Analytic Techniques for Robust Algorithm Design

Thesis Proposal

Ainesh Bakshi
Computer Science Department
Carnegie Mellon University
abakshi@cs.cmu.edu

Thesis Committee

Pravesh K. Kothari (CMU, Co-Chair)
David P. Woodruff (CMU, Co-Chair)
Ryan O'Donnell (CMU)
Boaz Barak (Harvard)
Santosh S. Vempala (Georgia Tech)

Abstract

Modern machine learning relies on algorithms that fit expressive models to large datasets. While such tasks are easy in low dimensions, real-world datasets are truly high-dimensional. Additionally, a prerequisite to deploying models in real-world systems is to ensure that their behavior degrades gracefully when the modelling assumptions no longer hold. Therefore, there is a growing need for *efficient algorithms* that fit reliable and robust models to data.

In this thesis proposal, we focus on designing such efficient, robust and provable algorithms for fundamental tasks in machine learning and statistics. In particular, we investigate two complementary themes arising in this area: *high-dimensional robust statistics* and *fast numerical linear algebra*. The first addresses how to fit expressive models to high-dimensional noisy datasets and the second develops fast algorithmic primitives to reduce dimensionality and de-noise large datasets. We resolve central open questions in robust statistics and randomized linear algebra, and introduce several new algorithmic ideas along the way. Finally, we make the case for analytic techniques, such as convex relaxations, being the natural choice for robust algorithm design.

Contents

1 Introduction

Modern machine learning relies on algorithms that fit expressive models to large datasets. While such tasks are easy in low dimensions, real-world datasets are truly high-dimensional. Furthermore, a prerequisite to deploying models in real-world systems is to ensure that their behavior degrades gracefully when the modelling assumptions no longer hold. Therefore, there is a growing need for *efficient algorithms* that fit reliable and robust models to data.

This thesis proposal focuses on the burgeoning area of designing efficient, robust and provable algorithms for fundamental tasks arising in machine learning. The long-term goal of this research program is to provide a unified algorithmic theory for such tasks. This program is an intensely active area of research and has already witnessed several breakthroughs, including adverasially robust high-dimensional statistics [DKK+19, LRV16], topic modelling [AGM12], learning mixture models [MV10], and tensor decompositions [BCMV14]. These breakthroughs have led to many new algorithmic insights and have been influential in designing the latest robust algorithms for sensitive practical applications.

In particular, in this thesis proposal, we focus on two complementary themes arising from this program: *high-dimensional robust statistics* and *fast numerical linear algebra*. The first investigates how to fit expressive models to high-dimensional noisy datasets and the second develops fast algorithmic primitives to reduce dimensionality and de-noise large datasets. For instance, fitting a linear model (regression) or a Gaussian mixture model to a high-dimensional dataset and reducing dimensionality of the dataset are oftentimes complementary and core algorithmic tasks in numerous machine learning and data analysis applications.

Ensuring robustness and reliability remains a long-standing challenge in such applications when the underlying data is noisy. We present several recent results on designing robust, efficient, and reliable algorithms for fundamental statistical and linear algebraic tasks in modern machine learning. The algorithms we develop draw upon tools from convex and polynomial optimization, high-dimensional probability, random matrix theory, functional analysis and convex geometry. Further, the algorithms are accompanied with provable guarantees on their correctness and performance. Finally, we distill unifying ideas across different settings that lead to robust algorithm design and discuss potential directions for future work.

1.1 Overview

Robust Statistics. Classical works in statistics study fitting a model to a set of data points, with the crucial assumption that the data we observe is noise-free and drawn i.i.d. from a reasonable distribution. However, as early as the 60's, statisticians already realized that real-world datasets are noisy and are unlikely to fit idealized statistics models [Hub64]. The sources of such noise can range from systematic bias and error in data collection to malicious tampering. Finding efficient algorithms to fit robust models to noisy data has spawned a decades-long research program spanning many academic disciplines (see [DK19, RSS18] for recent surveys).

One such fundamental task is to robustly fit a mixture of Gaussians to noisy input data. Finding an efficient algorithm for this task was also highlighted as a central open problem at the Foundations of Big Data workshop at the Simons Institute [DVW18]. Recent results addressed the special cases of robustly estimating a single Gaussian [DKK+19, LRV16], a mixture of mean-separated Gaussians [HL18, KSS18, DKS18], TV-distance separated Gaussians [BK20],[DHKK20], and a uni-

form mixture of two Gaussians [Kan20]. In joint work with Diakonikolas, Jia, Kane, Kothari and Vempala [BDJ+20], we completely resolved this problem and obtained the first polynomial-time algorithm for robustly estimating a mixture of k arbitrary Gaussians.

In a similar vein, yet another fundamental task in high-dimensional statistics is to robustly fit a linear function when the input is noisy and potentially drawn from a heavy-tailed distribution. In a sequence of breakthrough works in 2018, the first efficiently computable estimators for robust linear regression were obtained [KKM18, PSBR20, DKS19]. A central open question that came out of this sequence of works was whether the information-theoretically optimal rate of convergence can be obtained efficiently (see [KKM18]). In joint work with Adarsh Prasad [BP21], we obtained a polynomial time algorithm for robust regression that achieves the information-theoretically optimal convergence rate. This problem continues to be of significant interest, and several concurrent and followup works [ZJS20, CAT+20, PJL20, JLST21] improve the sample complexity and running time requirement in certain special cases, however do not obtain optimal rates always.

A closely related line of work, initiated by Balcan, Blum and Vempala [BBV08], considers learning from data in the presence of an overwhelming fraction of outliers. In this setting, it is information-theoretically impossible to recover the true parameters (such as mean or covariance). However, we can relax the goal to output a small list of candidate parameters, with at least one close to the true solution. In joint work with Pravesh Kothari [BK21], we give the first polynomial time algorithm for recovering the subspace spanned by the uncorrupted samples in the presence of overwhelming outliers. We also significantly improve the running time for regression, a problem originally considered in [RY20a, KKK19].

A common theme in all the aforementioned results is to identify analytic properties of probability distributions and design algorithms that encode these properties as polynomial constraints. Surprisingly, the algorithms we obtain by following this blueprint are automatically robust to corruptions in the input data. Therefore, we believe that the mathematical machinery we developed (based on the sum-of-squares convex hierarchy) lays the foundations of a unified theory of robust algorithmic statistics.

Fast Numerical Linear Algebra. Randomized algorithms for linear algebra problems have been at the center of the modern machine learning revolution. This is because matrices yield a simple representation of large-scale data and machine learning algorithms exploit efficient manipulations that can be performed on matrices (see recent surveys [Woo14, BHK20, KV17] and references therein). In this proposal, we focus on improving the running time of such key algorithmic primitives, including dimensionality reduction, principal component analysis (low-rank approximation) and testing properties of large matrices, such as positive semi-definiteness.

Low-rank approximation is one of the most common dimensionality reduction techniques, whereby one replaces a large matrix with a low rank factorization. It is easy to see that in the worst case, we must read the entire input matrix to compute such a factorization. However, when the input matrix has structure, we showed that we can avoid reading most of the matrix and compute a factorization in sub-linear time. In joint work with David Woodruff [BW18], we obtained the first sub-linear time algorithm for factoring distance matrices that arise from an arbitrary finite metric over n points. Further, in joint work with Chepurko and Woodruff [BCW20], we obtained an information-theoretically optimal algorithm for factoring positive semi-definite matrices and Euclidean distance matrices, improving prior work of Musco and Musco [MW17]. Additionally,

we obtained optimal algorithms for regression when the design matrix is PSD and initiated the study of robust low-rank approximation.

A closely related problem is to determine whether a given matrix is positive semi-definite (PSD). Classical algorithms to solve this problem require computing the entire spectrum, which can be computationally prohibitive for large matrices. In joint work with Chepurko and Jayaram [BCJ20], we formalized the problem of testing whether a matrix is PSD or separated from the PSD cone in Operator or Frobenius norm. We showed that a natural algorithm determines which of the two cases the input matrix belongs to, without reading most of the input matrix. In fact, the time/queries required is independent of the dimension of the matrix and only depends on the distance from the PSD cone.

In each of the aforementioned problems, we leverage the fact that the input isn't an arbitrary matrix, but rather drawn, potentially adversarially, from a structured family, such as PSD matrices. A similar phenomenon has been observed for several optimization problems, inlcuding solving linear systems for Laplacian/Diagonally Dominant matrices [ST14, KOSZ13, KMP14] and Block Henkel matrices [PV21], covariance estimation of Toeplitz matrices [ELMM20], and approximation the permanent of boolean [JS89], non-negative matrices [JSV04] and PSD [AGGS17, YP21] matrices. The algorithms we design for low-rank approximation are robust to small perturbations in the input that may falsify assumptions such as PSD-ness. However, the tools we introduce are specialized to each problem. A general theory of when structure in the input matrices can be exploited for a given class of optimization problems remains an outstanding research direction.

2 Robust Algorithmic Statistics

Given a collection of observations and a class of models, , the objective of a typical learning algorithm is to find the model in the class that best fits the data. Such learning algorithms often assume that the input data are i.i.d. samples generated by a statistical model in the given class. This is a simplifying assumption that is, at best, only approximately valid, as real datasets are typically exposed to some source of systematic noise. Robust statistics [Hub04, HRRS11] challenges this assumption by focusing on the design of *outlier-robust* estimators – algorithms that can tolerate a *constant fraction* of corrupted datapoints, and achieve error that is independent of the dimension. Despite significant effort over several decades starting with important early works of Tukey and Huber in the 60s, until fairly recently, even for the most basic high-dimensional estimation tasks, all known computationally efficient estimators were highly sensitive to outliers.

This state of affairs changed with two independent works [DKK+19, LRV16], which gave the first computationally efficient and outlier-robust learning algorithms for estimating the mean and covariance of a single high-dimensional Gaussian distribution. Since these initial works, we have witnessed substantial research progress on algorithmic aspects of robust high-dimensional estimation by several communities, including theoretical computer science, machine learning, and mathematical statistics (see [DK19] for a recent survey). However, several central questions remained open, as discussed in the subsequent subsections.

We begin by precisely defining the corruption model we consider. We work in the strong contamination model, which generalized several well-studied noise models, including the Huber contamination model [Hub64].

Definition 2.1 (Strong Contamination Model)**.** Given a parameter $\epsilon \in (0, 1/2)$ and a class of distributions $\mathcal{D}$ over $\mathbb{R}^d$, the adversary is computationally unbounded and operates as follows: the algorithm specifies a number of samples, n, and n i.i.d. samples are drawn from some unknown $D \in \mathcal{D}$. The adversary is allowed to inspect the samples, remove up to ϵn samples and replace them with arbitrary points in $\mathbb{R}^d$. The modified set is given as input to the algorithm. We call such a set an ϵ-corrupted sample.

Various communities have also considered less powerful adversaries, giving rise to weaker contamination models. For instance, an adversary may be adaptive or oblivious to the inliers, only allowed to add outliers, or only allowed to remove inliers. Next, we consider fitting basic models, such as Gaussian Mixture models and regression, in the presence of adversarial outliers.

2.1 Gaussian Mixture Models

The Gaussian Mixture Model (GMM) has been the subject of a century-old line of research beginning with Pearson [Pea94]. Progress on provable algorithms for learning GMMs began with the influential work of Dasgupta [Das99], yielding clustering algorithms that succeed under various separation assumptions [AK05, VW04, AM05, BV08]. These assumptions, however, do not capture natural separated instances of Gaussians, such as separation in distribution (total variation) distance. A more general approach [MV10, BS15] circumvents clustering altogether by giving an efficient algorithm for parameter estimation without any separation assumptions. However, this approach is brittle to even adversarially corrupting a single input point and crucially relies on the algebraic structure of Gaussians. A natural question to ask is then as follows:

Question 2.2. *Is there an efficient and robust algorithm to learn the parameters of arbitrary mixtures of k Gaussians?*

This question, and several special cases has received a lot of attention over the years. Clustering a mixture of k Gaussians is an important special case of this problem, where each pair of components of the mixture is nearly completely separated in total variation distance. Until recently, no efficient robust algorithm was known even for clustering a mixture of two well-separated Gaussians.

We begin by formally defining a Gaussian Mixture model:

Definition 2.3 (Gaussian Mixture Model)**.** A mixture of k Gaussians is a probability distributon, denoted by $\mathcal{D} = \sum_{i\in[k]} p_i \cdot \mathcal{N}(\mu_i, \Sigma_i)$, where for all $i \in [k]$, $\mu_i \in \mathbb{R}^d$ and $\Sigma_i \in \mathbb{R}^{d\times d}$ is a set of k means and covariances respectively, $p_i \geqslant 0$, and $\sum_{i\in[k]} p_i = 1$. A sample from $\mathcal{D}$ is generated by picking component i with probability p_i and then outputting an i.i.d. sample from $\mathcal{N}(\mu_i, \Sigma_i)$.

Additionally, to measure closeness between two distributions, we use total variation (TV) distance.

Definition 2.4 (Total Variation Distance)**.** Given two distributions p and q, we define the total variation distance between them as follows:

$$d_{\mathrm{TV}}(p, q) = \frac{1}{2}\int_{-\infty}^{\infty} |p(x) - q(x)|dx.$$

We first consider the special case where the input mixture is clusterable, i.e. all components of the mixture are pairwise separated in TV distance.

2.1.1 Robustly Clustering a Mixture of Gaussians

In recent work with Pravesh Kothari [BK20], we obtained the first polynomial-time algorithm based on the sum-of-squares (SoS) method for clustering TV-separated k-GMMs in the presence of a small fraction of fully adversarial outliers. Conceptually, our algorithm relies on analytic (as opposed to algebraic) properties of Gaussians, which we believe are closely tied to designing robust algorithms. Formally,

Theorem 2.5 (Outlier-Robust Clustering of k-GMMs)**.** *Fix $\eta, \epsilon > 0$. Let $\mathcal{D} = \sum_{i\in[k]} \frac{1}{k}\mathcal{N}(\mu_i, \Sigma_i)$ be a k-GMM such that for all $i \neq i'$, $d_{TV}(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_i, \Sigma_i)) \geqslant 1 - \exp(-(k/\eta)^c)$, for a fixed constant c. Then, there exists an algorithm that takes input an ϵ-corruption Y of a sample $X \sim \mathcal{D}$ such that $X = C_1 \cup C_2 \cup \ldots \cup C_k$, with equal sized clusters C_i corresponding to points drawn from $\mathcal{N}(\mu_i, \Sigma_i)$, and with probability at least 0.99, outputs an approximate clustering $Y = \hat{C}_1 \cup \hat{C}_2 \cup \ldots \cup \hat{C}_k$ satisfying $\min_{i\leqslant k} \frac{|\hat{C}_i \cap C_i|}{|C_i|} \geqslant 1 - O(k^{2k})(\epsilon + \eta)$. The algorithm succeeds whenever $n = |X| \geqslant d^{\mathrm{poly}(k/\eta)}$ and runs in time $n^{\mathrm{poly}(k/\eta)}$.*

We can use off-the-shelf robust estimators for mean and covariance of Gaussians([DKK+19]) in order to get statistically optimal estimates of the mean and covariances of the target k-GMM.

Corollary 2.6 (Parameter Recovery from Clustering)**.** *In the setting of Theorem 2.5, with the same running time, sample complexity and success probability, our algorithm can output $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$ such that for some permutation $\pi : [k] \to [k]$,*

$$d_{TV}\left(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}\left(\hat{\mu}_{\pi(i)}, \hat{\Sigma}_{\pi(i)}\right)\right) \leqslant \tilde{O}\left(k^{2k}(\epsilon + \eta)\right),$$

where $\tilde{O}$ surpresses polylogarithmic factors in k, η and ϵ.

We note that a similar result was independently and concurrently obtained by [DHKK20] resulting in a merge [BDH+20].

Discussion. We obtain the first outlier-robust algorithm that works for clustering k-GMMs under information-theoretically minimal separation assumptions. Such results were not known even for $k = 2$. To discuss the bottlenecks in prior works, it is helpful to use following consequence of two Gaussians with means μ_1, μ_2 and covariances Σ_1, Σ_2 being at a TV distance $\geqslant 1 - \exp(-O(\Delta^2))$ in terms of the distance between their parameters.

Definition 2.7 (Δ-Separated Mixture Model)**.** An equi-weighted mixture $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_k$ with parameters $\{\mu_i, \Sigma_i\}_{i\in[k]}$ is Δ-separated if for every pair of distinct components i, j, one of the following three conditions hold ($\Sigma^{\dagger/2}$ is the square root of pseudo-inverse of Σ):

1. **Mean-Separation:** $\exists v \in \mathcal{R}^d$ such that
$$\langle \mu_i - \mu_j, v\rangle^2 > \Delta^2 \cdot v^\top(\Sigma_i + \Sigma_j)v,$$
2. **Spectral-Separation:** $\exists v \in \mathcal{R}^d$ such that
$$v^\top \Sigma_i v > \Delta \cdot v^\top \Sigma_j v,$$
3. **Relative-Frobenius Separation:**[1] Σ_i and Σ_j have the same range space and
$$\left\|\Sigma_i^{\dagger/2}\Sigma_j\Sigma_i^{\dagger/2} - I\right\|_F^2 > \Delta^2 \cdot \left\|\Sigma_i^{\dagger/2}\Sigma_j^{1/2}\right\|_{op}^4.$$

We show that two Gaussians separated in TV distance can be separated in any of the aforementioned notions of parameter distance. The key bottleneck for known algorithms prior to our work was handling separation in Spectral and Relative Frobenius distance (cases 2 and 3 above).

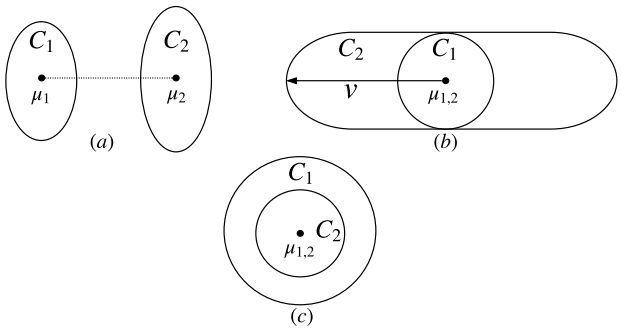

Figure 1: (a) Mean Separation (b) Spectral Separation (c) Relative Frobenius Separation

Often real-world data need not be Gaussian, and our algorithm does not overfit to this assumption. It succeeds for mixtures of all distributions that satisfy two well-studied analytic conditions:

[1] Unlike the other two distances, relative Frobenius distance is meaningful only for high-dimensional Gaussians. As an illustrative example, consider two 0 mean Gaussians with covariances $\Sigma_1 = I$ and $\Sigma_2 = (1 + \Theta(1/\sqrt{d}))I$. Then, for large enough d, the parameters are separated in relative Frobenius distance but not spectral or mean distance.

anti-concentration and *hypercontractivity*. In particular, we formulate these conditions as polynomial inequalities and obtain algorithms that can efficiently verify them. We thus move beyond Pearson's method of moments and consider identifying clean analytic conditions that enable the existence of efficient and robust clustering algorithms an important contribution of our work. We note that such a result for non-Gaussian distributions was not known, even with access to unbounded computation.

Next, we define the precise analytic conditions we require and we refer the reader to Appendix A for background on sum-of-squares proofs.

Definition 2.8 (Certifiable Hypercontractivity of Degree-2 polynomials)**.** An isotropic distribution $\mathcal{D}$ on $\mathcal{R}^d$ is said to be h-certifiably C-hypercontractive if there's a degree h sum-of-squares proof of the following unconstrained polynomial inequality in $d \times d$ matrix-valued indeterminate Q:

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Qx - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx\right]\right)^h\right] \leqslant (Ch)^h \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Qx - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx\right]\right)^2\right]\right)^{h/2},$$

A set of points $X \subseteq \mathcal{R}^d$ is said to be C-certifiably hypercontractive if the uniform distribution on X is h-certifiably C-hypercontractive.

Hypercontractivity is an important notion in high-dimensional probability and analysis on product spaces [O'D14]. Kauers, O'Donnell, Tan and Zhou [KOTZ14] showed certifiable hypercontractivity of Gaussians and more generally product distributions with subgaussian marginals. Certifiable hypercontractivity strictly generalizes the better known *certifiable subgaussianity* property (studied first in [KSS18]) that controls higher moments of linear polynomials.

In contrast to hypercontractivity, anti-concentration forces *lower-bounds* of the form $\mathbf{Pr}[\langle x, v\rangle^2 \geqslant \delta \|v\|_2^2] \geqslant \delta'$, for all directions v. Certifiable anti-concentration was recently introduced in independent works of Karmalkar, Klivans and Kothari [KKK19] and Raghavendra and Yau [RY20a] and later used [BK21],[RY20b] for the related problems of list-decodable linear regression and subspace recovery[2].

Following [KKK19], we formulate certifiable anti-concentration via a univariate, even polynomial $p_{\delta,\Sigma}$ that uniformly approximates the 0-1 core-indicator $1(\langle x, v\rangle^2 \geqslant \delta v^\top \Sigma v)$ over a large enough interval around 0. Let $q_{\delta,\Sigma}(x, v)$ be a multivariate (in v) polynomial defined by $q_{\delta,\Sigma}(x, v) = \left(v^\top \Sigma v\right)^{2s} p_{\delta,\Sigma}\left(\frac{\langle x,v\rangle}{\sqrt{v^\top \Sigma v}}\right)$.Since $p_{\delta,\Sigma}$ is an even polynomial, $q_{\delta,\Sigma}$ is a polynomial in v.

Definition 2.9 (Certifiable Anti-Concentration)**.** A mean 0 distribution D with covariance Σ is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated if for $q_{\delta,\Sigma}(x, v)$ defined above, there exists a degree-$2s$ sum-of-squares proof of the following two unconstrained polynomial inequalities in indeterminate v:

$$\left\{\langle x, v\rangle^{2s} + \delta^{2s} q_{\delta,\Sigma}(x, v)^2 \geqslant \delta^{2s}\left(v^\top \Sigma v\right)^{2s}\right\}, \left\{\mathop{\mathbb{E}}_{x\sim D}\left[q_{\delta,\Sigma}(x, v)^2\right] \leqslant C\delta\left(v^\top \Sigma v\right)^{2s}\right\},$$

An isotropic subset $X \subseteq \mathcal{R}^d$ is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated if the uniform distribution on X is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated.

[2]List-decodable versions of these problems generalize the "mixture" variants - mixed linear regression and subspace clustering - that are easily seen to be special cases of mixtures of k-Gaussians with TV separation 1.

Remark 2.10. For natural examples, $s(\delta) \leqslant 1/\delta^c$ for some fixed constant c. For e.g., $s(\delta) = O(\frac{1}{\delta^2})$ for standard Gaussian distribution and the uniform distribution on the unit sphere (see [KKK19] and [BK21]). To simplify notation, we will assume $s(\delta) \leqslant \mathrm{poly}(1/\delta)$ in the statement of our results.

Additionally, we need that the variance of degree-2 polynomials is bounded in terms of the Frobenius norm of the coefficients of the polynomial. Formally,

Definition 2.11 (Degree-2 Polynomials with Certifiably Bounded Variance)**.** A mean 0 distribution $\mathcal{D}$ with covariance Σ certifiably bounded variance degree 2 polynomials if there is a degree 2 sum-of-squares proof of the following inequality in the indeterminate $Q \in \mathbb{R}^{d\times d}$

$$\left\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\left(x^\top Q x - \mathbf{E}_{x\sim\mathcal{D}} x^\top Q x \right)^2 \right] \leqslant C \left\| \Sigma^{1/2} Q \Sigma^{1/2} \right\|_F^2 \right\},$$

Our general result gives an outlier-robust clustering algorithm for separated mixtures of *reasonable* distributions, i.e., distributions that satisfies both certifiable hypercontractivity, anti-concentration and have bounded variance of degree-2 polynomials. Even the information-theoretic (and without outliers, i.e., $\epsilon = 0$) clusterability of such distributions was not known prior to our work.

Theorem 2.12 (Outlier-Robust Clustering of Reasonable Mixtures)**.** *Fix $\eta > 0, \epsilon > 0$. Let $\mathcal{D}$ be a Δ-separated mixture of reasonable distributions. Then, there exists an algorithm that takes input an ϵ-corruption Y of a sample $X = C_1 \cup C_2 \cup \ldots C_k$, with true clusters C_i of size n/k drawn i.i.d. from $\mathcal{D}$ and outputs an approximate clustering $Y = \hat{C}_1 \cup \hat{C}_2 \cup \ldots \cup \hat{C}_k$ satisfying $\min_{i\leqslant k} \frac{|\hat{C}_i \cap C_i|}{|C_i|} \geqslant 1 - O(k^{2k})(\epsilon + \eta)$. The algorithm succeeds with probability at least 0.99 over the draw of the original sample X whenever $n \geqslant d^{\mathrm{poly}(k/\eta)}$ and runs in time $n^{\mathrm{poly}(k/\eta))}$ whenever $\Delta \geqslant \mathrm{poly}(k/\eta)^k$.*

Future Directions. The class of distributions that satisfy certifiable hypercontractivity of degree-2 polynomials is quite broad, and includes all strongly log-concave distributions. However, all existing approaches can establish certifiable anti-concentration only for rotiationally invariant distributions and affine transformations thereof. Therefore, a natural open question is as follows:

Open Question 2.13 (Charecterizing Certifiable Anti-Concentration)**.** What class of distributions (beyond rotationally invariant distributions) admit low-degree sum-of-squares certificates ?

Further, the certificates we establish, even for Gaussian distributions, require a degree that grows polynomialy with δ, the bound on the expectation. The running time of our algorithm scales exponentially in the degree required above and thus improved bounds lead to significantly faster algorithms. Moreover, such an improvement would lead to milder assumptions on the TV separation between the components.

Open Question 2.14 (Degree of Certifiable Anti-Concentration)**.** What is the minimum degree required to establish $(\delta, C\delta)$-certifiable anti-concentration for Gaussian distributions? Is a polynomial dependence on δ neccessary?

2.1.2 Robustly Learning a Mixture of Arbitrary Gaussians

Building on [BK20], in joint work with Diakonikolas, Jia, Kane, Kothari and Vempala, [BDJ+20] we were able to completely answer the aforementioned central question (Question 2.2) in the affirmative, by providing an efficient and robust algorithm that learns the parameters of all mixtures of k Gaussians, thereby resolving this central question in high-dimensional statistics. Our result requires the information-theoretically minimum assumptions on the input mixture, is robust a small fraction of adversarial corruptions and is provably faster than the existing non-robust algorithm of Moitra-Valiant [MV10]. Formally,

Theorem 2.15 (Robustly Learning k Arbitrary Gaussians)**.** *There is an algorithm with the following behavior: Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)} (1/\epsilon)^{c_k}$ samples from a distribution F on $\mathcal{R}^d$ such that $d_{TV}(F, \mathcal{M}) \leqslant \epsilon$, for an unknown target k-GMM $\mathcal{D} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, the algorithm runs in time $\mathrm{poly}(n)\,(1/\epsilon)^{c_k'}$ and outputs a k-GMM hypothesis $\widehat{\mathcal{D}} = \sum_{i=1}^{k} \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ such that with high probability we have that $d_{TV}\left(\widehat{\mathcal{D}}, \mathcal{D}\right) \leqslant \mathcal{O}\left(\varepsilon^{1/c_k''}\right)$, where c_k, c_k', c_k'' depends only on k.*

A number of works have made algorithmic progress on important special cases of the above problem, including faster robust clustering for the spherical case under minimal separation conditions [HL18, KSS18, DKS18], robust clustering for separated (and potentially non-spherical) Gaussian mixtures [BDH+20], and robustly learning *uniform* mixtures of two arbitrary Gaussian components [Kan20]. A similar result was independently and concurrently obtained by [LM21], under slightly stronger assumptions, and using completely different techniques.

Theorem 2.15 gives the first polynomial-time *robust proper learning algorithm*, with dimension-independent error guarantee, for *arbitrary* k-GMMs, for any fixed k. Known Statistical Query lower bounds [DKS17] suggest that $d^{\Omega(k)}$ samples are necessary for efficiently learning GMMs, for approximation to constant accuracy, even in the (much simpler) noiseless setting and when the components are pairwise well-separated in total variation distance. This provides evidence that the sample-time tradeoff achieved by our result is qualitatively optimal.

Further, we show that *the same algorithm* also achieves the stronger parameter estimation guarantee. We note that parameter estimation requires some assumptions on the underlying mixture. The following corollary applies under the standard assumption that any pair of components in the unknown mixture has total variation distance at least ϵ^{c_k}, where c_k only depends on k.

Corollary 2.16 (Robust Parameter Estimation)**.** *Let $\mathcal{D} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$ be an unknown target k-GMM satisfying the following conditions: (i) $d_{TV}(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_j, \Sigma_j)) \geqslant \varepsilon^{f_1(k)}$ for all $i \neq j$, and (ii) $S = \{i \in [k] : w_i \geqslant \varepsilon^{f_2(k)}\}$ is a subset of $[k]$, where $f_1(k), f_2(k)$ are sufficiently small functions of k. Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)} (1/\varepsilon)^{c_k}$ samples from a distribution F on $\mathcal{R}^d$ such that $d_{TV}(F, \mathcal{D}) \leqslant \epsilon$, there exists an algorithm that runs in time $\mathrm{poly}(n)\,(1/\varepsilon)^{c_k'}$ and outputs a k'-GMM hypothesis $\widehat{\mathcal{D}} = \sum_{i=1}^{k'} \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ with $k' \leqslant k$ such that with high probability there exists a bijection $\pi : S \to [k']$ satisfying the following: For all $i \in S$, it holds that $|w_i - \widehat{w}_{\pi(i)}| \leqslant \mathrm{poly}_k(\epsilon)$ and $d_{TV}\left(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\widehat{\mu}_{\pi(i)}, \widehat{\Sigma}_{\pi(i)})\right) \leqslant \epsilon^{1/c_k''}$.*

Discussion. *Handling Arbitrary Weights*: Our algorithm succeeds without any assumptions on the weights of the mixture components. We emphasize that this is an important feature and not a technicality. Prior and concurrent work cannot handle the case of general weights – even for the

case of $k = 2$ components! Obtaining a fully polynomial-time algorithm for the general case (i.e., one not incurring an exponential cost in $1/w_{\min}$) requires genuinely new algorithmic ideas and is one of the key technical innovations of the aforementioned result.
Handling Arbitrary Covariances: Our algorithm does not require assumptions on the eigenvalues of the component covariances, modulo basic limitations posed by numerical computation issues. Specifically, our algorithm works even if some of the component covariances are rank-deficient (i.e., have directions of 0 variance) with running time scaling polynomially in the bit-complexity of the unknown component means and covariances. Such a dependence on the bit complexity of the input parameters is unavoidable – there exist[3] examples of rank-deficient covariances with irrational entries such that the total variation distance between the corresponding Gaussian and every Gaussian with covariance matrix of rational entries is the maximum possible value of one.

Overview. In the non-robust setting (i.e., for $\epsilon = 0$), the algorithm of [MV10] solves this learning problem. The key idea of [MV10] is to observe that if a mixture of k Gaussians has every pair of components separated in total variation distance by at least δ, then a random univariate projection of the mixture has a pair of components that are $\delta/\sqrt{d}$-separated in total variation distance. Their algorithm uses this observation to piece together estimates of the mixture when projected to several carefully chosen directions to get an estimate of the high-dimensional mixture. Notice, however, that such a strategy meets with instant roadblock in the presence of outliers: the fraction of outliers, being a dimension-independent constant, completely overwhelms the total variation distance between components in any one direction making them indistinguishable.

Our robust algorithm is based on three new ingredients:

1. a new and efficient *partial clustering algorithm* based on the sum-of-squares (SoS) method,
2. a novel *list-decodable tensor decomposition* method, and
3. a recursive *spectral separation* method.

We briefly describe these ideas below and how they can be interleaved to obtain our algorithm.

Efficient Partial Clustering. We call a mixture partially clusterable if it contains a pair of components at total variation distance larger than $1 - \Omega_k(1)$. Interestingly, it turns out that the clustering algorithm of [BK20] (Theorem 2.12) can be generalized to the partial clustering setting, i.e., the setting where we are guaranteed to have a pair of components that are well-separated (with no guarantees on the remaining components). For a mixture with minimum mixing weight α, this gives an algorithm with running time of $d^{(k/\alpha)^{O(k)}}$ to partition the input sample into components so that each piece of the partition is (effectively) a $(\mathrm{poly}(\alpha/k) + \epsilon)$-corrupted sample from disjoint sub-mixtures.

By applying the above partial-clustering algorithm, we can effectively assume that the input is an ϵ-corrupted sample from a mixture with every pair of components *at most* $(1 - \Omega_k(1))$-far in total variation distance. Then, we use robust covariance estimation (see Theorem 7.1 in [BK20]) to make the mixture approximately isotropic, i.e. the mean of the mixture is ≈ 0 and the covariance of the mixture is $\approx I$ (in Frobenius norm).

[3]For example, for the unit vector $v = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, 0, \ldots, 0)$, for every choice of rational covariance Σ, the total variation distance between $\mathcal{N}(0, I - vv^\top)$ and $\mathcal{N}(0, \Sigma)$ is one.

After partial clustering and an approximate isotropic transformation, every pair of components are close in TV distance. Under this condition, in order to learn the unknown mixture with error guarantees in total variation distance, it suffices to obtain $\text{poly}_k(\epsilon)$-error estimates of the μ_i, Σ_i's in Frobenius norm. As we will see soon, this will suffice for our weaker result that has an exponential dependence on the minimum mixing weight.

To get a fully polynomial algorithm, we delve a bit deeper: the exponential dependence on the minimum mixing weight is incurred only when two components are spectrally separated (see Definition 2.7, which in turn relies on the degree required for certifiable anti-concentration. Instead, we give a new partial clustering algorithm that works in fixed polynomial time, whenever there is a pair of Gaussian components separated either via their means or the relative Frobenius distance. The resulting clusters might now have components that are spectrally separated, a difficulty that we address later.

List-decodable Tensor Decomposition. Kane [Kan20] gave a polynomial-time algorithm to robustly learn an *equiweighted* mixture of two Gaussians. For this special case, after isotropic transformation, one can effectively assume that the two means are $\pm\mu$ and the two covariances are $I \pm \Sigma$. Kane's idea was to use the Hermite tensor (which can be built using the 4-th and 6-th raw moments of the mixture). Since we must use outlier-robust estimates of these tensors, we can only obtain estimates that are accurate up to constant error in Frobenius norm of the tensor. Kane's key observation is that for the special case of $k = 2$ components, one can build two different Hermite tensors, one of which is rank-one with component $\approx \mu$ (and thus one can immediately "read off" μ); the other only has a tensor power of Σ. This second tensor is of the form $\hat{T}_4 = \text{Sym}((\Sigma - I) \otimes (\Sigma - I)) + E$, where $\|E\|_F = O_k(\sqrt{\epsilon})$ and Sym refers to symmetrizing over all possible permutations of the "4 modes of the tensor". Unlike the case of the mean, one cannot simply "read-off"[4] Σ from T_4, but Kane gives a simple method to accomplish this. As noted in [Kan20], it is not clear how to extend this to non-equiweighted mixtures of $k = 2$ Gaussians, and going to even $k = 3$ components requires substantially new ideas.

The surprising fact that we establish is that by looking at only the first four moments of our mixture, we can learn all of the components up to low-rank error, i.e., up to errors along a bounded number of hidden directions. Thus, the new tensor decomposition has both Frobenius norm error and low-rank error. To see the idea, it is helpful to focus on the simpler case where all the means are zero. In this case, the estimated 4th Hermite tensor (built from estimated raw moments of degree at most 4 of the mixture) has the following :

$$\hat{T}_4 = \sum_{i=1}^{k} w_i \text{Sym}((\Sigma_i - I) \otimes (\Sigma_i - I) + E) \ ,$$

where E is a 4-tensor with $\|E\|_F = O_k(\sqrt{\epsilon})$.

Given the form of this tensor, it is natural to consider tensor decomposition algorithms, by thinking of $\Sigma_i - I$ as a d^2-dimensional vector. However, we run into the issue of uniqueness of tensor decomposition, since we are dealing with 2nd order tensors (once we view $\Sigma_i - I$ as a d^2-dimensional vector). One might imagine computing higher-order tensors of similar forms to

[4] It is helpful to visualize a single entry of this tensor for, say, the case when i, j, k, ℓ are all distinct: $\hat{T}_4(i,j,k,\ell) = \frac{1}{3}(\Sigma(i,j)\Sigma(k,\ell) + \Sigma(i,k)\Sigma(j,\ell) + \Sigma(i,\ell)\Sigma(j,k)) + error$. Notice that obtaining entries of Σ from T_4 is formally a task of solving noisy quadratic equations.

overcome the uniqueness issues, but this runs into two major complications: first, the symmetrization operation introduces spurious terms that do not have the sum of tensor-power structure required for such an algorithm to succeed.

Second, even if one were to get hold of the tensor without the symmetrization operation, the only applicable tensor decomposition algorithm (recall that we do not make *any* genericity assumptions on the components that are typically required by tensor decomposition algorithms) is the result of Barak, Kelner, and Steurer [BKS15]. However, the [BKS15] result, while being efficient in its dependence on the number of components, has exponential dependence on the target error, which is prohibitively expensive for our application.

Rather than recovering the unique decomposition of the tensor $\hat{T}_4$ above, we instead produce a list of candidate decompositions. To do this, we start by applying an operation that is a common trick in most tensor decomposition algorithms. In our context, this trick amounts to taking a random matrix (with independent standard Gaussian entries) P and "collapsing" the last two modes of $\hat{T}_4$ with P (i.e., computing $\hat{S}(i,j) = \sum_{k,\ell} \hat{T}_4(i,j,k,\ell)P(k,\ell)$) to obtain a matrix Q. In the usual tensor decomposition procedures, we are interested in proving that one can recover all the information about the components of the tensor from Q.

Spectral Separation of Thin Components. While the running time of our partial clustering and tensor decomposition algorithms are now polynomial, the guarantees of the tensor decomposition subroutine we discussed above are no longer enough to guarantee a recovery of parameters that result in a mixture close in total variation distance. Because of the three conditions that we assumed in the working of the tensor decomposition algorithm, we can no longer guarantee the third one that gives a lower bound on the smallest eigenvalue of every covariance (relative to the covariance of the mixture). In particular, we can end up in a situation where, even though we have a list of parameters that contain Frobenius-norm-close estimates of the covariances, the estimates do not provide a total variation distance guarantee. (Consider a "skinny" direction where the variance of some component is very small, or even 0, forcing us to learn the parameters more precisely!)

It turns out that the above is the only way the algorithm can fail at this point — one or more covariance matrices have a very small eigenvalue (if not, the Frobenius norm error would imply TV-distance error). But since we have estimates of the covariances, we can find such a small eigenvector. Now we observe that since the mixture is nearly isotropic (i.e., the overall variance in each direction is ~ 1), if some component has very small variance along a direction, then the components must be separable along this direction. We show that it is possible to efficiently cluster the mixture after projecting it to this direction, so that each cluster has strictly fewer components. We then recursively apply the entire algorithm on the clusters obtained, which will each have strictly fewer components.

Future Directions. A natural question arising from our work is to characterize the class of distributions such that their mixtures can be learned, even information-theoretically.

Open Question 2.17. Are there mixtures of non-Gaussian distributions that can be learned robustly/non-robustly and information-theoretically/efficiently? Is there a statistical-computational gap between any of these settings?

We note that the aforementioned algorithm is not entirely captured by the sum-of-squares proof system. This leads to the following question:

Open Question 2.18. Can the sum-of-squares proof system efficiently learn a mixutre of k arbitrary Gaussians?

We hope that answering some of these question, along with the techniques we have developed can pave the way for robustly learning various popular latent variable models.

2.2 Linear Regression

Regression continues to be extensively studied under various models, including realizable regression (no noise), true linear models (independent noise), asymmetric noise, agnostic regression and generalized linear models (see [Wei05] and references therein). In each model, a variety of distributional assumptions are considered over the covariates and the noise. As a consequence, there exist innumerable estimators for regression achieving various trade-offs between sample complexity, running time and rate of convergence. The presence of adversarial outliers adds yet another dimension to design and compare estimators.

Seminal works on robust regression focused on designing non-convex loss functions, including M-estimators [Hub04], Theil-Sen estimators [The92, Sen68], R-estimators [Jae72], Least-Median-Squares [Rou84] and S-estimators [RY84]. These estimators have desirable statistical properties under disparate assumptions, yet remain computationally intractable in high dimensions. Further, recent works show that it is information-theoretically impossible to design robust estimators for linear regression without distributional assumptions [KKM18].

An influential recent line of work showed that when the data is drawn from the well studied and highly general class of *hypercontractive* distributions (see Definition 2.8), there exist robust and computationally efficient estimators for regression [KKM18, PSBR20, DKS19]. Several families of natural distributions fall into this category, including Gaussians, strongly log-concave distributions and product distributions on the hypercube. However, both estimators converge to the the true hyperplane (in ℓ_2-norm) at a sub-optimal rate, as a function of the fraction of corrupted points.

Given the vast literature on ad-hoc and often incomparable estimators for high-dimensional robust regression, the central question we address in this work is as follows:

Does there exist a unified approach to design robust and computationally efficient estimators achieving optimal rates for all linear regression models under mild distributional assumptions?

We address the aforementioned question by introducing a framework to design robust estimators for linear regression when the input is drawn from a *hypercontractive* distribution. Our estimators converge to the true hyperplanes at the information-theoretically optimal rate (as a function of the fraction of corrupted data) under various well-studied noise models, including independent and agnostic noise. Further, we show that our estimators can be computed in polynomial time using the *sum-of-squares* convex hierarchy.

In classical regression, we assume $\mathcal{D}$ is a distribution over $\mathcal{R}^d \times \mathcal{R}$ and for a vector $\Theta \in \mathcal{R}^d$, the least-squares loss is given by $\mathsf{err}_{\mathcal{D}}(\Theta) = \mathbb{E}_{x,y\sim\mathcal{D}}\left[\left(y - x^\top\Theta\right)^2\right]$. The goal is to learn $\Theta^* = \arg\min_\Theta \mathsf{err}_{\mathcal{D}}(\Theta)$. We assume sample access to $\mathcal{D}$, and given n i.i.d. samples, we want to obtain a vector Θ that approximately achieves optimal error, $\mathsf{err}_{\mathcal{D}}(\Theta^*)$. In contrast to the classical setting, we work in the *strong contamination model*, defined above.

Model 2.19 (Robust Regression Model)**.** *Let $\mathcal{D}$ be a distribution over $\mathcal{R}^d \times R$ such that the marginal distribution over $\mathcal{R}^d$ is centered and has covariance Σ^* and let $\Theta^* = \arg\min_\Theta \mathbb{E}_{x,y\sim\mathcal{D}}\left[(y - \langle\Theta, x\rangle)^2\right]$ be the optimal hyperplane for $\mathcal{D}$. Let $\{(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots (x_n^*, y_n^*)\}$ be n i.i.d. random variables drawn from $\mathcal{D}$. Given $\epsilon > 0$, the robust regression model $\mathcal{R}_\mathcal{D}(\epsilon, \Sigma^*, \Theta^*)$ outputs a set of n samples $\{(x_1, y_1), \ldots (x_n, y_n)\}$ such that for at least $(1-\epsilon)n$ points $x_i = x_i^*$ and $y_i = y_i^*$. The remaining ϵn points are arbitrary, and potentially adversarial w.r.t. the input and estimator.*

A natural starting point is to assume that the marginal distribution over the covariates (the x's above) is heavy-tailed and has bounded, finite covariance. However, we show that there is no robust estimator in this setting, even when the linear model has no noise and the uncorrupted points lie on a line.

Theorem 2.20 (Bounded Covariance does not suffice)**.** *For all $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{R}^d \times \mathcal{R}$ such that $d_{TV}(\mathcal{D}_1, \mathcal{D}_2) \leqslant \epsilon$ and the marginal distribution over the covariates has bounded covariance, denoted by $I \preceq \Sigma \preceq O(1)I$, yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega(1)$, where Θ_1 and Θ_2 are the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$.*

The aforementioned result precludes any statistical estimator that converges to the true hyperplane as the fraction of corrupted points tends to 0. Therefore, we strengthen the distributional assumption consider hypercontractive distributions instead.

Definition 2.21 (Certifiable Hypercontractivity)**.** A distribution $\mathcal{D}$ on $\mathcal{R}^d$ is (c_k, k)-certifiably hypercontractive if for all $r \leqslant k/2$, there exists a degree $\mathcal{O}(k)$ sum-of-squares proof (defined below) of the following inequality in the variable v

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\langle x, v\rangle^{2r}\right] \leqslant \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[c_r \langle x, v\rangle^2\right]^r$$

such that $c_r \leqslant c_k$.

Remark 2.22. Hypercontractivity captures a broad class of distributions, including Gaussian distributions, uniform distributions over the hypercube and sphere, affine transformations of isotropic distributions satisfying Poincare inequalities [KSS18] and strongly log-concave distributions. Further, hypercontractivity is preserved under natural closure properties like affine transformations, products and weighted mixtures [KSS18].

In this work we focus on the *rate of convergence* of our estimators to the true hyperplane, Θ^*, as a function of the fraction of corrupted points, denoted by ϵ. We measure convergence in both parameter distance (ℓ_2-distance between the hyperplanes) and least-squares error on the true distribution ($\mathsf{err}_\mathcal{D}$).

We introduce a simple analytic condition on the relationship between the noise (marginal distribution over $y - x^\top\Theta^*$) and covariates (marginal distribution over x) that can be considered as a proxy for independence of $y - x^\top\Theta^*$ and x :

Definition 2.23 (Negatively Correlated Moments)**.** Given a distribution $\mathcal{D}$ over $\mathcal{R}^d \times \mathcal{R}$, such that the marginal distribution on $\mathcal{R}^d$ is (c_k, k)-hypercontractive, the corresponding regression instance has negatively correlated moments if for all $r \leqslant k$, and for all v,

$$\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\langle v, x\rangle^r \left(y - x^\top\Theta^*\right)^r\right] \leqslant \mathcal{O}(1) \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\langle v, x\rangle^r\right] \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y - x^\top\Theta^*\right)^r\right]$$

Informally, the *negatively correlated moments* condition can be viewed as a polynomial relaxation of independence of random variables. Note, it is easy to see that when the noise is independent of the covariates, the above definition is satisfied.

Remark 2.24. We show that when this condition is satisfied by the true distribution, $\mathcal{D}$, we obtain rates that match the information theoretically optimal rate in a *true linear model*, where the noise (marginal distribution over $y - x^\top \Theta^*$) is independent of the covariates (marginal distribution over x). Further, when this condition is not satisfied, we show that there exist distributions for which obtaining rates matching the *true linear model* is impossible.

When the distribution over the input is hypercontractive and has negatively correlated moments, we obtain an estimator achieving *rate* proportional to $\epsilon^{1-1/k}$ for parameter recovery. Further, our estimator can be computed efficiently. Thus, our main algorithmic result is as follows:

Theorem 2.25 (Robust Regresssion with Negatively Correlated Noise)**.** *Given $\epsilon > 0, k \geqslant 4$, and $n \geqslant (d \log(d))^{\mathcal{O}(k)}$ samples from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$, such that $\mathcal{D}$ is (c,k)-certifiably hypercontractive and has negatively correlated moments, there exists an algorithm that runs in $n^{\mathcal{O}(k)}$ time and outputs an estimator $\tilde{\Theta}$ such that with high probability,*

$$\left\| (\Sigma^*)^{1/2} \left(\Theta^* - \tilde{\Theta} \right) \right\|_2 \leqslant \mathcal{O}\left(\epsilon^{1-1/k} \right) \left(err_{\mathcal{D}}(\Theta^*)^{1/2} \right)$$

and,

$$err_{\mathcal{D}}(\tilde{\Theta}) \leqslant \left(1 + \mathcal{O}\left(\epsilon^{2-2/k} \right) \right) err_{\mathcal{D}}(\Theta^*)$$

Remark 2.26. We note that prior work does not draw a distinction between the independent and dependent noise models. In comparison (see Table 1), Klivans, Kothari and Meka [KKM18] obtained a sub-optimal least-squares error scales proportional to $\epsilon^{1-2/k}$. For the special case of $k = 4$, Prasad et. al. [PSBR20] obtain least squares error proportional to $O(\epsilon \kappa^2(\Sigma))$, where κ is the condition number. In very recent independent work Zhu, Jiao and Steinhardt [ZJS20] obtained a sub-optimal least-squares error scales proportional to $\epsilon^{2-4/k}$.

Further, we show that the rate we obtained in Theorem 2.25 is information-theoretically optimal, even when the noise and covariates are independent:

Theorem 2.27 (Lower Bound for Independent Noise)**.** *For any $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{R}^2 \times \mathcal{R}$ such that the marginal distribution over $\mathcal{R}^2$ has covariance Σ and is (c,k)-hypercontractive for both distributions, and yet $\left\| \Sigma^{1/2}(\Theta_1 - \Theta_2) \right\|_2 = \Omega\left(\epsilon^{1-1/k} \sigma \right)$, where Θ_1, Θ_2 are the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, $\sigma = \max(err_{\mathcal{D}_1}(\Theta_1), err_{\mathcal{D}_2}(\Theta_2))$ and the noise is uniform over $[-\sigma, \sigma]$. Further, $|err_{\mathcal{D}_1}(\Theta_2) - err_{\mathcal{D}_1}(\Theta_1)| = \Omega\left(\epsilon^{2-2/k} \sigma^2 \right)$.*

Next, we consider the setting where the noise is allowed to arbitrary, and need not have negatively correlated moments with the covariates. A simple modification to our algorithm and analysis yields an efficient estimator that obtains rate proportional to $\epsilon^{1-2/k}$ for parameter recovery.

Corollary 2.28 (Robust Regresssion with Dependent Noise)**.** *Given $\epsilon > 0, k \geqslant 4$ and $n \geqslant (d \log(d))^{\mathcal{O}(k)}$ samples from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$, such that $\mathcal{D}$ is (c,k)-certifiably hypercontractive, there exists an algorithm that runs in $n^{\mathcal{O}(k)}$ time and outputs an estimator $\tilde{\Theta}$ such that with probability 9/10,*

$$\left\| (\Sigma^*)^{1/2} \left(\Theta^* - \tilde{\Theta} \right) \right\|_2 \leqslant \mathcal{O}\left(\epsilon^{1-2/k} \right) \left(err_{\mathcal{D}}(\Theta^*)^{1/2} \right),$$

Estimator	Independent Noise	Arbitrary Noise
Prasad et. al. [PSBR20], Diakonikolas et. al. [DKK+18]	$\epsilon\,\kappa^2$ (only $k=4$)	$\epsilon\,\kappa^2$ (only $k=4$)
Klivans, Kothari and Meka [KKM18]	$\epsilon^{1-2/k}$	$\epsilon^{1-2/k}$
Zhu, Jiao and Steinhardt [ZJS20]	$\epsilon^{2-4/k}$	$\epsilon^{2-4/k}$
Our Work Thm 2.25, Cor 2.28	$\epsilon^{2-2/k}$	$\epsilon^{2-4/k}$
Lower Bounds Thm 2.27, Thm 2.29	$\epsilon^{2-2/k}$	$\epsilon^{2-4/k}$

Table 1: Comparison of convergence rate (for least-squares error) achieved by various computationally efficient estimators for Robust Regression, when the underlying distribution is (c_k, k)-hypercontractive.

and,

$$err_{\mathcal{D}}(\tilde{\Theta}) \leqslant \left(1 + \mathcal{O}\left(\epsilon^{2-4/k}\right)\right) err_{\mathcal{D}}(\Theta^*).$$

Further, we show that the dependence on ϵ is again information-theoretically optimal:

Theorem 2.29 (Lower Bound for Dependent Noise)**.** *For any $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{R}^2 \times \mathcal{R}$ such that the marginal distribution over $\mathcal{R}^2$ has covariance Σ and is (c,k)-hypercontractive for both distributions, and yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega\left(\epsilon^{1-2/k}\sigma\right)$, where Θ_1, Θ_2 be the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively and $\sigma = \max(err_{\mathcal{D}_1}(\Theta_1), err_{\mathcal{D}_2}(\Theta_2))$. Further, $|err_{\mathcal{D}_1}(\Theta_2) - err_{\mathcal{D}_1}(\Theta_1)| = \Omega\left(\epsilon^{2-4/k}\sigma^2\right)$.*

Overview. Consider two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ over $\mathcal{R}^d \times \mathcal{R}$ such that the total variation distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ is ϵ and the marginals for both distributions over $\mathcal{R}^d$ are (c_k, k)-hypercontractive and have covariance Σ. Ignoring computational and sample complexity concerns, we can obtain the optimal hyperplanes corresponding to each distribution. Note, these hyperplanes need not be unique and are simply charecterized as minimzers of the least-squares loss : for $i \in \{1, 2\}$,

$$\Theta_i = \arg\min_{\Theta} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_i}\left[\left(y - x^\top\Theta\right)^2\right]$$

Our central contribution is to obtain an information theoretic proof that the optimal hyperplanes are indeed close in scaled ℓ_2 norm, i.e.

$$\left\|\Sigma^{1/2}\left(\Theta_1 - \Theta_2\right)\right\|_2 \leqslant \mathcal{O}\left(\epsilon^{1-1/k}\right)\left(\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_1}\left[\left(y - x^\top\Theta_1\right)^2\right]^{1/2} + \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_2}\left[\left(y - x^\top\Theta_2\right)^2\right]^{1/2}\right)$$

Further, we show that the $\epsilon^{1-1/k}$ dependence can be achieved even when the noise is not completely independent of the covariates but satisfies an analytic condition which we refer to as *negatively correlated moments* (see Definition 2.23). We provide an outline of the proof as it illustrates the techniques we introduced in this work.

Coupling and Decoupling. We begin by considering a maximal coupling, $\mathcal{G}$, between distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ such that they disagree on at most an ϵ-measure support (ϵ-fraction of the points for a discrete distribution). Let $(x, y) \sim \mathcal{D}_1$ and $(x', y') \sim \mathcal{D}_2$. Then, observe for any vector v,

$$\begin{aligned}\langle v, \Sigma(\Theta_1 - \Theta_2)\rangle &= \left\langle v, \mathop{\mathbb{E}}_{\mathcal{G}}\left[xx^\top\right](\Theta_1 - \Theta_2)\right\rangle \\ &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x\left(x^\top\Theta_1 - y\right)\right\rangle\right] + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x\left(y - x^\top\Theta_2\right)\right\rangle\right]\end{aligned} \tag{1}$$

While the first term in Equation (8) depends completely on $\mathcal{D}_1$, the second term requires using the properties of the maximal coupling. Since $1 = 1_{(x,y)=(x',y')} + 1_{(x,y)\neq(x',y')}$, we can rewrite the second term in Equation (8) as follows:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x\left(y - x^\top\Theta_2\right)\right\rangle\right] &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - (x')^\top\Theta_2\right)\right\rangle 1_{(x,y)=(x',y')}\right] \\ &\quad + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x\left(y - x^\top\Theta_2\right)\right\rangle 1_{(x,y)\neq(x',y')}\right]\end{aligned} \tag{2}$$

With a bit of effort, we can combine Equations (8) and (2), and upper bound them as follows:

$$\begin{aligned}\langle v, \Sigma(\Theta_1 - \Theta_2)\rangle \leqslant \mathcal{O}(1)\Bigg(&\underbrace{\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x\left(x^\top\Theta_1 - y\right)\right\rangle\right]}_{\text{(i)}} + \underbrace{\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left((x')^\top\Theta_2 - y'\right)\right\rangle\right]}_{\text{(ii)}} \\ &+ \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x\left(y - x^\top\Theta_1\right)\right\rangle 1_{(x,y)\neq(x',y')}\right] \\ &+ \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - (x')^\top\Theta_2\right)\right\rangle 1_{(x,y)\neq(x',y')}\right]\Bigg)\end{aligned} \tag{3}$$

Observe, since we have a maximal coupling, the last two terms appearing in Equation (3) are non-zero only on an ϵ-measure support. To bound them, we decouple the indicator using Hölder's inequality,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x(y - x^\top\Theta_1)\right\rangle 1_{(x,y)\neq(x',y')}\right] &\leqslant \mathbb{E}\left[1_{(x,y)\neq(x',y')}\right]^{\frac{k-1}{k}} \mathbb{E}\left[\langle v, x\rangle^k\left(y - x^\top\Theta_1\right)^k\right]^{\frac{1}{k}} \\ &\leqslant \epsilon^{1-1/k} \cdot \underbrace{\mathbb{E}\left[\langle v, x\rangle^k\left(y - x^\top\Theta_1\right)^k\right]^{\frac{1}{k}}}_{\text{(iii)}}\end{aligned} \tag{4}$$

where we used the maximality of the coupling $\mathcal{G}$ to bound $\mathbb{E}\left[1_{(x,y)\neq(x',y')}\right] \leqslant \epsilon$. The above analysis can be repeated verbatim for the second term in (3) as well. Going forward, we focus on bounding terms (i), (ii) and (iii).

Gradient Conditions. To bound terms (i) and (ii) in Equation (3), we crucially rely on *gradient information* provided by the least-squares objective. Concretely, a key observation in our information-theoretic proof is that the candidate hyperplanes are locally optimal: given least-squares loss, for

$i \in \{1, 2\}$ for all vectors v,

$$\left\langle \nabla \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_i} \left[\left(y - x^\top \Theta_i \right)^2 \right], v \right\rangle = \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_i} \left[\left\langle v, xx^\top \Theta_i - xy \right\rangle \right] = 0$$

where Θ_1 and Θ_2 are the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively. Therefore, both (i) and (ii) are identically 0. It remains to bound (iii).

Independence and Negatively Correlated Moments. We observe that term (iii) can be interpreted as the k-th order correlation between the distribution of the covariates projected along v and the distribution of the noise in the linear model. Here, we observe that if the linear model satisfies the *negatively correlated moments* condition (Definition 2.23), we can decouple the expectation and bound each term independently:

$$\mathbb{E} \left[\langle v, x \rangle^k \left(y - x^\top \Theta_1 \right)^k \right]^{1/k} \leqslant \mathbb{E} \left[\langle v, x \rangle^k \right]^{1/k} \mathbb{E} \left[\left(y - x^\top \Theta_1 \right)^k \right]^{1/k} \tag{5}$$

Observe, when the underlying linear model has independent noise, Equation (5) follows for any k. We thus crucially exploit the structure of the noise and require a considerably weaker notion than independence. Further, if the *negatively correlated moments* property is not satisfied, we can use Cauchy-Schwarz to decouple the expectation in Equation (5) and incur a $\epsilon^{1-2/k}$ dependence. Conceptually, we emphasize that the *negatively correlated moments* condition may be of independent interest to design estimators that exploit independence in various statistics problems.

Hypercontractivity. To bound the RHS in Equation (5), we use our central distributional assumption of hypercontractive k-th moments (Definition 2.8) of the covariates :

$$\mathbb{E} \left[\langle v, x \rangle^k \right]^{1/k} \leqslant \sqrt{c_k}\, \mathbb{E} \left[\langle v, x \rangle^2 \right]^{1/2} = \sqrt{c_k}\, \langle v, \Sigma v \rangle^{1/2}$$

We can bound the noise similarly, by assuming that the noise is hypercontractive and this considerably simplifies our statements. However, hypercontractivity of the noise is not a necessary assumption and prior work indeed incurs a term proportional to the k-th moment of the noise. Assuming boundedness of the regression vectors, Klivans, Kothari and Meka [KKM18] obtained a uniform upper bound on k-th moment of the noise by truncating large samples. We note that the same holds for our estimators and we refer the reader to Section 5.2.3 in their paper. Finally, substituting $v = \Theta_1 - \Theta_2$ and rearranging, completes the information-theoretic proof.

We note that our approach already differs from prior work [KKM18, PSBR20, ZJS19] and to our knowledge, we obtain the first information theoretic proof that being ϵ-close in TV distance implies that the optimal hyperplanes are $\mathcal{O}\left(\epsilon^{1-1/k}\right)$ close in ℓ_2 distance.

Future Directions. We note that our estimators obtain the rate matching recent work for Gaussians, albeit in quasi-polynomial time. In comparison, Diakonikolas, Kong and Stewart [DKS18] obtain the same rate in polynomial time, when the noise is independent of the covariates. This leads to the following question

Open Question 2.30 (Sub-Gaussian Rates in Polynomial Time)**.** Is there a polynomial time algorithm that achieves $O(\epsilon \cdot \mathrm{poly}(\log(1/\epsilon)))$ rates for all sub-Gaussian distributions? Is any extra $\log(1/\epsilon)$ factor neccesary?

Further, the sample complexity of our estimators scales proportional to $d^{\Omega(k)}$. Such large sample complexity may not be neccesary.

Open Question 2.31 (Sub-Gaussian Rates in Polynomial Time)**.** Can we achieve the optimal tradeoff between sample complexity, running time and rate for all hypercontractive distributions?

A natural generalization of our work is to consider robust algorithms for Generalized Linear Models, which capture linear, logistic and multi-response regression. Further, such algorithms would pave the way for robust estimators for learning Graphical Models that have received significant attention in various machine learning and computational biology domains. Thus far, obtaining the statistically optimal rate for learning simple Graphical Models remains open, even with unbounded computation [LSS$^+$]. A closely related problem is that of list-decodable regression and subspace recovery, where an overwhelming fraction of data is corrupted (see for example [KKK19, RY20a, RY20b][BK21]). Studying variants of regression and latent variable models in the list-decodable setting is ripe for future work.

3 Randomized Numerical Linear Algebra

In the second half of this thesis proposal, we study the power of randomization for several basic tasks in numerical linear algebra, including computing a low-rank approximation and testing PSD-ness. Traditionally, algorithms for such tasks involve computing the entire spectrum of the matrix, and even assuming access to fast matrix multiplication, the running time is at least n^ω. We study structural properties under which we can obtain algorithms that perform provably better than performing matrix multiplication. Further, we analyze how robust these algorithms are to perturbations of the corresponding structral properties.

3.1 Low-Rank Approximation

Low-rank approximation is one of the most common dimensionality reduction techniques, whereby one replaces a large matrix $\mathbf{A}$ with a low-rank factorization $\mathbf{U} \cdot \mathbf{V} \approx \mathbf{A}$. Such a factorization provides a compact way of storing $\mathbf{A}$ and allows one to multiply $\mathbf{A}$ quickly by a vector. It is used as an algorithmic primitive in clustering [DFK+04, McS01], recommendation systems [DKR02], web search [AFKM01, Kle99], and learning mixtures of distributions [AM05, KSV05], and has numerous other applications.

A large body of recent work has looked at *relative-error* low-rank approximation, whereby given an $n \times n$ matrix $\mathbf{A}$, an accuracy parameter $\epsilon > 0$, and a rank parameter k, one seeks to output a rank-k matrix $\mathbf{B}$ for which

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2, \tag{6}$$

where for a matrix $\mathbf{C}$, $\|\mathbf{C}\|_F^2 = \sum_{i,j} \mathbf{C}_{i,j}^2$, and $\mathbf{A}_k$ denotes the best rank-k approximation to $\mathbf{A}$ in Frobenius norm. $\mathbf{A}_k$ can be computed exactly using the singular value decomposition, but takes time $O(n^\omega)$, where ω is the matrix multiplication constant. We refer the reader to the survey [Woo14] and references therein.

For worst-case matrices, it is not hard to see that any algorithm achieving (6) must spend at least $\Omega(\text{nnz}(\mathbf{A}))$ time, where $\text{nnz}(\mathbf{A})$ denotes the number of non-zero entries (sparsity) of $\mathbf{A}$. Indeed, without reading most of the non-zero entries of $\mathbf{A}$, one could fail to read a single large entry, thus making one's output matrix $\mathbf{B}$ an arbitrarily bad approximation.

A flurry of recent work [KP16, MW17, CLW18, Tan19, RSML18, GLT18, IVWW19, SW19, GSLW19] has looked at the possibility of achieving *sublinear* time algorithms (classical and quantum) for low-rank approximation. In particular, Musco and Woodruff [MW17] consider the important case of positive-semidefinite (PSD) matrices. PSD matrices include as special cases covariance matrices, correlation matrices, graph Laplacians, kernel matrices and random dot product models. Further, the special case where the input itself is low-rank (PSD Matrix Completion) has applications in quantum state tomography [GLF+10]. Subsequently, Bakshi and Woodruff [BW18] considered low-rank approximation of the closely related family of Negative-type (Euclidean Squared) distance matrices. Negative-type metrics include as special cases ℓ_1 and ℓ_2 metrics, spherical metrics and hypermetrics, as well as effective resistances in graphs [DL09, TD87, CRR+96, CKM+11]. Negative-type metrics have found various applications in algorithm design and optimization [ALN08, SS11, KMP14].

Musco and Woodruff show that it is possible to output a low-rank matrix $\mathbf{B}$ in factored form achieving (6) in $\widetilde{O}(nk/\epsilon^{2.5} + nk^{\omega-1}/\epsilon^{2(\omega-1)})$ time, while reading only $\widetilde{O}(nk/\epsilon^{2.5})$ entries of $\mathbf{A}$. They

also showed a lower bound that any algorithm achieving (6) must read $\Omega(nk/\epsilon)$ entries, and closing the gap between these bounds has remained an open question. Similarly, in joint work with David Woodruff, we exploit the structure of Negative-type metrics to reduce to the PSD case and obtain a bi-criteria algorithm that requires $\widetilde{O}(nk/\epsilon^{2.5})$ queries. The gap in the sample complexity and the requirement of a bi-criteria guarantee remained open. We resolve these both these questions here.

Next we consider PSD matrices that have been corrupted by a small amount of noise. A drawback of algorithms achieving (6) is that they cannot tolerate any amount of unstructured noise. For instance, if one slightly corrupts a few off-diagonal entries, making the input matrix $\mathbf{A}$ no longer PSD, then it is impossible to detect such corruptions in sublinear time, making the relative-error guarantee (6) information-theoretically impossible. Motivated by this, we also introduce a new framework where an adversary corrupts the input by adding a noise matrix $\mathbf{N}$ to a psd matrix $\mathbf{A}$. We assume that the Frobenius norm of the corruption is bounded relative to the Frobenius norm of $\mathbf{A}$, i.e., $\|\mathbf{N}\|_F^2 \leqslant \eta\|\mathbf{A}\|_F^2$. We also assume the corruption is well-spread, i.e., each row of $\mathbf{N}$ has ℓ_2^2-norm at most a fixed constant factor larger than ℓ_2^2-norm of the corresponding row of $\mathbf{A}$.

This model captures small perturbations to PSD matrices that we may observe in real-world datasets, as a consequence of round-off or numerical errors in tasks such as computing Laplacian pseudoinverses, and systematic measurement errors when computing a covariance matrix. One important application captured by our model is low-rank approximation of corrupted *correlation matrices*. Finding a low-rank approximation of such matrices occurs when measured correlations are asynchronous or incomplete, or when models are stress-tested by adjusting individual correlations. Low-rank approximation of correlation matrices also has many applications in finance [Hig02].

Given that it is information-theoretically impossible to obtain the relative-error guarantee (6) in the *robust model*, we relax our notion of approximation to the following well-studied additive-error guarantee:

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leqslant \|\mathbf{A} - \mathbf{A}_k\|_F^2 + (\epsilon + \eta)\|\mathbf{A}\|_F^2. \tag{7}$$

This additive-error guarantee was introduced by the seminal work of Frieze et. al. [FKV04], and triggered a long line of work on low-rank approximation from a computational perspective. Frieze et al. showed that it is possible to achieve (7) in $O(\text{nnz}(\mathbf{A}))$ time. Further, given access to an oracle for computing row norms of $\mathbf{A}$, 7 is achievable in sublinear time. More recently, the same notion of approximation was used to obtain sublinear sample complexity and running time algorithms for *distance matrices* [BW18],[IVWW19], and a quantum algorithm for recommendation systems [KP16], which was subsequently dequantized [Tan19].

This raises the question of how robust are our sublinear low-rank approximation algorithms for structured matrices, if we relax to additive-error guarantees and allow for corruption. In particular, can we obtain additive-error low-rank approximation algorithms for PSD matrices that achieve sublinear time and sample complexity in the presence of noise? We characterize when such robust algorithms are achievable in sublinear time.

Our Results. We begin with stating our results for low-rank approximation for structured matrices. Our main result is an optimal algorithm for low-rank approximation of PSD matrices:

Problem	Prior Work		Our Results		Query Lower Bound
	Query	Run Time	Query	Run Time	
PSD LRA	$O\left(\frac{nk}{\epsilon^{2.5}}\right)$	$O\left(\frac{nk^{\omega-1}}{\epsilon^{2\omega-2}}+\frac{nk}{\epsilon^{2.5}}\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^\dagger\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	[MW17]		Thm. 3.1		[MW17]
PSD LRA PSD Output	$O\left(nk\left(\frac{k}{\epsilon^2}+\frac{1}{\epsilon^3}\right)\right)$	$O\left(nk^{\omega-1}\left(\frac{k}{\epsilon^\omega}+\frac{1}{\epsilon^{3\omega-3}}\right)\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^\dagger\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	[MW17]		Thm. 3.1		[MW17]
Negative-Type LRA	$O\left(\frac{nk}{\epsilon^{2.5}}\right)$	$O\left(\frac{nk^{\omega-1}}{\epsilon^{2\omega-2}}+\frac{nk}{\epsilon^{2.5}}\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^\dagger\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	Bi-criteria, [BW18]		No Bi-criteria, Thm. 3.4		[BW18]
Coreset Ridge Regression	$O\left(\frac{ns_\lambda^2}{\epsilon^4}\right)$	$O\left(\frac{ns_\lambda^\omega}{\epsilon^\omega}\right)$	$O^*\left(\frac{ns_\lambda}{\epsilon^2}\right)$	$O^\dagger\left(\frac{ns_\lambda^{\omega-1}}{\epsilon^{2\omega-2}}\right)$	$\Omega\left(\frac{ns_\lambda}{\epsilon^2}\right)$
	[MW17]		Thm. 3.6		

Table 2: Comparison with prior work. The notation O^* and $O^\dagger$ represent existence of matching lower bounds for query complexity and running time (assuming the fast matrix multiplication exponent ω is 2) respectively. The notation s_λ is used to denote the statistical dimension of ridge regression. All bounds are stated ignoring polylogarithmic factors in n, k and ϵ.

Theorem 3.1 (Sample-Optimal PSD LRA)**.** *Given a PSD matrix* $\mathbf{A}$*, there exists an algorithm that queries* $\widetilde{O}(nk/\epsilon)$ *entries in* $\mathbf{A}$ *and outputs a rank* k *matrix* $\mathbf{B}$ *such that with probability* $99/100$*,* $\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_F^2$*, and the algorithm runs in time* $\widetilde{O}(n\cdot(k/\epsilon)^{\omega-1})$*.*

Remark 3.2. Our algorithm matches the sample complexity lower bound of Musco and Woodruff, up to logarithmic factors, which shows that any randomized algorithm that outputs a $(1+\epsilon)$-relative-error low-rank approximation for a PSD matrix $\mathbf{A}$ must read $\Omega(nk/\epsilon)$ entries. Our running time also improves that of Musco and Woodruff and is optimal if the matrix multiplication exponent ω is 2.

Remark 3.3. We can extend our algorithm such that the low-rank matrix $\mathbf{B}$ we output is also PSD with the same query complexity and running time. In comparison, the algorithm of Musco and Woodruff accesses $\widetilde{O}(nk/\epsilon^3+nk^2/\epsilon^2)$ entries in $\mathbf{A}$ and runs in time $\widetilde{O}(n(k/\epsilon)^\omega+nk^{\omega-1}/\epsilon^{3(\omega-1)})$.

At the core of our analysis is a sample optimal algorithm for Spectral Regression: $\min_{\mathbf{X}}\|\mathbf{DX}-\mathbf{E}\|_2^2$. We show that when $\mathbf{D}$ has orthonormal columns and $\mathbf{E}$ is arbitrary, we can sketch the problem by sampling rows proportional to the leverage scores of $\mathbf{D}$ and approximately preserve the minimum cost. This is particularly surprising since our sketch only computes sampling probabilities by reading entries in $\mathbf{D}$, while being completely agnostic to the entries in $\mathbf{E}$. Here, we also prove a spectral approximate matrix product guarantee for our one-sided leverage score sketch, which may be of independent interest. We note that such a guarantee for leverage score sampling does not appear in prior work, and we discuss the technical challenges we need to overcome in the subsequent section.

The techniques we develop for PSD low-rank approximation also extend to computing a low-rank approximation for distance matrices that arise from negative-type (Euclidean-squared) metrics. Here, our input is a pair-wise distance matrix $\mathbf{A}$ corresponding to a point set $\mathcal{P}=\{x_1,x_2,\ldots,x_n\}\in\mathbb{R}^d$ such that $\mathbf{A}_{i,j}=\|x_i-x_j\|_2^2$. We obtain an optimal algorithm for computing a low-rank approximation of such matrices:

Theorem 3.4 (Sample-Optimal LRA for Negative-Type Metrics)**.** *Given a negative-type distance matrix* $\mathbf{A}$*, there exists an algorithm that queries* $\widetilde{O}(nk/\epsilon)$ *entries in* $\mathbf{A}$ *and outputs a rank* k *matrix* $\mathbf{B}$ *such that with probability* $99/100$*,* $\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_F^2$*, and the algorithm runs in time* $\widetilde{O}(n\cdot(k/\epsilon)^{\omega-1})$*.*

Remark 3.5. In prior work with David Woodruff [BW18], we obtained a $\widetilde{O}(nk/\epsilon^{2.5})$ query algorithm that outputs a rank-$(k+4)$ matrix $\mathbf{B}$ such that $\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_F^2$. We show that the bi-criteria guarantee is not necessary, thereby resolving an open question in their paper.

Structured Regression. The sample-optimal algorithm for PSD Low-Rank Approximation also leads to a faster algorithm for Ridge Regression, when the design matrix is PSD. Given a PSD matrix $\mathbf{A}$, a vector y and a regularization parameter λ, we consider the following optimization problem: $\min_{x\in\mathbb{R}^n}\|\mathbf{A}x-y\|_2^2+\lambda\|x\|_2^2$. This problem is often referred to as Ridge Regression and has been the focus of numerous theoretical and practical works.

Theorem 3.6 (PSD Ridge Regression.)**.** *Given a PSD matrix* $\mathbf{A}$*, a regularization parameter* λ *and statistical dimension* $s_\lambda = Tr(\mathbf{A}^2+\lambda\mathbf{I})^{-1}\mathbf{A}^2$*, there exists an algorithm that queries* $\widetilde{O}(ns_\lambda/\epsilon^2)$ *entries of* $\mathbf{A}$ *and with probability* $99/100$ *outputs a* $(1+\epsilon)$ *approximate solution to the Ridge Regression objective and runs in* $\widetilde{O}(n(s_\lambda/\epsilon^2)^{\omega-1})$ *time.*

Remark 3.7. Our result improves on prior work by Musco and Woodruff [MW17], who obtain an algorithm that queries $\widetilde{O}(ns_\lambda^2/\epsilon^4)$ entries in $\mathbf{A}$ and runs in $\widetilde{O}(n(s_\lambda/\epsilon^2)^\omega)$ time.

Robust Low-Rank Approximation. Next, we consider a robust form of low-rank approximation problem, where the input is a PSD matrix corrupted by noise. In this setting, we have query access to the corrupted matrix $\mathbf{A}+\mathbf{N}$, where $\mathbf{A}$ is PSD and $\mathbf{N}$ is such that $\|\mathbf{N}\|_F^2 \leqslant \eta\|\mathbf{A}\|_F^2$. Further, for all $i\in[n]$ $\|\mathbf{N}_{i,*}\|_2^2 \leqslant c\|\mathbf{A}_{i,*}\|_2^2$, for a fixed constant c. The diagonal of a PSD matrix carries crucial information since the largest diagonal entry upper bounds all off-diagonal entries. Therefore, a reasonable adversarial strategy is to corrupt the largest diagonal entries and make them close to the small diagonal entries, which enables the resulting matrix to have large off-diagonal entries that are hard to find. Capturing this intuition we parameterize our algorithms and lower bounds by the largest ratio between a diagonal entry of $\mathbf{A}$ and $\mathbf{A}+\mathbf{N}$, denoted by $\phi_{\max} = \max_{j\in[n]}\mathbf{A}_{j,j}/|(\mathbf{A}+\mathbf{N})_{j,j}|$.

Theorem 3.8 (Robust LRA Lower Bound)**.** *Let* $\epsilon > \eta > 0$*. Given* $\mathbf{A}+\mathbf{N}$ *such that* $\mathbf{A}$ *is PSD and* $\mathbf{N}$ *is a corruption matrix as defined above, any randomized algorithm that with probability at least* $2/3$ *outputs a rank-*k *approximation up to additive error* $(\epsilon+\eta)\|\mathbf{A}\|_F^2$ *must read* $\Omega\left(\phi_{\max}^2 nk/\epsilon\right)$ *entries of* $\mathbf{A}+\mathbf{N}$*.*

Remark 3.9. Any algorithm must incur additive error $\eta\|\mathbf{A}\|_F^2$, since $\mathbf{A}$ is not even identifiable below additive-error $\eta\|\mathbf{A}\|_F^2$.

Remark 3.10. In our hard instance, $\phi_{\max}^2$ can be as large as $\epsilon n/k$, which implies a sample-complexity lower bound of $\Omega(n^2)$. While this lower bound precludes sublinear algorithms for arbitrary PSD matrices, we observe that in many applications $\phi_{\max}$ can be significantly smaller. For instance, if $\mathbf{A}$ is a correlation matrix, we know that the true diagonal entries of $\mathbf{A}+\mathbf{N}$ are 1 and can ignore any corruption on them to bound $\phi_{\max}$ by 1.

Motivated by the aforementioned observation, we introduce algorithms for robust low-rank approximation, parameterized by the corruption on the diagonal entries. We obtain the following theorem:

Theorem 3.11 (Robust Low-Rank Approximation)**.** *Given* $\mathbf{A}+\mathbf{N}$*, which satisfies our noise model, there exists an algorithm that queries* $\widetilde{O}(\phi_{\max}^2 nk/\epsilon)$ *entries in* $\mathbf{A}+\mathbf{N}$ *and computes a rank* k *matrix* $\mathbf{B}$ *such that with probability at least* $99/100$*,* $\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant \|\mathbf{A}-\mathbf{A}_k\|_F^2 + (\epsilon+\sqrt{\eta})\|\mathbf{A}\|_F^2$.

Remark 3.12. While the sample complexity of this algorithm matches the sample complexity in the lower bound, it incurs additive-error $\sqrt{\eta}\|\mathbf{A}\|_F^2$ as opposed to $\eta\|\mathbf{A}\|_F^2$. An interesting open question here is whether we can achieve additive-error $o(\sqrt{\eta}\|\mathbf{A}\|_F^2)$, though we note that when $\eta^2 \leqslant \epsilon$, this just changes the additive error guarantee of our low-rank approximation by a constant factor.

Remark 3.13. Our techniques extend to low-rank approximation of correlation matrices, and we obtain a sample complexity of $\widetilde{O}(nk/\epsilon)$, which is optimal. In fact, the hard instance in [MW17] implies an $\Omega(nk/\epsilon)$ lower bound on the sample complexity, even in the presence of no noise. Surprisingly, corrupting a correlation matrix does not increase the sample complexity and only incurs an additive error of $\sqrt{\eta}\|\mathbf{A}\|_F^2$.

Future Directions. A nascent area in algorithm design is developing fast algorithms for structured linear algebra problems. This area has seen rapid progress for problems including low-rank approximation (see above), regression and covariance estimation. Considering structured matrices can also be an avenue for progress on major open problems like spectral low-rank approximation. An open ended research direction is as follows:

Open Question 3.14. When does structure in the input lead to faster algorithms for fundamental problems in numerical linear algebra? How robust are the corresponding algorithms to perturbations of the structure in the input?

As mentioned above, exploiting structure of the input matrices has lead to several algorithmic breakthroughts: solving linear systems for Laplacian/Diagonally Dominant matrices [ST14, KOSZ13, KMP14] and Block Henkel matrices [PV21], covariance estimation of Toeplitz matrices [ELMM20], and approximation the permanent of boolean [JS89], non-negative matrices [JSV04] and PSD [AGGS17, YP21] matrices. Obtaining provable guarantees for the aforementioned tasks, even when the input matrix is perturbed by noise, is an intruiging research direction.

A closely related model to study algorihms for numerical linear algebra is the matrix-vector query model. It captures a large family of iterative algorithms, such as Krylov methods, and typically the number of matrix-vector queries are the main bottleneck in real-world problems. Therefore, a natural question to consider is as follows:

Open Question 3.15. What is the matrix-vector product complexity of low-rank approximation for the Frobenius norm, and more generally, for other matrix norms?

Currently, we do not know any lower bounds for the matrix-vector complexity of low-rank approximation under any matrix norm. For Frobenius and Spectral low-rank approximation, Musco and Musco [MM15] provide an upper bound of $O\left(k\log(d/\epsilon)/\sqrt{\epsilon}\right)$ matrix-vector products.

More broadly, the tools we developed in these works have been useful for a myriad of machine learning applications, including provable guarantees for training two layer ReLU networks [BJW19], distriuted clustering [ABB+17] , learning a latent simplex in input sparsity time [BBK+21], and quantum-inspired algorithms for machine learning [CCH+20]. Looking forward, we hope to understand the power and applicability of these tools to learning other latent models as well as quantum-inspired algorithms.

3.2 PSD Testing

Testing whether a matrix is Positive Semi-Definite often provides crucial insights into the structure of metric spaces, arises as a separation oracles in Semi-Definite Programming (SDP), leads to faster algorithms for solving linear systems and linear algebra problems detects existence of community structure in random graphs, and is used to ascertain local convexity of functions. Furthermore, testing PSDness is also useful when studying the rate of dissipation in the heat equation and the behavior of non-oscillatory, exponentially stable modes of linear differential equations. For these applications, in addition to testing the existence of negative eigenvalues, it is often important to provide a *certificate* that the matrix is not PSD, by exhibiting a direction in which the quadratic form is negative.

While efficient, numerically stable algorithms for computing the spectrum of a matrix have been known since Turing [Tur48], such algorithms require reading the entire matrix and incur a cubic running time in practice. Computing the eigenvalues of a matrix is often the bottleneck in applications, especially when just determining the existence of negative eigenvalues suffices. For instance, checking embeddability of a finite metric into Euclidean space, feasibility of a SDP, convexity of a function, and if specialized solvers are applicable for linear algebraic problems, all only require knowledge of whether a given matrix is PSD. We initiate the study of when we can test PSD-ness without reading the entire matrix.

We approach the problem from the perspective of property testing, where the input matrix $\mathbf{A}$ is promised to be either a PSD matrix, or "ε-far" from PSD under an appropriate notion of distance (discussed below). Specifically, we work in the *bounded-entry model*, proposed by Balcan, Li, Woodruff, and Zhang [BLWZ19], where the input matrix has bounded entries: $\|\mathbf{A}\|_\infty \leqslant 1$. Boundedness is often a natural assumption in practice, and has numerous real world applications, such as recommender systems as in the Netflix Challenge [KBV09], unweighted or bounded weight graphs [Gol10, GGR98], correlation matrices, distance matrices with bounded radius, and others [LWW14, KIDP16, BLWZ19]. Further, the boundedness of entries avoids degenerate instances where an arbitrarily large entry is hidden in $\mathbf{A}$, thereby drastically changing the spectrum of $\mathbf{A}$, while being impossible to test without reading the entire matrix.

Our starting point is a simple fact: a matrix $\mathbf{A}$ is PSD if and only if all *principal*[5] submatrices of $\mathbf{A}$ are PSD. However, a much more interesting direction is: if $\mathbf{A}$ is not PSD, what can be said about the eigenvalues of the submatrices of $\mathbf{A}$? Specifically, if $\mathbf{A}$ is far from PSD, how large of a submatrix must one sample in order to find a negative eigenvalue? Note that given a principal submatrix $\mathbf{A}_{T\times T}$ with $x^\top \mathbf{A}_{T\times T} x < 0$ for some $x \in \mathcal{R}^{|T|}$, this direction x can be used as a certificate that the input matrix is not PSD, since $y^\top \mathbf{A} y = x^\top \mathbf{A}_{T\times T} x < 0$, where y is the result of padding x with 0's. Further, it leads us to a natural algorithm to test definiteness: sample multiple principal submatrices and compute their eigenvalues. If any are negative, then $\mathbf{A}$ must not be PSD. Determining the query complexity of this task is the principal focus of this paper. Specifically, we ask:

Can the positive semi-definiteness of a bounded matrix be tested via the semi-definiteness of a small random submatrix?

The Testing Models. The distance from $\mathbf{A}$ to the PSD cone is given by $\min_{\mathbf{B}\succeq 0} \|\mathbf{A}-\mathbf{B}\|$, where $\|\cdot\|$ is a norm, and $\mathbf{B}\succeq 0$ denotes that $\mathbf{B}$ is PSD. To instantiate $\|\cdot\|$, we consider two natural norms

[5]Recall that a principal submatrix $\mathbf{A}_{T\times T}$ for $T \subseteq [n]$ is the restriction of $\mathbf{A}$ to the rows and columns indexed by T.

over $n \times n$ matrices: the spectral norm ($\|\cdot\|_2$) and the Euclidean norm ($\|\cdot\|_F$). Perhaps surprisingly, the distance of a symmetric matrix $\mathbf{A}$ to the PSD cone under these norms can be characterized in terms of the eigenvalues of $\mathbf{A}$. In particular, let $\lambda \in \mathcal{R}^n$ be the vector of eigenvalues of $\mathbf{A}$. Then, the spectral norm distance corresponds to the ℓ_∞ distance between λ and the positive orthant. Similarly, the squared Frobenius distance corresponds to the ℓ_2^2 distance between λ and the positive orthant.

Therefore, we will refer to the two resulting gap problems as the ℓ_∞-gap and the ℓ_2^2-gap, respectively. This connection between matrix norms of $\mathbf{A}$ and vector norms of eigenvalues λ will be highly useful for the analysis of random submatrices. Next, we formally define the testing problems:

Problem 3.16 (PSD Testing with Spectral norm/ℓ_∞-gap)**.** Given $\varepsilon \in (0, 1]$ and a symmetric matrix $\mathbf{A} \in \mathcal{R}^{n \times n}$ such that $\|\mathbf{A}\|_\infty \leqslant 1$, distinguish whether $\mathbf{A}$ satisfies:

(1) $\mathbf{A}$ is PSD.

(2) $\mathbf{A}$ is ε-far from the PSD cone in Spectral norm: $\min_{\mathbf{B} \succeq 0} \|\mathbf{A} - \mathbf{B}\|_2 = \max_{i:\lambda_i < 0} |\lambda_i(\mathbf{A})| \geqslant \varepsilon n$.

The fact that the spectral norm distance from $\mathbf{A}$ to the PSD cone ($\min_{\mathbf{B} \succeq 0} \|\mathbf{A} - \mathbf{B}\|_2$) is equivalent to the magnitude of the smallest negative eigenvalue of $\mathbf{A}$ is a consequence of the variational principle for eigenvalues. For general non-symmetric matrices $\mathbf{A}$, one can replace (2) above with the condition $x^\top \mathbf{A} x < -\varepsilon n$ for some unit vector $x \in \mathcal{R}^n$, which is equivalent to (2) if $\mathbf{A}$ is symmetric (again by the variational principle). We note that our results for the ℓ_∞-gap hold in this more general setting.[6]

Next, if we instantiate $\|\cdot\|$ with the (squared) Euclidean norm, we obtain the ℓ_2^2 gap problem.

Problem 3.17 (PSD Testing with ℓ_2^2-gap)**.** Given $\varepsilon \in (0, 1]$ and a symmetric matrix $\mathbf{A} \in \mathcal{R}^{n \times n}$ such that $\|\mathbf{A}\|_\infty \leqslant 1$, distinguish whether $\mathbf{A}$ satisfies:

(1) $\mathbf{A}$ is PSD.

(2) $\mathbf{A}$ is ε-far from the PSD cone in squared Euclidean norm:

$$\min_{\mathbf{B} \succeq 0} \|\mathbf{A} - \mathbf{B}\|_F^2 = \sum_{i:\lambda_i(\mathbf{A}) < 0} \lambda_i^2(\mathbf{A}) \geqslant \varepsilon n^2 \tag{8}$$

We first state our result for the ℓ_∞ gap problem in its most general form, which is equivalent to Problem 3.16 in the special case when $\mathbf{A}$ is symmetric.

Theorem 3.18 (ℓ_∞-gap Upper Bound)**.** *There is a non-adaptive sampling algorithm which, given $\mathbf{A} \in \mathcal{R}^{n \times n}$ with $\|\mathbf{A}\|_\infty \leqslant 1$ and $\varepsilon \in (0, 1)$, returns PSD if $x^\top \mathbf{A} x \geqslant 0$ for all $x \in \mathcal{R}^n$, and with probability $2/3$ returns Not PSD if $x^\top \mathbf{A} x \leqslant -\varepsilon n$ for some unit vector $x \in \mathcal{R}^n$. The algorithm make $\tilde{O}(1/\varepsilon^2)$ queries to the entries of $\mathbf{A}$, and runs in time $\tilde{O}(1/\varepsilon^\omega)$.*

We demonstrate that the algorithm of Theorem 3.18 is optimal up to $\log(1/\varepsilon)$ factors, even for adaptive algorithms with two-sided error. Formally, we show:

[6] Also note that given query access to any $\mathbf{A} \in \mathcal{R}^{n \times n}$, one can always run a tester on the symmetrization $\mathbf{B} = (\mathbf{A} + \mathbf{A}^\top)/2$, which satisfies $x^\top \mathbf{A} x = x^\top \mathbf{B} x$ for all x, with at most a factor of 2 increase in query complexity.

Theorem 3.19 (ℓ_∞-gap Lower Bound)**.** *Any adaptive or non-adaptive algorithm which solves the PSD testing problem with ε-ℓ_∞ gap with probability at least $2/3$, even with two-sided error and if $\mathbf{A}$ is promised to be symmetric, must query $\widetilde{\Omega}(1/\varepsilon^2)$ entries of $\mathbf{A}$.*

Next, we present our algorithm for the ℓ_2^2-gap problem. Our algorithm crucially relies on first running our tester for the ℓ_∞-gap problem, which allows us to demonstrate that if $\mathbf{A}$ is far from PSD in ℓ_2^2 but close in ℓ_∞, then it must be far, under other notions of distance such as Schatten norms or residual tail error, from any PSD matrix.

Theorem 3.20 (ℓ_2^2-gap Upper Bound)**.** *here is a non-adaptive sampling algorithm which, given a symmetric matrix $\mathbf{A} \in \mathcal{R}^{n\times n}$ with $\|\mathbf{A}\|_\infty \leqslant 1$ and $\varepsilon \in (0,1)$, returns PSD if $\mathbf{A}$ is PSD, and with probability $2/3$ returns Not PSD if $\min_{\mathbf{B}\succeq 0} \|\mathbf{A}-\mathbf{B}\|_F^2 \geqslant \varepsilon n^2$. The algorithm make $\tilde{O}(1/\varepsilon^4)$ queries to $\mathbf{A}$, and runs in time $\tilde{O}(1/\varepsilon^{2\omega})$.*

We complement our upper bound by a $\widetilde{\Omega}(\frac{1}{\varepsilon^2})$ lower bound for PSD-testing with ε-ℓ_2^2 gap, which holds even for algorithms with two sided error. Our lower bound demonstrates a separation between the complexity of PSD testing with $\sqrt{\varepsilon}$-ℓ_∞ gap and PSD testing with ε-ℓ_2^2-gap, showing that the concentration of negative mass in large eigenvalues makes PSD testing a strictly easier problem.

Theorem 3.21 (ℓ_2^2-gap Lower Bound)**.** *Any non-adaptive algorithm which solves the PSD testing problem with ε-ℓ_2^2 gap with probability at least $2/3$, even with two-sided error, must query $\widetilde{\Omega}(1/\varepsilon^2)$ entries of $\mathbf{A}$.*

Our lower bound is built on discrete hard instances which are "locally indistinguishable", in the sense that the distribution of any small set of samples is completely identical between the PSD and ε-far cases. At the heart of the lower bound is a key combinatorial Lemma about arrangements of paths on cycle graphs. Our construction is highly general, and we believe will likely be useful for proving other lower bounds for matrix and graph property testing problems.

References

[ABB+17] Pranjal Awasthi, Ainesh Bakshi, Maria-Florina Balcan, Colin White, and David Woodruff. Robust communication-optimal distributed clustering algorithms. arXiv preprint arXiv:1703.00830, 2017. 24

[AFKM01] Dimitris Achlioptas, Amos Fiat, Anna R Karlin, and Frank McSherry. Web search via hub synthesis. In Proceedings 42nd IEEE Symposium on Foundations of Computer Science, pages 500–509. IEEE, 2001. 20

[AGGS17] Nima Anari, Leonid Gurvits, Shayan Oveis Gharan, and Amin Saberi. Simply exponential approximation of the permanent of positive semidefinite matrices. In 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), pages 914–925. IEEE, 2017. 3, 24

[AGM12] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In 2012 IEEE 53rd annual symposium on foundations of computer science, pages 1–10. IEEE, 2012. 1

[AK05] Sanjeev Arora and Ravi Kannan. Learning mixtures of separated nonspherical gaussians. The Annals of Applied Probability, 15(1A):69–92, 2005. 4

[ALN08] Sanjeev Arora, James Lee, and Assaf Naor. Euclidean distortion and the sparsest cut. Journal of the American Mathematical Society, 21(1):1–21, 2008. 20

[AM05] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In International Conference on Computational Learning Theory, pages 458–469. Springer, 2005. 4, 20

[Bar] Boaz Barak. Proofs, beliefs, and algorithms through the lens of sum-of-squares. 36

[BBK+21] Ainesh Bakshi, Chiranjib Bhattacharyya, Ravi Kannan, David Woodruff, and Samson Zhou. Learning a latent simplex in input sparsity time. In International Conference on Learning Representations, 2021. 24

[BBV08] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In Proceedings of the fortieth annual ACM symposium on Theory of computing, pages 671–680, 2008. 2

[BCJ20] Ainesh Bakshi, Nadiia Chepurko, and Rajesh Jayaram. Testing positive semi-definiteness via random submatrices. In 2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS), pages 1191–1202. IEEE, 2020. 3

[BCMV14] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In Proceedings of the forty-sixth annual ACM symposium on Theory of computing, pages 594–603, 2014. 1

[BCW20] Ainesh Bakshi, Nadiia Chepurko, and David P Woodruff. Robust and sample optimal algorithms for psd low rank approximation. In 2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS), pages 506–516. IEEE, 2020. 2

[BDH+20] Ainesh Bakshi, Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, Sushrut Karmalkar, and Pravesh K Kothari. Outlier-robust clustering of gaussians and other non-spherical mixtures. In 2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS), pages 149–159. IEEE, 2020. 6, 9

[BDJ+20] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M Kane, Pravesh K Kothari, and Santosh S Vempala. Robustly learning mixtures of k arbitrary gaussians. arXiv preprint arXiv:2012.02119, 2020. 2, 9

[BHK20] Avrim Blum, John Hopcroft, and Ravindran Kannan. Foundations of data science. Cambridge University Press, 2020. 2

[BJW19] Ainesh Bakshi, Rajesh Jayaram, and David P Woodruff. Learning two layer rectified neural networks in polynomial time. In Conference on Learning Theory, pages 195–268. PMLR, 2019. 24

[BK20] Ainesh Bakshi and Pravesh Kothari. Outlier-robust clustering of non-spherical mixtures. arXiv preprint arXiv:2005.02970, 2020. 1, 5, 9, 10

[BK21] Ainesh Bakshi and Pravesh K Kothari. List-decodable subspace recovery: Dimension independent error in polynomial time. In Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 1279–1297. SIAM, 2021. 2, 7, 8, 19

[BKS15] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In Proceedings of the forty-seventh annual ACM symposium on Theory of computing, pages 143–151, 2015. 12

[BLWZ19] Maria-Florina Balcan, Yi Li, David P Woodruff, and Hongyang Zhang. Testing matrix rank, optimally. In Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 727–746. SIAM, 2019. 25

[BP21] Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC), pages 102–115, 2021. 2

[BS15] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. SIAM Journal on Computing, 44(4):889–911, 2015. 4

[BV08] S Charles Brubaker and Santosh S Vempala. Isotropic pca and affine-invariant clustering. In Building Bridges, pages 241–281. Springer, 2008. 4

[BW18] Ainesh Bakshi and David Woodruff. Sublinear time low-rank approximation of distance matrices. In Advances in Neural Information Processing Systems, pages 3782–3792, 2018. 2, 20, 21, 22, 23

[CAT+20] Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I Jordan, Nicolas Flammarion, and Peter L Bartlett. Optimal robust linear regression in nearly linear time. arXiv preprint arXiv:2007.08137, 2020. 2

[CCH+20] Nadiia Chepurko, Kenneth L Clarkson, Lior Horesh, Honghao Lin, and David P Woodruff. Quantum-inspired algorithms from randomized numerical linear algebra. arXiv preprint arXiv:2011.04125, 2020. 24

[CKM+11] Paul Christiano, Jonathan A Kelner, Aleksander Madry, Daniel A Spielman, and Shang-Hua Teng. Electrical flows, laplacian systems, and faster approximation of maximum flow in undirected graphs. In Proceedings of the forty-third annual ACM symposium on Theory of computing, pages 273–282. ACM, 2011. 20

[CLW18] Nai-Hui Chia, Han-Hsuan Lin, and Chunhao Wang. Quantum-inspired sublinear classical algorithms for solving low-rank linear systems. arXiv preprint arXiv:1811.04852, 2018. 20

[CRR+96] Ashok K Chandra, Prabhakar Raghavan, Walter L Ruzzo, Roman Smolensky, and Prasoon Tiwari. The electrical resistance of a graph captures its commute and cover times. Computational Complexity, 6(4):312–340, 1996. 20

[Das99] Sanjoy Dasgupta. Learning mixtures of gaussians. In 40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039), pages 634–644. IEEE, 1999. 4

[DFK+04] Petros Drineas, Alan Frieze, Ravi Kannan, Santosh Vempala, and V Vinay. Clustering large graphs via the singular value decomposition. Machine learning, 56(1-3):9–33, 2004. 20

[DHKK20] Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, and Sushrut Karmalkar. Robustly learning any clusterable mixture of gaussians. arXiv preprint arXiv:2005.06417, 2020. 1, 6

[DK19] Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. arXiv preprint arXiv:1911.05911, 2019. 1, 4

[DKK+18] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. arXiv preprint arXiv:1803.02815, 2018. 16

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. SIAM Journal on Computing, 48(2):742–864, 2019. 1, 4, 5

[DKR02] Petros Drineas, Iordanis Kerenidis, and Prabhakar Raghavan. Competitive recommendation systems. In Proceedings of the thiry-fourth annual ACM symposium on Theory of computing, pages 82–90. ACM, 2002. 20

[DKS17] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), pages 73–84. IEEE, 2017. 9

[DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, pages 1047–1060, 2018. 1, 9, 18

[DKS19] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 2745–2754. SIAM, 2019. 2, 13

[DL09] Michel Marie Deza and Monique Laurent. Geometry of cuts and metrics, volume 15. Springer, 2009. 20

[DVW18] Ilias Diakonikolas, Santosh Vempala, and David Woodruff. Research vignette: Foundations of data science. Simons Institute, Semester on Foundations of Big Data, 2018. 1

[ELMM20] Yonina C Eldar, Jerry Li, Cameron Musco, and Christopher Musco. Sample efficient toeplitz covariance estimation. In Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 378–397. SIAM, 2020. 3, 24

[FKV04] Alan M. Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. J. ACM, 51(6):1025–1041, 2004. 21

[GGR98] Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. Journal of the ACM (JACM), 45(4):653–750, 1998. 25

[GLF+10] David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. Physical review letters, 105(15):150401, 2010. 20

[GLS81] Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. Combinatorica, 1(2):169–197, 1981. 36

[GLT18] András Gilyén, Seth Lloyd, and Ewin Tang. Quantum-inspired low-rank stochastic regression with logarithmic dependence on the dimension. arXiv preprint arXiv:1811.04909, 2018. 20

[Gol10] Oded Goldreich. Introduction to testing graph properties. In Property testing, pages 105–141. Springer, 2010. 25

[GSLW19] András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics. In Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, pages 193–204. ACM, 2019. 20

[Hig02] Nicholas J Higham. Computing the nearest correlation matrix—a problem from finance. IMA journal of Numerical Analysis, 22(3):329–343, 2002. 21

[HL18] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, pages 1021–1034, 2018. 1, 9

[HRRS11] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. Robust statistics: the approach based on influence functions, volume 196. John Wiley & Sons, 2011. 4

[Hub64] Peter J Huber. Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35(1):73–101, 1964. 1, 4

[Hub04] Peter J Huber. Robust statistics, volume 523. John Wiley & Sons, 2004. 4, 13

[IVWW19] Piotr Indyk, Ali Vakilian, Tal Wagner, and David Woodruff. Sample-optimal low-rank approximation of distance matrices. arXiv preprint arXiv:1906.00339, 2019. 20, 21

[Jae72] Louis A Jaeckel. Estimating regression coefficients by minimizing the dispersion of the residuals. The Annals of Mathematical Statistics, pages 1449–1458, 1972. 13

[JLST21] Arun Jambulapati, Jerry Li, Tselil Schramm, and Kevin Tian. Robust regression revisited: Acceleration and improved estimation rates. arXiv preprint arXiv:2106.11938, 2021. 2

[JS89] Mark Jerrum and Alistair Sinclair. Approximating the permanent. SIAM journal on computing, 18(6):1149–1178, 1989. 3, 24

[JSV04] Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. Journal of the ACM (JACM), 51(4):671–697, 2004. 3, 24

[Kan20] Daniel M Kane. Robust learning of mixtures of gaussians. arXiv preprint arXiv:2007.05912, 2020. 2, 9, 11

[KBV09] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. Computer, 42(8):30–37, 2009. 25

[KIDP16] Ramakrishnan Kannan, Mariya Ishteva, Barry Drake, and Haesun Park. Bounded matrix low rank approximation. In Non-negative Matrix Factorization Techniques, pages 89–118. Springer, 2016. 25

[KKK19] Sushrut Karmalkar, Adam Klivans, and Pravesh Kothari. List-decodable linear regression. In Advances in Neural Information Processing Systems, pages 7423–7432, 2019. 2, 7, 8, 19, 36

[KKM18] Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. arXiv preprint arXiv:1803.03241, 2018. 2, 13, 15, 16, 18, 36

[Kle99] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5):604–632, 1999. 20

[KMP14] Ioannis Koutis, Gary L Miller, and Richard Peng. Approaching optimality for solving sdd linear systems. SIAM Journal on Computing, 43(1):337–354, 2014. 3, 20, 24

[KOSZ13] Jonathan A Kelner, Lorenzo Orecchia, Aaron Sidford, and Zeyuan Allen Zhu. A simple, combinatorial algorithm for solving sdd systems in nearly-linear time. In Proceedings of the forty-fifth annual ACM symposium on Theory of computing, pages 911–920, 2013. 3, 24

[KOTZ14] Manuel Kauers, Ryan O'Donnell, Li-Yang Tan, and Yuan Zhou. Hypercontractive inequalities via sos, and the frankl–rödl graph. In Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1644–1658. SIAM, 2014. 7

[KP16] Iordanis Kerenidis and Anupam Prakash. Quantum recommendation systems. arXiv preprint arXiv:1603.08675, 2016. 20, 21

[KSS18] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, pages 1035–1046, 2018. 1, 7, 9, 14, 36

[KSV05] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In International Conference on Computational Learning Theory, pages 444–457. Springer, 2005. 20

[KV17] Ravindran Kannan and Santosh Vempala. Randomized algorithms in numerical linear algebra. Acta Numerica, 26:95–135, 2017. 2

[Las01] Jean B Lasserre. New positive semidefinite relaxations for nonconvex quadratic programs. In Advances in Convex Analysis and Global Optimization, pages 319–331. Springer, 2001. 36

[LM21] Allen Liu and Ankur Moitra. Settling the robust learnability of mixtures of gaussians. In Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, pages 518–531, 2021. 9

[LRV16] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), pages 665–674. IEEE, 2016. 1, 4

[LSS$^+$] Erik M Lindgren, Vatsal Shah, Yanyao Shen, Alexandros G Dimakis, and Adam Klivans. On robust learning of ising models. 19

[LWW14] Yi Li, Zhengyu Wang, and David P Woodruff. Improved testing of low rank matrices. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 691–700, 2014. 25

[McS01] Frank McSherry. Spectral partitioning of random graphs. In Proceedings 42nd IEEE Symposium on Foundations of Computer Science, pages 529–537. IEEE, 2001. 20

[MM15] Cameron Musco and Christopher Musco. Randomized block Krylov methods for stronger and faster approximate singular value decomposition. In Advances in Neural Information Processing Systems, pages 1396–1404, 2015. 24

[MSS16] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), pages 438–446. IEEE, 2016. 36

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In 2010 IEEE 51st Annual Symposium on Foundations of Computer Science, pages 93–102. IEEE, 2010. 1, 4, 9, 10

[MW17] Cameron Musco and David P. Woodruff. Sublinear time low-rank approximation of positive semidefinite matrices. In 58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017, pages 672–683, 2017. 2, 20, 22, 23, 24

[Nes00] Yurii Nesterov. Squared functional systems and optimization problems. In High performance optimization, pages 405–440. Springer, 2000. 36

[O'D14] Ryan O'Donnell. Analysis of boolean functions. Cambridge University Press, 2014. 7

[Par00] Pablo A Parrilo. Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization. PhD thesis, California Institute of Technology, 2000. 36

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London. A, 185:71–110, 1894. 4

[PJL20] Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. arXiv preprint arXiv:2009.12976, 2020. 2

[PSBR20] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(3):601–627, 2020. 2, 13, 15, 16, 18

[PV21] Richard Peng and Santosh Vempala. Solving sparse linear systems faster than matrix multiplication. In Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 504–521. SIAM, 2021. 3, 24

[Rou84] Peter J Rousseeuw. Least median of squares regression. Journal of the American statistical association, 79(388):871–880, 1984. 13

[RSML18] Patrick Rebentrost, Adrian Steffens, Iman Marvian, and Seth Lloyd. Quantum singular-value decomposition of nonsparse low-rank matrices. Physical review A, 97(1):012327, 2018. 20

[RSS18] Prasad Raghavendra, Tselil Schramm, and David Steurer. High-dimensional estimation via sum-of-squares proofs. arXiv preprint arXiv:1807.11419, 6, 2018. 1

[RY84] Peter Rousseeuw and Victor Yohai. Robust regression by means of s-estimators. In Robust and nonlinear time series analysis, pages 256–272. Springer, 1984. 13

[RY20a] Prasad Raghavendra and Morris Yau. List decodable learning via sum of squares. In Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 161–180. SIAM, 2020. 2, 7, 19

[RY20b] Prasad Raghavendra and Morris Yau. List decodable subspace recovery. In Conference on Learning Theory, pages 3206–3226. PMLR, 2020. 7, 19

[Sen68] Pranab Kumar Sen. Estimates of the regression coefficient based on kendall's tau. Journal of the American statistical association, 63(324):1379–1389, 1968. 13

[Sho87] Naum Z Shor. Quadratic optimization problems. Soviet Journal of Computer and Systems Sciences, 25:1–11, 1987. 36

[SS11] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. SIAM Journal on Computing, 40(6):1913–1926, 2011. 20

[ST14] Daniel A Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. SIAM Journal on Matrix Analysis and Applications, 35(3):835–885, 2014. 3, 24

[SW19] Xiaofei Shi and David P. Woodruff. Sublinear time numerical linear algebra for structured matrices. In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019., pages 4918–4925, 2019. 20

[Tan19] Ewin Tang. A quantum-inspired classical algorithm for recommendation systems. In Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, pages 217–228. ACM, 2019. 20, 21

[TD87] Paul Terwilliger and Michel Deza. The classification of finite connected hypermetric spaces. Graphs and Combinatorics, 3(1):293–298, 1987. 20

[The92] Henri Theil. A rank-invariant method of linear and polynomial regression analysis. In Henri Theil's contributions to economics and econometrics, pages 345–381. Springer, 1992. 13

[Tur48] Alan M Turing. Rounding-off errors in matrix processes. The Quarterly Journal of Mechanics and Applied Mathematics, 1(1):287–308, 1948. 25

[VW04] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. Journal of Computer and System Sciences, 68(4):841–860, 2004. 4

[Wei05] Sanford Weisberg. Applied linear regression, volume 528. John Wiley & Sons, 2005. 13

[Woo14] David P. Woodruff. Sketching as a tool for numerical linear algebra. Foundations and Trends in Theoretical Computer Science, 10(1-2):1–157, 2014. 2, 20

[YP21] Chenyang Yuan and Pablo A Parrilo. Maximizing products of linear forms, and the permanent of positive semidefinite matrices. Mathematical Programming, pages 1–12, 2021. 3, 24

[ZJS19] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. arXiv preprint arXiv:1909.08755, 2019. 18

[ZJS20] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Robust estimation via generalized quasi-gradients. arXiv preprint arXiv:2005.14073, 2020. 2, 15, 16

A The Sum-of-Squares Proof System

In this section, we provide the necessary background for the sum-of-squares proof system. We follow the exposition as it appears in lecture notes by Barak [Bar], the Appendix of Ma, Shi and Steurer [MSS16], and the preliminary sections of several recent works [KSS18, KKK19, KKM18].

Pseudo-Distributions. We can represent a discrete probability distribution over $\mathcal{R}^n$ by its probability mass function $D\colon \mathcal{R}^n \to \mathcal{R}$ such that $D \geqslant 0$ and $\sum_{x\in\mathrm{supp}(D)} D(x) = 1$. Similarly, we can describe a pseudo-distribution by its mass function by relaxing the non-negativity constraint, while still passing certain low-degree non-negativity tests.

Definition A.1 (Pseudo-distribution)**.** A level-ℓ pseudo-distribution is a finitely-supported function $D : \mathcal{R}^n \to \mathcal{R}$ such that $\sum_x D(x) = 1$ and $\sum_x D(x)f(x)^2 \geqslant 0$ for every polynomial f of degree at most $\ell/2$, where the summation is over all x in the support of D.

Next, we define the notion of pseudo-expectation.

Definition A.2 (Pseudo-expectation)**.** The pseudo-expectation of a function f on $\mathcal{R}^d$ with respect to a pseudo-distribution D, denoted by $\tilde{\mathbb{E}}_{D(x)}[f(x)]$, is defined as

$$\tilde{\mathbb{E}}_{D(x)}[f(x)] = \sum_x D(x)f(x) \tag{9}$$

We use the notation $\tilde{\mathbb{E}}_{D(x)}[(1, x_1, x_2, \ldots, x_n)^{\otimes \ell}]$ to denote the degree-ℓ moment tensor of the pseudo-distribution D. In particular, each entry in the moment tensor corresponds to the pseudo-expectation of a monomial of degree at most ℓ in x. Crucially, there's an efficient separation oracle for moment tensors of pseudo-distributions.

Fact A.3 ([Sho87, Par00, Nes00, Las01])**.** *For any $n, \ell \in \mathbb{N}$, the following set has a $n^{O(\ell)}$-time weak separation oracle (in the sense of [GLS81]):*

$$\left\{\tilde{\mathbb{E}}_{D(x)}(1, x_1, x_2, \ldots, x_n)^{\otimes \ell} \mid \textit{degree-}\ell\textit{ pseudo-distribution } D \textit{ over } \mathcal{R}^n\right\} \tag{10}$$

This fact, together with the equivalence of weak separation and optimization [GLS81] forms the basis of the sum-of-squares algorithm, as it allows us to efficiently approximately optimize over pseudo-distributions.

Given a system of polynomial constraints, denoted by $\mathcal{A}$, we say that it is *explicitly bounded* if it contains a constraint of the form $\{\|x\|^2 \leqslant M\}$. Then, the following fact follows from Fact A.3 and [GLS81]:

Fact A.4 (Efficient Optimization over Pseudo-distributions). *There exists an $(n+m)^{O(\ell)}$-time algorithm that, given any explicitly bounded and satisfiable system*[7] *$\mathcal{A}$ of m polynomial constraints in n variables, outputs a level-ℓ pseudo-distribution that satisfies $\mathcal{A}$ approximately.*

[7] Here, we assume that the bit complexity of the constraints in $\mathcal{A}$ is $(n+m)^{O(1)}$.