *For every $v \in \{\pm 1\}^n$ with at most ℓ sign changes and every $i \in [N]$, $\langle \hat{\mathbf{p}}_i - \hat{\mathbf{p}}, v\rangle \le 5\delta$.*
4. *$\sum_{i\in[N]} w_i X_i = \hat{\mathbf{p}} \sum_{i\in[N]} w_i$.*
5. *For every $v \in \{\pm 1\}^n$ with at most ℓ sign changes,*
$$\sum_{i\in[N]} w_i \langle X_i - \hat{\mathbf{p}}_i, v\rangle^t \le (8t/k)^{t/2} \cdot \sum_{i\in[N]} w_i.$$
6. *$\hat{\mathbf{p}}_i \ge 0$ for all $i \in [n]$ and $\sum_i \hat{\mathbf{p}}_i = 1$.*

Lemma 4.5. *There is a system $\hat{\mathcal{P}}'$ of degree-$O(t)$ polynomial equations and inequalities in the variables $\{w_i\}$, $\{\hat{\mathbf{p}}_i\}$, $\hat{\mathbf{p}}$, and $n^{O(t)}$ other variables, whose coefficients depend on $\epsilon, t, X_1, \ldots, X_n$ such that*

1. *(Satisfiability) With probability at least $1 - 1/$poly(n), $\hat{\mathcal{P}}'$ has a solution in which $\hat{\mathbf{p}} = \mathbf{p}$ and for each $i \in [N]$, $\hat{\mathbf{p}}_i = \mathbf{p}_i$ and w_i is the indicator for whether X_i is an uncorrupted point.*
2. *(Encodes Moment Bounds) $\hat{\mathcal{P}}' \vdash_{O(t)} \mathcal{P}'$.*
3. *(Solvability) If $\hat{\mathcal{P}}'$ is satisfied, then for every integer $C > 0$, there is an $n^{O(Ct)}$-time algorithm which outputs a degree-Ct pseudodistribution which satisfies $\hat{\mathcal{P}}$ up to additive error 2^{-n}.*

Together with Lemma 4.4, this suggests the following algorithm:

Algorithm 2. PiecewiseLearn
Input: Corruption parameter ϵ, diversity parameter δ, support size n, batch size k, samples $\{X_i\}_{i\in[N]}$, degree t, (η, s, d) for which $\mathbf{p}$ is (η, s)-piecewise degree-d
Output: Estimate $\mathbf{p}^*$

(1) Run SDP solver to find a pseudodistribution $\tilde{\mathbb{E}}$ of degree $O(t)$ satisfying the constraints of Program $\hat{\mathcal{P}}$.
(2) Set $\tilde{\mathbf{p}} \triangleq \tilde{\mathbb{E}}[\hat{\mathbf{p}}]$.
(3) Let $K = s(d+1)$. Using the algorithm of [2], output the s-piecewise degree-d distribution $\mathbf{p}^*$ that minimizes $\|\tilde{\mathbf{p}} - \mathbf{p}^*\|_{\mathcal{A}_K}$ (up to additive error η)

4.3 Deterministic Conditions and Identifiability

We will condition on the following deterministic conditions holding simultaneously:

(I) The "Satisfiability" condition of Lemma 4.5 holds.
(II) The mean of the uncorrupted points concentrates in $\mathcal{A}_\ell$ norm, i.e. $\left\|\frac{1}{N}\sum_{i\in S_g}(X_i - \mathbf{p}_i)\right\|_{\mathcal{A}_\ell} \le O(\epsilon^{1-1/t}\sqrt{t/k})$.
(III) For every $v \in \{\pm 1\}^n$ with at most ℓ sign changes,
$$\left|\frac{1}{N}\sum_{i\in[N]}\left(\langle Y_i - \mathbf{p}_i, v\rangle^t - \mathop{\mathbb{E}}_{Y_i\sim\mathcal{D}_i}\langle Y_i - \mathbf{p}_i, v\rangle^t\right)\right| \le (8t/k)^{t/2}$$

Lemma 4.6. *(I), (II), (III) all hold with probability $1 - 1/$poly(n).*

The SoS proof of identifiability given Program $\mathcal{P}'$ is identical to the proof of identifiability given Program $\mathcal{P}$ in Section 3.3, the only difference being that all intermediate steps in the proof are quantified over $v \in \{\pm 1\}^n$ with at most ℓ sign changes, rather than over all $v \in \{\pm 1\}^n$. This yields the following:

Lemma 4.7. *Suppose Conditions (I)-(III) hold. Then for any $v \in \{\pm 1\}^n$ with at most ℓ sign changes, we have that*
$$\mathcal{P}' \vdash_{O(t)} \langle \hat{\mathbf{p}} - \mathbf{p}, v\rangle^t \le O\left(\delta^t + \epsilon^{t-1}(t/k)^{t/2}\right).$$

Once we have Lemma 4.7, the rounding step can be analyzed in essentially the same way as Lemma 3.9, and this together with Lemma 4.5 completes the proof of Theorem 4.1.

5 ENCODING MOMENT CONSTRAINTS

In this section we will prove Lemmas 3.2 and 4.5. The programs $\hat{\mathcal{P}}$ and $\hat{\mathcal{P}}'$ referenced in those Lemmas will involve systems of inequalities among matrix-valued polynomials. We defer a formal discussion of matrix SoS proofs to the full version.

5.1 Moment Constraints for Program $\mathcal{P}$

We first show how to encode Constraint 3 of Program $\mathcal{P}$, namely that for each $i \in [N]$
$$\{v_i^2 = 1 \; \forall \; 1 \le i \le n\} \vdash_2 \langle \hat{\mathbf{p}}_i - \hat{\mathbf{p}}, v\rangle \le 5\delta. \tag{8}$$
This would hold if there existed SoS polynomials $q_S(v, \hat{p}_i, \hat{p})$ for which $5\delta - \langle \hat{\mathbf{p}}_i - \hat{\mathbf{p}}, v\rangle = \sum_S \prod_{i\in S}(1 - v_i^2) \cdot q_S(v, \hat{p}_i, \hat{p})$ such that each summand on the right-hand side is of degree at most 2. So let Q^S be an $n \times n$ matrix of indeterminates, with entries indexed by $i, j \in [n]$, which will correspond to the matrix of coefficients of $q(v, \hat{p}_i, \hat{p})$ as a quadratic polynomial in v.

Next we show how to encode Constraint 5 of Program $\mathcal{P}$. For every $S \subset [n]$ of size at most $O(t)$, let M^S be an $n^{t/2} \times n^{t/2}$ matrix of indeterminates, one for each pair of multi-indices γ, ρ over $[n]$ both of degree at most $t/2$. We would like to impose constraints on the entries $M^S_{\gamma,\rho}$ so that psd-ness of the matrices in $\{M^S : S \subseteq [n]\}$ encodes the fact that the Booleanity axioms degree-$O(t)$ imply
$$\sum_{i\in[N]} w_i \langle X_i - \hat{\mathbf{p}}_i, v\rangle^t \le 2 \cdot (8t/k)^{t/2} \sum_{i\in[N]} w_i \tag{9}$$
Condition (9) means that there exist polynomials p_S for which
$$2 \cdot (8t/k)^{t/2} \sum_{i\in[N]} w_i - \sum_{i\in[N]} w_i \langle X_i - \hat{\mathbf{p}}_i, v\rangle^t = \sum_{S:|S|\le O(t)} p_S(v, \{w_i\}, \{\hat{\mathbf{p}}_i\}, \hat{\mathbf{p}}) \cdot \prod_{i\in S}(1 - v_i^2),$$
where each p_S is a SoS polynomial such that $p_S(v, \{w_i\}, \{\hat{\mathbf{p}}_i\}, \hat{\mathbf{p}}) \cdot \prod_{i\in S}(1 - v_i^2)$ is degree $O(t)$. M^S will correspond to the matrix of coefficients of $p_S(v, \{w_i\}, \{\hat{\mathbf{p}}_i\}, \hat{\mathbf{p}})$ as a degree-t polynomial in v. Specifically, we will consider the following program.

Program $\hat{\mathcal{P}}$. *The variables are* $\{w_i\}_{i\in[N]}$, $\hat{\mathbf{p}}$, $\{\hat{\mathbf{p}}_i\}_{i\in[N]}$, $\{Q^S_{i,j}\}$, *and* $\{M^S_{\gamma,\rho}\}$ *and the constraints are*

(1) $w_i^2 = w_i$ *for all* $i \in [N]$.
(2) $\sum w_i = (1-\epsilon)N$.
(3) $5\delta - \langle \hat{\mathbf{p}}_i - \hat{\mathbf{p}}, v\rangle = \sum_S \prod_{i\in S}(1-v_i^2)\cdot\langle v, Q^S v\rangle$
(4) $\sum_{i\in[N]} w_i X_i = \hat{\mathbf{p}}\cdot\sum_{i\in[N]} w_i$
(5)

$$2\cdot(8t/k)^{t/2} - \frac{1}{(1-\epsilon)N}\sum_{i\in[N]} w_i\langle X_i - \hat{\mathbf{p}}, v\rangle^t = \sum_{S:|S|\le O(t)}\prod_{i\in S}(1-v_i^2)\cdot\langle v^{\otimes t/2}, M^S v^{\otimes t/2}\rangle$$

(6) $Q^S \succeq 0$ *for all* $S \subset [n]$ *for which* $|S| \le 2$
(7) $M^S \succeq 0$ *for all* $S \subset [n]$ *for which* $|S| \le O(t)$.
(8) $\hat{\mathbf{p}}_i \ge 0$ *for all* $i \in [n]$ *and* $\sum_i \hat{\mathbf{p}}_i = 1$.

Definition 5.1. Define the *canonical assignment* to the variables $\{w_i\}_{i\in[N]}$, $\hat{\mathbf{p}}$, and $\{\hat{\mathbf{p}}_i\}_{i\in[N]}$ to be as follows: for each $i \in [N]$, $w_i = \mathbb{1}[X_i \text{ is uncorrupted}]$, $\hat{\mathbf{p}}_i = \mathbf{p}_i$, and $\hat{\mathbf{p}} = \frac{1}{(1-\epsilon)N}\sum_i w_i X_i$.

Proof of Lemma 3.2. The fact that $\hat{\mathcal{P}} \vdash_{O(t)} \mathcal{P}$ follows by standard facts about matrix SoS, and solvability follows from the fact that the problem of outputting a degree-$O(t)$ pseudodistribution satisfying a system of degree-$O(t)$ polyomial constraints can be encoded as a semidefinite program of size $n^{O(t)}$.

It remains to show satisfiability of Program $\hat{\mathcal{P}}$. Constraints 1, 2, and 4 are clearly satisfied by the canonical assignment.

For Constraints 3 and 6, we want to show that for each $i \in [N]$, the SoS proof (8) exists as a polynomial inequality only in the variable v, with $\{\hat{\mathbf{p}}_i\}$ and $\hat{\mathbf{p}}$ now fixed. Fix any $i \in [N]$ and for convenience define $\alpha_j = (\hat{\mathbf{p}}_i - \hat{\mathbf{p}})_j$. We have that

$$\{v_i^2 = 1\ \forall\ 1 \le i \le n\} \vdash_2 \langle \hat{\mathbf{p}}_i - \hat{\mathbf{p}}, v\rangle = \sum_{j=1}^n \alpha_j v_j \le \sum_{j=1}^n |\alpha_j| = \|\hat{\mathbf{p}}_i - \hat{\mathbf{p}}\|_1$$

Because $d_{\mathrm{TV}}(\mathbf{p}_i, \mathbf{p}_j) \le 2\delta$ for all $j \in [N]$, $\|\hat{\mathbf{p}}_i - \hat{\mathbf{p}}\|_1$ is at most

$$\left\|\frac{1}{(1-\epsilon)N}\sum_{j\in S_g: j\ne i}(\mathbf{p}_i - \mathbf{p}_j)\right\|_1 + \left\|\frac{1}{(1-\epsilon)N}\sum_{j\in S_g}(X_j - \mathbf{p}_j)\right\|_1 \le 4\delta + \left\|\frac{1}{(1-\epsilon)N}\sum_{j\in S_g}(X_j - \mathbf{p}_j)\right\|_1$$

By Hoeffding's and the fact that $\{X_j\}_{j\in S_g}$ is a collection of independent draws from $\{\mathrm{Mul}_k(\mathbf{p}_j)\}_{j\in S_g}$ respectively, we know that $\left\|\frac{1}{(1-\epsilon)N}\sum_{j\in S_g}(X_j - \mathbf{p}_j)\right\|_1 \le \delta$ with probability at least $1 - n\cdot e^{-2\delta^2 N/n^2}$, from which (8) follows.

For Constraints 5 and 7, suppose a degree-$O(t)$ SoS proof of the following exists using Booleanity:

$$\frac{1}{(1-\epsilon)N}\sum_{i\in S_g}\langle X_i - \hat{\mathbf{p}}_i, v\rangle^t \le 2\cdot(8t/k)^{t/2}, \tag{10}$$

where v is the only variable and $\{w_i\}$, $\{\hat{\mathbf{p}}_i\}$, and $\mathbf{p}$ have all been fixed. By definition, this means that there exist sum-of-squares polynomials $p_S(v)$ for every $S \subset [n]$ of size at most $O(t)$ such that $p_S(v)\cdot\prod_{i\in S}(1-v_i^2)$ is degree $O(t)$ and

$$2\cdot(8t/k)^{t/2} - \frac{1}{(1-\epsilon)N}\sum_{i\in S_g} w_i\langle X_i - \hat{\mathbf{p}}_i, v\rangle^t = \sum_{|S|\le O(t)} p_S(v)\prod_{i\in S}(1-v_i^2).$$

By taking M^S to be the matrix of coefficients for which we have $\langle v^{\otimes t/2}, M^S v^{\otimes t/2}\rangle = p_S(v)$ and noting that $M^S \succeq 0$ because p_S is SoS, we satisfy the remaining Constraints 5 and 7 of $\hat{\mathcal{P}}$.

It remains to verify that the SoS proof (10) exists with high probability. Because $\hat{\mathbf{p}}_i = \mathbf{p}_i$, it is enough to show that the SoS proof

$$\{v_i^2 = 1\ \forall 1 \le i \le n\} \vdash_{O(t)} \frac{1}{(1-\epsilon)N}\sum_{i\in S_g}\langle X_i - \mathbf{p}_i, v\rangle^t \le 2\cdot(8t/k)^{t/2},$$

exists. It is enough to bound the quantity

$$b(v) \triangleq \frac{1}{(1-\epsilon)N}\sum_{i\in S_g}\langle X_i - \mathbf{p}_i, v\rangle^t - \frac{1}{(1-\epsilon)N}\sum_{i\in S_g}\mathop{\mathbb{E}}_{X\sim\mathcal{D}_i}\langle X - \mathbf{p}_i, v\rangle^t$$

by $b(v) \le (8t/k)^{t/2}$. Together with Lemma 2.6, this will conclude the proof. But the desired bound on $b(v)$ follows by condition (III) in Lemma 3.3, with probability $1 - 1/\mathrm{poly}(n)$. □

5.2 Moment Constraints for Program $\mathcal{P}'$

The only changes in going from Program $\mathcal{P}$ to Program $\mathcal{P}'$ are Constraints 3 and 5. In this section, we explain how to succinctly quantify over all $v \in \{\pm1\}^n$ with at most ℓ sign changes. To describe this encoding, we first recall some basic facts about the (discretized) Haar wavelet basis.

Haar Wavelets.

Definition 5.2. Let $m \in \mathbb{N}$ and let $n = 2^m$. The *Haar wavelet basis* is an orthonormal basis over $\mathbb{R}^n$ consisting of the *father wavelet* $\psi_{0_{\text{father}},0} = n^{-1/2}\cdot\mathbf{1}$, the *mother wavelet* $\psi_{0_{\text{mother}},0} = n^{-1/2}\cdot(1,\dots,1,-1,\dots,-1)$, and for every $1 \le i < m$ and $0 \le j < 2^i$, the wavelet $\psi_{i,j}$ whose $2^{m-i}\cdot j + 1,\dots,2^{m-i}\cdot j + 2^{m-i-1}$-th coordinates are $2^{-(m-i)/2}$, whose $2^{m-i}\cdot j + (2^{m-i-1}+1),\dots,2^{m-i}\cdot j + 2^{m-i}$-th coordinates are $-2^{-(m-i)/2}$, and whose other coordinates are 0.

Let H_m denote the $n\times n$ matrix whose rows consist of the vectors of the Haar wavelet basis for $\mathbb{R}^n$. When the context is clear, we will omit the subscript and refer to this matrix as H.

The key observation is that there is an orthonormal basis under which any $v \in \{\pm1\}^n$ with at most ℓ sign changes has an $(\ell\log n + 1)$-sparse representation.

Define $\mathcal{T} \triangleq \{0_{\text{father}}, 0_{\text{mother}}, 1, \dots, m-1\}$. By abuse of notation, we will sometimes identify the indices 0_{father} and 0_{mother} with their numerical value of 0.

Lemma 5.3. *Let* $v \in \{\pm1\}^n$ *have at most* ℓ *sign changes, and let* $\ell' \triangleq \ell\log n + 1$. *Then*

$$\sum_{i\in\mathcal{T}}\sum_{j=0}^{2^i-1} 2^{-(m-i)/2}|\langle\psi_{i,j}, v\rangle| \le \ell'. \tag{11}$$

.

Proof. We first show that Hv has at most ℓ' nonzero entries. For any $\psi_{i,j}$ with nonzero entries at indices $[a,b] \subset [n]$ and such that $i \ne 0_{\text{father}}$, if v has no sign change in the interval $[a,b]$, then

$\langle\psi_{i,j},v\rangle = 0$. For every index $\nu \in [n]$ at which v has a sign change, there are at most $m = \log n$ choices of i,j for which $\psi_{i,j}$ has a nonzero entry at index ν, from which the claim follows by a union bound over all ℓ choices of ν, together with the fact that $\langle\psi_{0_{\text{father}},0},v\rangle$ may be nonzero.

Now for each (i,j) for which $\langle\psi_{i,j},v\rangle \neq 0$, note that

$$2^{-(m-i)/2}\cdot|\langle\psi_{i,j},v\rangle| \leq 2^{-(m-i)/2}\cdot\left(2^{-(m-i)/2}\cdot 2^{m-i}\right) = 1,$$

from which (11) follows. □

For notational simplicity in the arguments below, for $\nu \in [n]$, if the ν-th element of the Haar wavelet basis for $\mathbb{R}^n$ is some $\psi_{i,j}$, then let $\mu^{(\nu)}$ denote the weight $2^{-(m-i)/2}$. Also, for any $i \in \mathcal{T}$, let $T_i \subset [n]$ denote the set of all indices ν for which the ν-th Haar wavelet is of the form $\psi_{i,j}$ for some j.

The Matrix SoS Encoding. By Lemma 5.3, instead of quantifying over all $v \in \{\pm1\}^n$ with at most ℓ sign changes in Constraints 3 and 5 of $\mathcal{P}'$, we can quantify over all $v \in \mathbb{R}^n$ with Frobenius norm at most n and for which (11) is satisfied. Specifically, we can ask for an SoS proof of

$$\langle\hat{\mathbf{p}}_i - \hat{\mathbf{p}}, v\rangle \leq 5\delta \tag{12}$$

using Axioms 5.4.

Axioms 5.4 (Axioms for Constraint 3). *Let $\mathbf{W}_1,\ldots,\mathbf{W}_n$ be auxiliary scalar variables.*

(1) $v_i^2 = 1$ *for all* $i \in [n]$
(2) $-\mathbf{W}_i \leq (Hv)_i \leq \mathbf{W}_i$ *for all* $i \in [n]$
(3) $\sum_i \mu^{(i)}\cdot\mathbf{W}_i \leq \ell'$,

Likewise, we can ask for an SoS proof of

$$\frac{1}{(1-\epsilon)N}\sum_{i\in[N]} w_i\langle X_i - \hat{\mathbf{p}}_i, v\rangle^t \leq 2\cdot(8t/k)^{t/2}, \tag{13}$$

using Axioms 5.5.

Axioms 5.5 (Axioms for Constraint 5). *Let $\{\mathbf{U}_\alpha\}$, where α ranges over all monomials in the indices $[n]$ of degree $t/2$.*

(1) $v_i^2 = 1$ *for all* $i \in [n]$
(2) $-\mathbf{U}_\alpha \leq (H^{\otimes t/2}v^{\otimes t/2})_\alpha \leq \mathbf{U}_\alpha$ *for all monomials* α *of degree* $t/2$
(3) $\sum_\alpha \mu^{(\alpha)}\mathbf{U}_\alpha \leq \ell'^{t/2}$,

where $\mu^{(\alpha)} \triangleq \prod_{i\in\alpha}\mu^{(i)}$.

As in the proof of Lemma 3.2, the values of $\{\hat{\mathbf{p}}_i\}$ and $\{w_i\}$ will be given by the canonical assignment, so the only variables in the SoS proofs of (12) and (13) will be $v_1,\ldots,v_n$ and, respectively, $\{\mathbf{W}_i\}_{i\in[n]}$ and $\{\mathbf{U}_\alpha\}_{|\alpha|\leq t/2}$.

By definition, the existence of a degree-d SoS proof for (12) using Axioms 5.4 is equivalent to the existence of polynomials $f_J^{K_1,K_2}(v,\mathbf{W},\{\hat{\mathbf{p}}_i\},\hat{\mathbf{p}})$ and $g_J^{K_1,K_2}(v,\mathbf{W},\{\hat{\mathbf{p}}_i\},\hat{\mathbf{p}})$ for $J,K_1,K_2 \subset [n]$ for which

$$5\delta - \langle\hat{\mathbf{p}}_i - \hat{\mathbf{p}}, v\rangle =$$

$$\sum_{J,K_1,K_2} f_J^{K_1,K_2}h_J^{K_1,K_2} + \left(\ell' - \sum_{i\in[n]}\mu^{(i)}\mathbf{W}_i\right)\sum_{J,K_1,K_2} g_J^{K_1,K_2}\cdot h_J^{K_1,K_2},$$

where

$$h_J^{K_1,K_2} \triangleq \prod_{i\in J}(1-v_i^2)\cdot\prod_{k_1\in K_1}(\mathbf{W}_{k_1} - (Hv)_{k_1})\cdot\prod_{k_2\in K_2}(\mathbf{W}_{k_2} + (H_v)_{k_2}),$$

and where each $f_J^{K_1,K_2}$ and $g_J^{T_1,T_2}$ is a sum-of-squares polynomial such that $f_J^{K_1,K_2}\cdot h_J^{K_1,K_2}$ and $(\mathbf{W}_j - (Hv)_j)\cdot g_J^{K_1,K_2}\cdot h_J^{K_1,K_2}$ is degree d. We will take this degree to be $d = O(1)$.

Completely analogously, the existence of a degree-d SoS proof for (13) using Axioms 5.5 is equivalent to the existence of polynomials $p_S^{T_1,T_2}(v,\mathbf{U},\{w_i\},\{\hat{\mathbf{p}}_i\},\hat{\mathbf{p}})$ and $q_S^{T_1,T_2}(v,U,\{w_i\},\{\hat{\mathbf{p}}_i\},\hat{\mathbf{p}})$ for $S \subset [n]$, $T_1,T_2 \subseteq \{\alpha : |\alpha| \leq t/2\}$ for which

$$2\cdot(8t/k)^{t/2}\sum_{i\in[N]} w_i - \sum_{i\in[N]} w_i\langle X_i - \hat{\mathbf{p}}_i, v\rangle^t = \sum_{S,T_1,T_2} p_S^{T_1,T_2}r_S^{T_1,T_2}$$
$$+\left(\ell'^{t/2} - \sum_\alpha \mu^{(\alpha)}\cdot\mathbf{U}_\alpha\right)\sum_{S,T_1,T_2}\cdot q_S^{T_1,T_2}\cdot r_S^{T_1,T_2}$$

where $r_S^{T_1,T_2}$ is defined to be $\prod_{i\in S}\left(1-v_i^2\right)$ times

$$\prod_{\alpha\in T_1}\left(\mathbf{U}_\alpha - (H^{\otimes t/2}v^{\otimes t/2})_\alpha\right)\prod_{\beta\in T_2}\left(\mathbf{U}_\beta + (H^{\otimes t/2}v^{\otimes t/2})_\beta\right)$$

and where each $p_S^{T_1,T_2}$ and $q_S^{T_1,T_2}$ is a sum-of-squares polynomial such that $p_S^{T_1,T_2}\cdot r_S^{T_1,T_2}$ and $\left(\mathbf{U}_\alpha - (H^{\otimes t/2}v^{\otimes t/2})_\alpha\right)\cdot q_S^{T_1,T_2}\cdot r_S^{T_1,T_2}$ is degree d. We will take this degree to be $d = O(t)$.

Let $F_J^{K_1,K_2}$ and $G_J^{K_1,K_2}$ respectively denote the matrices of coefficients of $f_J^{K_1,K_2}$ and $g_J^{K_1,K_2}$ as degree-$O(1)$ polynomials solely in the variables $\{v_i\}$ and $\{\mathbf{W}_i\}$, with entries denoted by $(F_J^{K_1,K_2})_{\gamma,\rho}$ and $(G_J^{K_1,K_2})_{\gamma,\rho}$. Likewise, let $P_S^{T_1,T_2}$ and $Q_S^{T_1,T_2}$ respectively denote the matrices of coefficients of $p_S^{T_1,T_2}$ and $q_S^{T_1,T_2}$ as degree-$O(t)$ polynomials solely in the variables $\{v_i\}$ and $\{\mathbf{U}_\alpha\}$, with entries denoted by $(P_S^{T_1,T_2})_{\gamma,\rho}$ and $(Q_S^{T_1,T_2})_{\gamma,\rho}$.

Remark 5.6. *As we will demonstrate in the course of our analysis, we only need consider K_1,K_2 of size at most 1, and T_1,T_2 of size at most 2, so the total number of constraints in the overall program will only be singly-exponential in t.*

We will consider the following program.

Program $\hat{\mathcal{P}}'$. *The variables are $\{w_i\}_{i\in[N]}$, $\hat{\mathbf{p}}$, $\{\hat{\mathbf{p}}_i\}_{i\in[N]}$, $\{Q_{ij}\}$, $\{(P_S^{T_1,T_2})_{\gamma,\rho}\}$, $\{(Q_S^{T_1,T_2})_{\gamma,\rho}\}$, and the constraints are*

(1) $w_i^2 = w_i$ *for all* $i \in [N]$
(2) $(1-\epsilon)N \leq \sum w_i \leq (1-\epsilon)N$
(3)

$$5\delta - \langle\hat{\mathbf{p}}_i - \hat{\mathbf{p}}, v\rangle = \sum_{J,K_1,K_2} h_J^{K_1,K_2}\cdot\langle(v,\mathbf{W})^{\otimes t/2}, F_K^{J_1,J_2}(v,\mathbf{W})^{\otimes t/2}\rangle$$
$$+\left(\ell' - \sum_i \mu^{(i)}\mathbf{W}_i\right)\sum_{J,K_1,K_2} h_J^{K_1,K_2}\cdot\langle(v,\mathbf{W})^{\otimes t/2}, G_K^{J_1,J_2}(v,\mathbf{W})^{\otimes t/2}\rangle$$

(4) $\sum_{i\in[N]} w_iX_i = \hat{\mathbf{p}}\cdot\sum_{i\in[N]} w_i$

(5)

$$2(8t/k)^{t/2}\sum_{i\in[N]} w_i - \sum_{i\in[N]} w_i\langle X_i-\hat{\mathbf{p}}_i, v\rangle^t = \sum_{S,T_1,T_2} r_S^{T_1,T_2}\cdot\langle (v,\mathbf{U})^{\otimes t/2}, P_S^{T_1,T_2}(v,\mathbf{U})^{\otimes t/2}\rangle$$

$$+\left(\ell'^{t/2} - \sum_\alpha \mu^{(\alpha)}\cdot \mathbf{U}_\alpha\right)\sum_{S,T_1,T_2} r_S^{T_1,T_2}\cdot\langle (v,\mathbf{U})^{\otimes t/2}, Q_S^{T_1,T_2}(v,\mathbf{U})^{\otimes t/2}\rangle$$

(6) $F_S^{T_1,T_2}, G_S^{T_1,T_2} \geq 0$ *for all* $T_1,T_2,S\subset[n]$ *for which* $|T_1|,|T_2|,|S|\leq O(t)$..

(7) $P_S^{T_1,T_2}, Q_S^{T_1,T_2} \geq 0$ *for all* $T_1,T_2,S\subset[n]$ *for which* $|T_1|,|T_2|,|S|\leq O(t)$.

(8) $\hat{\mathbf{p}}_i \geq 0$ *for all* $i\in[n]$ *and* $\sum_i \hat{\mathbf{p}}_i = 1$.

Proof of Lemma 4.5. As before, solvability follows from the fact that the problem of outputting a degree-$O(t)$ pseudodistribution satisfying a system of degree-$O(t)$ polynomial constraints can be encoded a a semidefinite program of size $n^{O(t)}$.

The fact that $\hat{\mathcal{P}}' \vdash_{O(t)} \mathcal{P}'$ follows by definition and by Lemma 11.

Finally, we verify that under the canonical assignment, with high probability over $X_1,\ldots,X_N$ there exists a satisfying assignment to the remaining variables of $\hat{\mathcal{P}}'$. As in the proof of Lemma 3.2, the canonical assignment clearly satisfies Constraints 1, 2, and 4.

We prove that Constraints 3 and 6 are satisfiable with high probability in Lemma 5.8, and we prove that Constraints 5 and 7 are satisfiable with high probability in Lemma 5.12. □

The following will be useful in proving Lemmas 5.8 and 5.12.

Lemma 5.7 ("Shelling trick"). *If* $v\in\mathbb{R}^m$ *satisfies* $\|v\|_2\leq C$ *and* $\|v\|_1 = C\cdot\sqrt{k}$, *then there exist* k*-sparse vectors* $v_1,\ldots,v_{m/k}$ *with disjoint supports for which* $v=\sum_{i=1}^{m/k} v_i$ *and* $\sum_{i=1}^{m/k}\|v_i\|_2\leq 2C$.

Lemma 5.8. *Under the canonical assignment, with high probability there is some choice of* $\{(F_K^{J_1,J_2})_{\gamma,\rho}\}$ *and* $\{(G_K^{J_1,J_2})_{\gamma,\rho}\}$ *for which Constraints 3 and 6 are satisfied.*

Proof. We first write

$$\langle\hat{\mathbf{p}}_i-\hat{\mathbf{p}}, v\rangle = \frac{1}{m}\sum_{j\neq i}\langle \mathbf{p}_i-\mathbf{p}_j, v\rangle + \frac{1}{m}\sum_{j\in S_g}\langle X_j-\mathbf{p}_j, v\rangle.$$

Booleanity degree-2 implies that $\langle \mathbf{p}_i-\mathbf{p}_j, v\rangle\leq 4\delta$. It remains to show that with high probability, there is a degree-$O(t)$ proof that Axioms 5.4 imply $\frac{1}{m}\sum_{j\in S_g}\langle X_j-\mathbf{p}_j, v\rangle\leq\delta$.

Equivalently, we must show that for any degree-t pseudodistribution $\tilde{\mathbb{E}}$ over the variables v and $\mathbf{U}$ which satisfies Axioms 5.4,

$$\frac{1}{m}\sum_{j\in S_g}\langle X_j-\mathbf{p}_j, \tilde{\mathbb{E}}[v]\rangle\leq\delta. \tag{14}$$

The set of vectors $\tilde{\mathbb{E}}[v]$ arising from pseudodistributions $\tilde{\mathbb{E}}$ satisfying Axioms 5.4 is some convex set $\mathcal{J}\subset\mathbb{R}^n$.

Lemma 5.9. *Let* $\mathcal{J}$ *be the convex set of all vectors* $\tilde{\mathbb{E}}[v]$ *for some degree-*t $\tilde{\mathbb{E}}$ *over the variables* $v,\mathbf{W}$ *satisfying Axioms 5.4. Additionally, let* $\mathcal{J}_1,\mathcal{J}_2\subset\mathbb{R}^n$ *consist of all vectors* u *for which* $\sum_i\mu^{(i)}|u_i|\leq\ell'$ *and for which* $\|u\|_2\leq\sqrt{n}$ *respectively. Then* $\mathcal{J}\subset H^{-1}(\mathcal{J}_1\cap\mathcal{J}_2)$.

Proof. Take any $u\in\mathcal{J}$. We first show that $u\in H^{-1}\cdot\mathcal{J}_1$. By linearity of $\tilde{\mathbb{E}}$, we may write u as $u=H^{-1}\cdot\tilde{\mathbb{E}}[Hv]$. For any $i\in[n]$, the second of Axioms 5.4 immediately implies that $-\mathbf{W}_i\leq\tilde{\mathbb{E}}[(Hv)_i]\leq\mathbf{W}_i$. This is the only place where we use the second of Axioms 5.4, and only in a linear fashion, hence Remark 5.6.

So $\sum_i\mu^{(i)}|(Hu)_i|\leq\tilde{\mathbb{E}}[\sum_i\mu^{(i)}\mathbf{W}_i]\leq\ell'$, where the last inequality follows by the third of Axioms 5.4.

Finally, to show that $u\in H^{-1}\cdot\mathcal{J}_2$, note first that by orthonormality of H, it is enough to show that $u\in\mathcal{J}_2$. But this follows immediately from the fact that $\tilde{\mathbb{E}}$ satisfies the first of Axioms 5.4, which implies that $-1\leq\tilde{\mathbb{E}}[v_i]\leq 1$ for all $i\in[n]$, from which we conclude that $\|u\|_2^2=n$ and thus $u\in\mathcal{J}_2$. □

Lemma 5.10. *For every* $\eta\leq(\ell')^{-1}$, *there exists a set* $\mathcal{N}\subset\mathbb{P}_{n-1}(\mathbb{R})$ *of size* $O(n^{3/2}/\eta)^s$ *such that for every* $u\in H^{-1}(\mathcal{J}_1\cap\mathcal{J}_2)$, *there exists some* $\tilde{u}=\sum_\nu\alpha_\nu\cdot u_\nu^*$ *for* $u_\nu^*\in\mathcal{N}$ *such that 1)* $\|u-\tilde{u}\|_2\leq\eta$, *2)* $\sum_\nu\alpha_\nu\leq 1$, *and 3)* $\|u_\nu^*\|_\infty\leq 2(\ell')$ *for all* ν.

Proof. Let $s=\ell'$, and let $m=\log n$. Let $\mathcal{N}'$ be an $\frac{\eta}{(m+1)\sqrt{n}}$-net in L_2 norm for all s^2-sparse vectors in $\S^{n-1}$. Because $\S^{s^2-1}$ has an $\frac{\eta}{(m+1)\sqrt{n}}$-net in L_2 norm of size $(3(m+1)\sqrt{n}/\eta)^{s^2}$, by a union bound we have that $|\mathcal{N}'|\leq\binom{n}{s^2}\cdot(3(m+1)\sqrt{n}/\eta)^{s^2}=O(n^{3/2}\log n/\eta)^{s^2}$.

Take any $u\in H^{-1}(\mathcal{J}_1\cap\mathcal{J}_2)$ and consider $w\triangleq Hu\in\mathcal{J}_1\cap\mathcal{J}_2$. We may write w as $\sum_{i\in\mathcal{T}}w[i]$, where $w[i]=\sum_{\nu\in T_i}w_\nu\cdot e_\nu$ for e_ν the ν-th standard basis vector in $\mathbb{R}^n$.

As the nonzero entries of $w[i]$ are just a subset of those of w, we clearly have $\|w[i]\|_2\leq\sqrt{n}$ for all $i\in\mathcal{T}$. Moreover, because $w\in\mathcal{J}_1$, we have that

$$\sum_i 2^{-(m-i)/2}|w[i]|\leq s, \tag{15}$$

so in particular $\|w[i]\|_1\leq 2^{(m-i)/2}\cdot s=2^{-i/2}\cdot s\sqrt{n}$. We can thus apply Lemma 5.7 to conclude that for each $i\in[m]$, $w[i]=\sum_j w^{i,j}$ for some vectors $\{w^{i,j}\}_j$ of sparsity at most $\lceil 2^{-i}\cdot s^2\rceil\leq s^2$ and for which $\sum_j\left\|w^{i,j}\right\|_2\leq\sqrt{n}$. For each $w^{i,j}$, there is some $(w')^{i,j}\in\mathcal{N}'$ such that if we define $\tilde{w}^{i,j}\triangleq\left\|w^{i,j}\right\|_2\cdot(w')^{i,j}$, then we have

$$\left\|w^{i,j}-\tilde{w}^{i,j}\right\|_2\leq\frac{\eta}{(m+1)\sqrt{n}}\cdot\left\|w^{i,j}\right\|_2. \tag{16}$$

Defining $\tilde{w}[i]\triangleq\sum_j\tilde{w}^{i,j}$, we get that

$$\|w[i]-\tilde{w}[i]\|_2\leq\frac{\eta}{(m+1)\sqrt{n}}\sum_j\left\|w^{i,j}\right\|_2\leq\frac{\eta}{m+1}.$$

So if we define $\tilde{w}\triangleq\sum_{i\in\mathcal{T}}\tilde{w}[i]=\sum_{i\in\mathcal{T}}\sum_j\tilde{w}^{i,j}$, we have that $\|w-\tilde{w}\|_2\leq\eta$.

Now let $\mathcal{N}\triangleq\mathbb{P}\left(H^{-1}\mathcal{N}'\right)$. As $u=H^{-1}w$ and H^{-1} is an isometry, if we define $\tilde{u}^{i,j}\triangleq H^{-1}\tilde{w}^{i,j}$ and $\tilde{u}\triangleq\sum_{i\in\mathcal{T}}\sum_j\tilde{u}^{i,j}$, then we likewise get that $\|u-\tilde{u}\|_2\leq\eta$, and clearly $\tilde{u}^{i,j}\in\mathcal{N}$, concluding the proof of part 1) of the lemma.

For each $\tilde{u}^{i,j}$, define

$$u_*^{i,j}\triangleq\tilde{u}^{i,j}/\alpha_{i,j}\quad\text{for}\quad\alpha_{i,j}\triangleq s^{-1}\cdot 2^{-(m-i)/2}\left\|w^{i,j}\right\|_\infty \tag{17}$$

so that

$$\tilde{u}=\sum_{i,j}\alpha_{i,j}u_*^{i,j}.$$

Note that $\sum_{i,j} \alpha_{i,j}$ is at most

$$\frac{1}{s}\sum_i 2^{-(m-i)/2}\sum_j \left\|w^{i,j}\right\|_\infty \le \frac{1}{s}\sum_i 2^{-(m-i)/2}\|w[i]\|_1 \le 1,$$

where the penultimate inequality follows by the fact that for fixed i, the supports of the vectors $w^{i,j}$ are disjoint for different j so that $\sum_j \left\|w^{i,j}\right\|_\infty \le \|w[i]\|_1$, and the last inequality follows from (15). This concludes the proof of part 2) of the lemma.

Finally, we need to bound $\left\|u_*^{i,j}\right\|_\infty$. Note first that for any vector z supported only on indices $\nu \in T_i$,

$$\left\|H^{-1}z\right\|_\infty \le 2^{-(m-i)/2}\cdot\|z\|_\infty \tag{18}$$

because the Haar wavelets $\{\psi_{i,j}\}_j$ have disjoint supports and L_∞ norm $2^{-(m-i)/2}$. It follows that

$$\begin{aligned}
\left\|\tilde{u}^{i,j}\right\|_\infty &\le \left\|H^{-1}w^{i,j}\right\|_\infty + \left\|H^{-1}(w^{i,j}-\tilde{w}^{i,j})\right\|_\infty \\
&\le 2^{-(m-i)/2}\cdot\left\|w^{i,j}\right\|_\infty + 2^{-(m-i)/2}\left\|w^{i,j}-\tilde{w}^{i,j}\right\|_2 \\
&\le 2^{-(m-i)/2}\cdot\left\|w^{i,j}\right\|_\infty + 2^{-(m-i)/2}\cdot\frac{\eta}{(m+1)\sqrt{n}}\left\|w^{i,j}\right\|_2 \\
&\le 2^{-(m-i)/2}\cdot\left\|w^{i,j}\right\|_\infty + 2^{-(m-i)/2}\cdot\frac{\eta}{(m+1)\sqrt{n}}\left\|w^{i,j}\right\|_\infty\cdot s \\
&= 2^{-(m-i)/2}\cdot\left\|w^{i,j}\right\|_\infty\left(1+\frac{\eta\cdot s}{(m+1)\sqrt{n}}\right) \\
&\le 2\cdot 2^{-(m-i)/2}\cdot\left\|w^{i,j}\right\|_\infty,
\end{aligned}$$

where the first inequality is triangle inequality, the second inequality follows by (18), the third inequality follows from monotonicity of L_p norms, the fourth inequality follows from (16), the fifth inequality follows from the fact that $w^{i,j}$ is s^2-sparse, and the final inequality follows from the hypothesis that $\eta \le 1/s$. Recalling (17), we conclude that $\left\|u_*^{i,j}\right\|_\infty \le 2s$ as claimed. □

By Hoeffding's, we can control $\frac{1}{m}\sum_{j\in S_g}\langle X_j - \mathbf{p}_j, u\rangle$ $\forall u \in \mathcal{N}$:

Lemma 5.11. *Let $\xi > 0$ and let $\mathcal{N} \in \mathbb{P}_{n-1}(\mathbb{R})$ be any collection of M directions. Then*

$$\Pr\left[\frac{1}{m}\sum_{j\in S_g}\langle X_j-\mathbf{p}_j,u\rangle > \xi\cdot\|u\|_\infty \ \ \forall u\in\mathcal{N}\right] < 2M\cdot e^{-2m\xi^2},$$

where the probability is over the samples X_j for $j \in S_g$.

We may now proceed with the proof of (14). For $u \in \mathcal{J}$, by Lemmas 5.9 and 5.10, there is some $\tilde{u} = \sum_\nu \alpha_\nu u_\nu^*$ such that $u_\nu^* \in \mathcal{N}$ and $\|u-\tilde{u}\|_2 \le \eta$. We may write

$$\begin{aligned}
\frac{1}{m}\sum_{j\in S_g}\langle X_j-\mathbf{p}_j,u\rangle &\le \frac{1}{m}\sum_{j\in S_g}\langle X_j-\mathbf{p}_j,\tilde{u}\rangle + \left\|\frac{1}{m}\sum_{j\in S_g}X_j\right\|_2\cdot\|u-\tilde{u}\|_2 \\
&\le \frac{1}{m}\sum_{j\in S_g}\langle X_j-\mathbf{p}_j,\tilde{u}\rangle + \eta = \sum_\nu\alpha_\nu\left(\frac{1}{m}\sum_{j\in S_g}\langle X_j-\mathbf{p}_j,u_\nu^*\rangle\right)+\eta \\
&\le \sum_\nu \alpha_\nu\cdot\xi\cdot\|u_\nu^*\|_\infty + \eta \le 2\xi(\ell')(\log n+1)+\eta,
\end{aligned}$$

where the second inequality follows from the fact that $\frac{1}{m}\sum_{j\in S_g}X_j$ is a vector in $\triangle^n$ and thus has L_2 norm at most 1, and the penultimate step holds with probability $2|\mathcal{N}|e^{-8m\xi^2}$.

So if $\eta = \delta/2$ and $\xi = \frac{\delta}{4(\ell')(\log n+1)}$, then as long as

$$m = \Omega(\xi^{-2}\log|\mathcal{N}|) = \Omega\left(\frac{\log(1/\delta)}{\delta}\cdot\ell^4\log^7 n\right),$$

then with probability at least $1-\text{poly}(n)$, there exists an SoS proof of (14) using Axioms 5.4. □

Lemma 5.12. *Under the canonical assignment, with high probability there is some choice of $\{(P_S^{T_1,T_2})_{\gamma,\rho}\}$ and $\{(Q_S^{T_1,T_2})_{\gamma,\rho}\}$ for which Constraints 5 and 7 are satisfied.*

The proof of Lemma 5.12 is conceptually very similar to that of Lemma 5.8, so we defer it to the appendix of the full version.

REFERENCES

[1] J. Acharya, I. Diakonikolas, C. Hegde, J. Li, and L. Schmidt. 2015. Fast and Near-Optimal Algorithms for Approximating Distributions by Histograms. In *PODS*.

[2] Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. 2017. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 1278–1289.

[3] D. Achlioptas and F. McSherry. 2005. On Spectral Learning of Mixtures of Distributions. In *Proceedings of the Eighteenth Annual Conference on Learning Theory (COLT)*. 458–469.

[4] Frank J Anscombe. 1960. Rejection of outliers. *Technometrics* 2, 2 (1960), 123–146.

[5] S. Arora and R. Kannan. 2001. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd Symposium on Theory of Computing*. 247–257.

[6] F. Balabdaoui, K. Rufibach, and J. A. Wellner. 2009. Limit Distribution Theory for Maximum Likelihood Estimation of a Log-Concave Density. *The Annals of Statistics* 37, 3 (2009), pp. 1299–1331.

[7] F. Balabdaoui and J. A. Wellner. 2007. Estimation of a k-Monotone Density: Limit Distribution Theory and the Spline Connection. *The Annals of Statistics* 35, 6 (2007), pp. 2536–2564.

[8] Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. 2017. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*. 169–212.

[9] Richard E Barlow, David J Bartholomew, James M Bremner, and H Daniel Brunk. 1972. *Statistical inference under order restrictions: The theory and application of isotonic regression*. Technical Report. Wiley New York.

[10] L. Birgé. 1987. Estimating a density under order restrictions: Nonasymptotic minimax risk. *Annals of Statistics* 15, 3 (1987), 995–1012.

[11] Hugh D Brunk. 1955. Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics* (1955), 607–616.

[12] H. D. Brunk. 1958. On the Estimation of Parameters Restricted by Inequalities. *The Annals of Mathematical Statistics* 29, 2 (1958), pp. 437–454.

[13] K.S. Chan and H. Tong. 2004. Testing for multimodality with dependent data. *Biometrika* 91, 1 (2004), 113–123.

[14] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. 2014. Near-Optimal Density Estimation in Near-Linear Time Using Variable-Width Histograms. In *NIPS*. 1844–1852.

[15] Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. 2013. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 1380–1394.

[16] Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. 2014. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. ACM, 604–613.

[17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. 2017. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 47–60.

[18] Sitan Chen, Jerry Li, and Ankur Moitra. 2020. Learning Structured Distributions From Untrusted Batches: Faster and Simpler. *arXiv preprint arXiv:2002.10435* (2020).

[19] Sitan Chen and Ankur Moitra. 2019. Beyond the low-degree algorithm: mixtures of subcubes and their applications. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 869–880.

[20] Richard Cole and Tim Roughgarden. 2014. The sample complexity of revenue maximization. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. ACM, 243–252.

[21] S. Dasgupta. 1999. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*. 634–644.

[22] S. Dasgupta and L. Schulman. 2000. A two-round variant of EM for Gaussian mixtures. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*. 143–151.

[23] Constantinos Daskalakis, Anindya De, Gautam Kamath, and Christos Tzamos. 2016. A size-free CLT for poisson multinomials and its applications. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, 1074–1086.

[24] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R.A. Servedio, and L. Tan. 2013. Learning Sums of Independent Integer Random Variables. In *FOCS*. 217–226.

[25] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. 2012. Learning k-modal distributions via testing. In *SODA*. 1371–1385.

[26] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. 2012. Learning Poisson Binomial Distributions. In *STOC*. 709–728.

[27] Luc Devroye and Gabor Lugosi. 2001. *Combinatorial Methods in Density Estimation*. Springer Science & Business Media.

[28] Ilias Diakonikolas. 2016. Learning Structured Distributions. *Handbook of Big Data* 267 (2016).

[29] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2019. Robust Estimators in High-Dimensions Without the Computational Intractability. *SIAM J. Comput.* 48, 2 (2019), 742–864.

[30] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2017. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 999–1008.

[31] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. 2016. The fourier transform of poisson multinomial distributions and its algorithmic applications. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, 1060–1073.

[32] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. 2016. Optimal learning via the fourier transform for sums of independent integer random variables. In *Conference on Learning Theory*. 831–849.

[33] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. 2016. Properly learning poisson binomial distributions in almost polynomial time. In *Conference on Learning Theory*. 850–878.

[34] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. 2018. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 1047–1060.

[35] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. 1995. Wavelet shrinkage: asymptopia. *Journal of the Royal Statistical Society, Ser. B* (1995), 371–394.

[36] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. 1996. Density estimation by wavelet thresholding. *Ann. Statist.* 24, 2 (1996), 508–539.

[37] L. Dumbgen and K. Rufibach. 2009. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli* 15, 1 (2009), 40–68.

[38] J. Feldman, R. O'Donnell, and R. Servedio. 2005. Learning mixtures of product distributions over discrete domains. In *FOCS 2005*. 501–510.

[39] A.-L. Fougères. 1997. Estimation de densités unimodales. *Canadian Journal of Statistics* 25 (1997), 375–387.

[40] F. Gao and J. A. Wellner. 2009. On the rate of convergence of the maximum likelihood estimator of a k-monotone density. *Science in China Series A: Mathematics* 52 (2009), 1525–1538. Issue 7.

[41] U. Grenander. 1956. On the theory of mortality measurement. *Skand. Aktuarietidskr.* 39 (1956), 125–153.

[42] P. Groeneboom. 1985. Estimating a monotone density. In *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*. 539–555.

[43] D. L. Hanson and G. Pledger. 1976. Consistency in Concave Regression. *The Annals of Statistics* 4, 6 (1976), pp. 1038–1050.

[44] Clifford Hildreth. 1954. Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.* 49, 267 (1954), 598–619.

[45] Samuel B Hopkins and Jerry Li. 2018. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 1021–1034.

[46] Zhiyi Huang, Yishay Mansour, and Tim Roughgarden. 2018. Making the most of your samples. *SIAM J. Comput.* 47, 3 (2018), 651–674.

[47] Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics*. Springer, 492–518.

[48] Ayush Jain and Alon Orlitsky. 2019. Robust Learning of Discrete Distributions from Batches. *arXiv preprint arXiv:1911.08532* (2019).

[49] Ayush Jain and Alon Orlitsky. 2020. A General Method for Robust Learning from Batches. *arXiv preprint arXiv:2002.11099* (2020).

[50] H. K. Jankowski and J. A. Wellner. 2009. Estimation of a discrete monotone density. *Electronic Journal of Statistics* 3 (2009), 1567–1605.

[51] A. T. Kalai, A. Moitra, and G. Valiant. 2010. Efficiently learning mixtures of two Gaussians. In *STOC*. 553–562.

[52] Sushrut Karmalkar, Pravesh Kothari, and Adam Klivans. 2019. List-Decodable Linear Regression. *arXiv preprint arXiv:1905.05679* (2019).

[53] G. Kerkyacharian, D. Picard, and K. Tribouley. 1996. Lp Adaptive Density Estimation. *Bernoulli* 2, 3 (1996), pp. 229–247.

[54] Adam Klivans, Pravesh K Kothari, and Raghu Meka. 2018. Efficient Algorithms for Outlier-Robust Regression. In *Conference On Learning Theory*. 1420–1430.

[55] R. Koenker and I. Mizera. 2010. Quasi-concave density estimation. *Ann. Statist.* 38, 5 (2010), 2998–3027. https://doi.org/10.1214/10-AOS814

[56] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).

[57] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. 2018. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 1035–1046.

[58] Kevin A Lai, Anup B Rao, and Santosh Vempala. 2016. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 665–674.

[59] Rafał Latała. 1997. Estimation of moments of sums of independent real random variables. *The Annals of Probability* 25, 3 (1997), 1502–1513.

[60] Reut Levi, Dana Ron, and Ronitt Rubinfeld. 2013. Testing properties of collections of distributions. *Theory of Computing* 9, 1 (2013), 295–347.

[61] Jerry Li and Ludwig Schmidt. 2017. Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. In *Conference on Learning Theory*. 1302–1382.

[62] Jerry Zheng Li. 2018. *Principled approaches to robust machine learning and beyond*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[63] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. http://arxiv.org/abs/1602.05629

[64] A. Moitra and G. Valiant. 2010. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*. 93–102.

[65] Carl M O'Brien. 2016. Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics. *International Statistical Review* 84, 2 (2016), 318–319.

[66] Mingda Qiao and Gregory Valiant. 2017. Learning discrete distributions from untrusted batches. *arXiv preprint arXiv:1711.08113* (2017).

[67] Prasad Raghavendra, Tselil Schramm, and David Steurer. 2018. High-dimensional estimation via sum-of-squares proofs. *arXiv preprint arXiv:1807.11419* (2018).

[68] Prasad Raghavendra and Morris Yau. 2019. List Decodable Learning via Sum of Squares. *arXiv preprint arXiv:1905.04660* (2019).

[69] B.L.S. Prakasa Rao. 1969. Estimation of a unimodal density. *Sankhya Ser. A* 31 (1969), 23–36.

[70] R. A. Redner and H. F. Walker. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* 26 (1984), 195–202.

[71] Jacob Steinhardt. 2018. *Robust Learning: Information Theory and Algorithms*. Ph.D. Dissertation. Stanford University.

[72] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. 2018. Resilience: A Criterion for Learning in the Presence of Arbitrary Outliers. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[73] C. J. Stone. 1994. The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation. *The Annals of Statistics* 22, 1 (1994), pp. 118–171.

[74] C. J. Stone, M. H. Hansen, C. Kooperberg, and Y. K. Truong. 1997. Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture. *Ann. Statist.* 25, 4 (1997), 1371–1470.

[75] Kevin Tian, Weihao Kong, and Gregory Valiant. 2017. Learning populations of parameters. In *Advances in Neural Information Processing Systems*. 5778–5787.

[76] Aleksandr Filippovich Timan. 2014. *Theory of approximation of functions of a real variable*. Vol. 34. Elsevier.

[77] John W Tukey. 1960. A survey of sampling from contaminated distributions. *Contributions to probability and statistics* (1960), 448–485.

[78] John W Tukey. 1975. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, Vol. 2. 523–531.

[79] Vladimir Vapnik and Alexey Chervonenkis. 1974. Theory of pattern recognition.

[80] S. Vempala and G. Wang. 2002. A Spectral Algorithm for learning mixtures of distributions. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*. 113–122.

[81] G. Walther. 2009. Inference and Modeling with Log-concave Distributions. *Statist. Sci.* 24, 3 (2009), 319–327.

[82] E.J. Wegman. 1970. Maximum likelihood estimation of a unimodal density. I. and II. *Ann. Math. Statist.* 41 (1970), 457–471, 2169–2174.

[83] Edward J Wegman. 1970. Maximum likelihood estimation of a unimodal density function. *The Annals of Mathematical Statistics* 41, 2 (1970), 457–471.

[84] E. J. Wegman and I. W. Wright. 1983. Splines in Statistics. *J. Amer. Statist. Assoc.* 78, 382 (1983), pp. 351–365.

[85] R. Willett and R. D. Nowak. 2007. Multiscale Poisson Intensity and Density Estimation. *IEEE Transactions on Information Theory* 53, 9 (2007), 3171–3187.