List-decodeable Linear Regression

Anonymous Author(s)
Affiliation
Address
email

Abstract

We give the first polynomial-time algorithm for robust regression in the list-decodable setting where an adversary can corrupt a greater than $1/2$ fraction of examples. For any $\alpha < 1$, our algorithm takes as input a sample $\{(x_i, y_i)\}_{i \leqslant n}$ of n linear equations where αn of the equations satisfy $y_i = \langle x_i, \ell^* \rangle + \zeta$ for some small noise ζ and $(1-\alpha)n$ of the equations are *arbitrarily* chosen. It outputs a list L of size $O(1/\alpha)$ - a fixed constant - that contains an ℓ that is close to ℓ^*.

Our algorithm succeeds whenever the inliers are chosen from a *certifiably* anti-concentrated distribution D. To complement our result, we prove that the anti-concentration assumption on the inliers is information-theoretically necessary. As a corollary of our algorithmic result, we obtain a $(d/\alpha)^{O(1/\alpha^8)}$ time algorithm to find a $O(1/\alpha)$ size list when the inlier distribution is standard Gaussian. For discrete product distributions that are anti-concentrated only in *regular* directions, we give an algorithm that achieves similar guarantee under the promise that ℓ^* has all coordinates of same magnitude.

To solve the problem we introduce a new framework for list-decodable learning that strengthens the "identifiability to algorithms" paradigm based on the sum-of-squares method.

In an independent work, Raghavendra and Yau [RY19] have obtained a similar result for list-decodable regression also using the sum-of-squares method.

1 Introduction

In this work, we design algorithms for the problem of linear regression that are robust to training sets with an overwhelming ($\gg 1/2$) fraction of adversarially chosen outliers.

Outlier-robust learning algorithms have been extensively studied (under the name *robust statistics*) in mathematical statistics [Tuk75, MMY06, Hub11, HRRS11]. However, the algorithms resulting from this line of work usually run in time exponential in the dimension of the data [Ber06]. An influential line of recent work [KLS09, ABL13, DKK+16b, LRV16, CSV17, KS17a, KS17b, HL17, DKK+17, DKS17, KKM18] has focused on designing *efficient* algorithms for outlier-robust learning.

Our work extends this line of research. Our algorithms work in the *list-decodable learning* framework. In this model, the majority of the training data (a $1-\alpha$ fraction) can be

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

adversarially corrupted leaving only an $\alpha \ll 1/2$ fraction of *inliers*. Since uniquely recovering the underlying parameters is information-theoretically *impossible* in such a setting, the goal is to output a list (with an absolute constant size) of parameters, one of which matches the ground truth. This model was introduced in [BBV08] to give a discriminative framework for clustering. More recently, beginning with [CSV17], various works [DKS18, KS17a] have considered this as a model of *untrusted* data.

There has been a phenomenal progress in developing techniques for outlier-robust learning with a *small* ($\ll 1/2$)-fraction of outliers (e.g. outlier *filters* [DKK+16a, DKK+17], separation oracles for inliers [DKK+16a] or the *sum-of-squares* method [KS17b, HL17, KS17a, KKM18]). In contrast, progress on algorithms that tolerate the significantly harsher conditions in the list-decodable setting has been slower. The only prior works [CSV17, DKS18, KS17a] in this direction designed list-decodable algorithms for mean estimation via somewhat *ad hoc*, problem-specific methods.

In this paper, we develop a principled technique to give the first efficient list-decodable learning algorithm for the fundamental problem of *linear regression*. Our algorithm takes a corrupted set of linear equations with an $\alpha \ll 1/2$ fraction of inliers and outputs a $O(1/\alpha)$-size list of linear functions, one of which is guaranteed to be close to the ground truth (i.e., the linear function that correctly labels the inliers). Our key conceptual observation shows that list-decodable regression information-theoretically requires the inlier-distribution to be *anti-concentrated*. Our algorithm succeeds whenever the distribution satisfies a stronger *certifiable anti-concentration* condition. This class includes the standard gaussian distribution and more generally, any spherically symmetric distribution with strictly sub-exponential tails.

Prior to our work[1], the state-of-the-art outlier-robust algorithms for linear regression [KKM18, DKS19, DKK+18, PSBR18] could handle only a small (< 0.1)-fraction of outliers even under strong assumptions on the underlying distributions.

List-decodable regression generalizes the well-studied [DV89, JJ94, FS10, YCS13, BWY14, CYC14, ZJD16, SJA16, LL18] and *easier* problem of *mixed linear regression*: given k "clusters" of examples that are labeled by one out of k distinct unknown linear functions, find the unknown set of linear functions. All known techniques for the problem rely on faithfully estimating *moment tensors* from samples and thus, cannot tolerate the overwhelming fraction of outliers in the list-decodable setting. On the other hand, since we can take any cluster as inliers and treat rest as outliers, our algorithm immediately yields new efficient algorithms for mixed linear regression. Unlike all prior works, our algorithms work without any pairwise separation or bounded condition-number assumptions on the k linear functions.

List-Decodable Learning via the Sum-of-Squares Method Our algorithm relies on a strengthening of the robust-estimation framework based on the sum-of-squares (SoS) method. This paradigm has been recently used for clustering mixture models [HL17, KS17a] and obtaining algorithms for moment estimation [KS17b] and linear regression [KKM18] relies on a strengthening of robust-estimation framework based on the sum-of-squares (SoS) method. This paradigm has been recently used for clustering mixture models [HL17, KS17a] and obtaining algorithms for moment estimation [KS17b] and linear regression [KKM18] that are resilient to a small ($\ll 1/2$) fraction of outliers under the mildest known assumptions on the underlying distributions. Ths method reduces outlier-robust algorithm design to finding "simple" proofs of unique *identifiability* of the unknown parameter of the original distribution from a corrupted sample. However, this principled method works only in the setting with a small ($\ll 1/2$) fraction of outliers. As a consequence, the work of [KS17a] for

[1]There's a long line of work on robust regression algorithms (see for e.g. [BJKK17, KP19]) that can tolerate corruptions only in the *labels*. We are interested in algorithms robust against corruptions in both examples and labels.

mean estimation in the list-decodable setting relied on "supplementing" the SoS method with a somewhat *ad hoc*, problem-dependent technique.

As an important conceptual contribution, our work yields a framework for list-decodable learning that recovers some of the simplicity of the general blueprint. To do this, we give a general method for *rounding pseudo-distributions* in the setting with $\gg 1/2$ fraction outliers. A key step in our rounding builds on the work of [KS19] who developed such a method to give a simpler proof of the list-decodable mean estimation result of [KS17a]. In Section **??**, we explain our ideas in detail.

The results in all the works above hold whenever the underlying distribution satisfies a certain *certified concentration* condition formulated within the SoS system via higher moment bounds. An important contribution of this work is formalizing an *anti-concentration* condition within the SoS system. Unlike the bounded moment condition, there is no canonical phrasing within SoS for such statements. We choose a form that allows proving "certified anti-concentration" for a distribution by showing the existence of a certain approximating polynomial. This allows showing certified anti-concentration of natural distributions via a completely modular approach that relies on a beautiful line of works that construct "weighted " polynomial approximators [Lub07].

We believe that our framework for list-decodable estimation and our formulation of certified anti-concentration condition will likely have further applications in outlier-robust learning.

1.1 Our Results

We first define our model for generating samples for list-decodable regression.

Model 1.1 (Robust Linear Regression)**.** For $0 < \alpha < 1$ and $\ell^* \in \mathbb{R}^d$ with $\|\ell^*\|_2 \leqslant 1$, let $\mathrm{Lin}_D(\alpha, \ell^*)$ denote the following probabilistic process to generate n noisy linear equations $\mathcal{S} = \{\langle x_i, a\rangle = y_i \mid 1 \leqslant i \leqslant n\}$ in variable $a \in \mathbb{R}^d$ with αn *inliers* $\mathcal{I}$ and $(1-\alpha)n$ *outliers* $\mathcal{O}$:

1. Construct $\mathcal{I}$ by choosing αn i.i.d. samples $x_i \sim D$ and set $y_i = \langle x_i, \ell^*\rangle + \zeta$ for additive and independent noise ζ,
2. Construct $\mathcal{O}$ by choosing the remaining $(1-\alpha)n$ equations arbitrarily and potentially adversarially w.r.t the inliers $\mathcal{I}$.

The bound on the norm of ℓ^* is without any loss of generality. Note that α is measure of the "signal" (fraction of inliers) and can be $\ll 1/2$.

An η-approximate algorithm for list-decodable regression takes input a sample from $\mathrm{Lin}_D(\alpha, \ell^*)$ and outputs a *constant* (depending only on α) size list L of linear functions such that there is some $\ell \in L$ that is η-close to ℓ^*.

One of our key conceptual contributions is to identify the strong relationship between *anti-concentration inequalities* and list-decodable regression. Anti-concentration inequalities are well-studied [Erd45, TV12, RV08] in probability theory and combinatorics. The simplest of these inequalities upper bound the probability that a high-dimensional random variable has zero projections in any direction.

Definition 1.2 (Anti-Concentration)**.** A $\mathbb{R}^d$-valued zero-mean random variable Y has a δ-*anti-concentrated* distribution if $\mathbb{P}[\langle Y, v\rangle = 0] < \delta$.

In Proposition **??**, we provide a simple but conceptually illuminating proof that anti-concentration is *sufficient* for list-decodable regression. In Theorem **??**, we prove a sharp converse and show that anti-concentration is information-theoretically *necessary* for even noiseless list-decodable regression. This lower bound surprisingly holds for a natural distribution: uniform distribution on $\{0,1\}^d$ and more generally, uniform distribution on $[q]^d$ for $q = \{0, 1, 2 \ldots, q\}$.

Theorem 1.3 (See Proposition **??** and Theorem **??**). *There is a (inefficient) list-decodable regression algorithm for* $\mathrm{Lin}_D(\alpha, \ell^*)$ *with list size* $O(\frac{1}{\alpha})$ *whenever* D *is* α*-anti-concentrated. Further, there exists a distribution* D *on* $\mathbb{R}^d$ *that is* $(\alpha + \varepsilon)$*-anti-concentrated for every* $\varepsilon > 0$ *but there is no algorithm for* $\frac{\alpha}{2}$*-approximate list-decodable regression for* $\mathrm{Lin}_D(\alpha, \ell^*)$ *that returns a list of size* $< d$.

For our efficient algorithms, we need a *certified* version of the anti-concentration condition. To handle additive noise of variance ζ^2, we need a control of $\mathbb{P}[|\langle x, v\rangle| \leqslant \zeta]$. Thus, we extend our notion of anti-concentration and then define a *certified* analog of it:

Definition 1.4 (Certifiable Anti-Concentration). A random variable Y has a *k-certifiably* (C, δ)-anti-concentrated distribution if there is a univariate polynomial p satisfying $p(0) = 1$ such that there is a degree k sum-of-squares proof of the following two inequalities:

1. $\forall v, \langle Y, v\rangle^2 \leqslant \delta^2\, \mathbb{E}\langle Y, v\rangle^2$ implies $(p(\langle Y, v\rangle) - 1)^2 \leqslant \delta^2$.
2. $\forall v, \|v\|_2^2 \leqslant 1$ implies $\mathbb{E}\, p^2(\langle Y, v\rangle) \leqslant C\delta$.

Intuitively, certified anti-concentration asks for a *certificate* of the anti-concentration property of Y in the "sum-of-squares" proof system (see Section 2 for precise definitions). SoS is a proof system that reasons about polynomial inequalities. Since the "core indicator" $\mathbf{1}(|\langle x, v\rangle| \leqslant \delta)$ is not a polynomial, we phrase the condition in terms of an approximating polynomial p. We are now ready to state our main result.

Theorem 1.5 (List-Decodable Regression). *For every* $\alpha, \eta > 0$ *and a k-certifiably* $(C, \alpha^2\eta^2/10C)$*-anti-concentrated distribution* D *on* $\mathbb{R}^d$*, there exists an algorithm that takes input a sample generated according to* $\mathrm{Lin}_D(\alpha, \ell^*)$ *and outputs a list* L *of size* $O(1/\alpha)$ *such that there is an* $\ell \in L$ *satisfying* $\|\ell - \ell^*\|_2 < \eta$ *with probability at least* 0.99 *over the draw of the sample. The algorithm needs a sample of size* $n = (kd)^{O(k)}$ *and runs in time* $n^{O(k)} = (kd)^{O(k^2)}$.

Remark 1.6 (Tolerating Additive Noise). For additive noise of variance ζ^2 in the inlier labels, our algorithm, in the same running time and sample complexity, outputs a list of size $O(1/\alpha)$ that contains an ℓ satisfying $\|\ell - \ell^*\|_2 \leqslant \frac{\zeta}{\alpha} + \eta$. Since we normalize ℓ^* to have unit norm, this guarantee is meaningful only when $\zeta \ll \alpha$.

Remark 1.7 (Exponential Dependence on $1/\alpha$). List-decodable regression algorithms immediately yield algorithms for mixed linear regression (MLR) without any assumptions on the components. The state-of-the-art algorithm for MLR with gaussian components [LL18] has an exponential dependence on $k = 1/\alpha$ in the running time in the absence of strong pairwise separation or small condition number of the components. Liang and Liu [LL18] (see Page 10) use the relationship to learning mixtures of k gaussians (with an $\exp(k)$ lower bound [MV10]) to note that algorithms with polynomial dependence on $1/\alpha$ for MLR and thus, also for list-decodable regression might not exist.

Certifiably anti-concentrated distributions In Section **??**, we show certifiable anti-concentration of some well-studied families of distributions. This includes the standard gaussian distribution and more generally any anti-concentrated spherically symmetric distribution with strictly sub-exponential tails. We also show that simple operations such as scaling, applying well-conditioned linear transformations and sampling preserve certifiable anti-concentration. This yields:

Corollary 1.8 (List-Decodable Regression for Gaussian Inliers). *For every* $\alpha, \eta > 0$ *there's an algorithm for list-decodable regression for the model* $\mathrm{Lin}_D(\alpha, \ell^*)$ *with* $D = \mathcal{N}(0, \Sigma)$ *with* $\lambda_{\max}(\Sigma)/\lambda_{min}(\Sigma) = O(1)$ *that needs* $n = (d/\alpha\eta)^{O\left(\frac{1}{\alpha^4\eta^4}\right)}$ *samples and runs in time* $n^{O\left(\frac{1}{\alpha^4\eta^4}\right)} = (d/\alpha\eta)^{O\left(\frac{1}{\alpha^8\eta^8}\right)}$.

We note that certifiably anti-concentrated distributions are more restrictive compared to the families of distributions for which the most general robust estimation algorithms

work [KS17b, KS17a, KKM18]. To a certain extent, this is inherent. The families of distributions considered in these prior works do not satisfy anti-concentration in general. And as we discuss in more detail in Section **??**, anti-concentration is information-theoretically *necessary* (see Theorem **??**) for list-decodable regression. This surprisingly rules out families of distributions that might appear natural and "easy", for example, the uniform distribution on $\{0,1\}^n$. In fact, our lower bound shows the impossibility of even the "easier" problem of mixed linear regression on this distribution.

We rescue this to an extent for the special case when ℓ^* in the model $\mathrm{Lin}(\alpha, \ell^*)$ is a "Boolean vector", i.e., has all coordinates of equal magnitude. Intuitively, this helps because while the the uniform distribution on $\{0,1\}^n$ (and more generally, any discrete product distribution) is badly anti-concentrated in sparse directions, they are well anti-concentrated [Erd45] in the directions that are far from any sparse vectors.

As before, for obtaining efficient algorithms, we need to work with a *certified* version (see Definition **??**) of such a restricted anti-concentration condition. As a specific Corollary (see Theorem **??** for a more general statement), this allows us to show:

Theorem 1.9 (List-Decodable Regression for Hypercube Inliers)**.** *For every $\alpha, \eta > 0$ there's an η-approximate algorithm for list-decodable regression for the model $\mathrm{Lin}_D(\alpha, \ell^*)$ with D is uniform on $\{0,1\}^d$ that needs $n = (d/\alpha\eta)^{O(\frac{1}{\alpha^4\eta^4})}$ samples and runs in time $n^{O(\frac{1}{\alpha^4\eta^4})} = (d/\alpha\eta)^{O(\frac{1}{\alpha^8\eta^8})}$.*

In Section **??**, we obtain similar results for general product distributions. It is an important open problem to prove certified anti-concentration for a broader family of distributions.

1.2 Concurrent Work

In an independent and concurrent work, Raghavendra and Yau have given similar results for list-decodable linear regression and also use the sum-of-squares paradigm [RY19].

2 Preliminaries

In this section, we define pseudo-distributions and sum-of-squares proofs. See the lecture notes [**?**] for more details and the appendix in [**?**] for proofs of the propositions appearing here.

Let $x = (x_1, x_2, \ldots, x_n)$ be a tuple of n indeterminates and let $\mathbb{R}[x]$ be the set of polynomials with real coefficients and indeterminates $x_1, \ldots, x_n$. We say that a polynomial $p \in \mathbb{R}[x]$ is a *sum-of-squares (sos)* if there are polynomials $q_1, \ldots, q_r$ such that $p = q_1^2 + \cdots + q_r^2$.

2.1 Pseudo-distributions

Pseudo-distributions are generalizations of probability distributions. We can represent a discrete (i.e., finitely supported) probability distribution over $\mathbb{R}^n$ by its probability mass function $D\colon \mathbb{R}^n \to \mathbb{R}$ 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. Here, we relax the constraint $D \geqslant 0$ and only require that D passes certain low-degree non-negativity tests.

Concretely, a *level-ℓ pseudo-distribution* is a finitely-supported function $D : \mathbb{R}^n \to \mathbb{R}$ 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 $\ell/2$. (Here, the summations are over the support of D.) A straightforward polynomial-interpolation argument shows that every level-∞-pseudo distribution satisfies $D \geqslant 0$ and is thus an actual probability distribution. We define the *pseudo-expectation* of a function f on $\mathbb{R}^d$ with respect to a pseudo-distribution D, denoted $\tilde{\mathbb{E}}_{D(x)} f(x)$, as

$$\tilde{\mathbb{E}}_{D(x)} f(x) = \sum_x D(x) f(x) \ . \tag{2.1}$$

The degree-ℓ moment tensor of a pseudo-distribution D is the tensor $\mathbb{E}_{D(x)}(1, x_1, x_2, \ldots, x_n)^{\otimes \ell}$. In particular, the moment tensor has an entry corresponding to the pseudo-expectation of all monomials of degree at most ℓ in x. The set of all degree-ℓ moment tensors of probability distribution is a convex set. Similarly, the set of all degree-ℓ moment tensors of degree d pseudo-distributions is also convex. Key to the algorithmic utility of pseudo-distributions is the fact that while there can be no efficient separation oracle for the convex set of all degree-ℓ moment tensors of an actual probability distribution, there's a separation oracle running in time $n^{O(\ell)}$ for the convex set of the degree-ℓ moment tensors of all level-ℓ pseudodistributions.

Fact 2.1 ([**?** **?** **?** **?**]). *For any $n, \ell \in \mathbb{N}$, the following set has a $n^{O(\ell)}$-time weak separation oracle (in the sense of [**?**]):*

$$\left\{ \tilde{\mathbb{E}}_{D(x)}(1, x_1, x_2, \ldots, x_n)^{\otimes d} \mid \textit{degree-d pseudo-distribution } D \textit{ over } \mathbb{R}^n \right\} . \tag{2.2}$$

This fact, together with the equivalence of weak separation and optimization [**?**] allows us to efficiently optimize over pseudo-distributions (approximately)—this algorithm is referred to as the sum-of-squares algorithm.

The *level-ℓ sum-of-squares algorithm* optimizes over the space of all level-ℓ pseudo-distributions that satisfy a given set of polynomial constraints—we formally define this next.

Definition 2.2 (Constrained pseudo-distributions)**.** Let D be a level-ℓ pseudo-distribution over $\mathbb{R}^n$. Let $\mathcal{A} = \{f_1 \geqslant 0, f_2 \geqslant 0, \ldots, f_m \geqslant 0\}$ be a system of m polynomial inequality constraints. We say that D *satisfies the system of constraints $\mathcal{A}$ at degree r*, denoted $D \sststile{r}{} \mathcal{A}$, if for every $S \subseteq [m]$ and every sum-of-squares polynomial h with $\deg h + \sum_{i \in S} \max\{\deg f_i, r\}$,

$$\tilde{\mathbb{E}}_D \, h \cdot \prod_{i \in S} f_i \geqslant 0 \, .$$

We write $D \models \mathcal{A}$ (without specifying the degree) if $D \sststile{0}{} \mathcal{A}$ holds. Furthermore, we say that $D \sststile{r}{} \mathcal{A}$ holds *approximately* if the above inequalities are satisfied up to an error of $2^{-n^\ell} \cdot \|h\| \cdot \prod_{i \in S} \|f_i\|$, where $\|\cdot\|$ denotes the Euclidean norm[2] of the cofficients of a polynomial in the monomial basis.

We remark that if D is an actual (discrete) probability distribution, then we have $D \models \mathcal{A}$ if and only if D is supported on solutions to the constraints $\mathcal{A}$.

We say that a system $\mathcal{A}$ of polynomial constraints is *explicitly bounded* if it contains a constraint of the form $\{\|x\|^2 \leqslant M\}$. The following fact is a consequence of Fact 2.1 and [**?**],

Fact 2.3 (Efficient Optimization over Pseudo-distributions)**.** *There exists an $(n+m)^{O(\ell)}$-time algorithm that, given any explicitly bounded and satisfiable system[3] $\mathcal{A}$ of m polynomial constraints in n variables, outputs a level-ℓ pseudo-distribution that satisfies $\mathcal{A}$ approximately.*

2.2 Sum-of-squares proofs

Let $f_1, f_2, \ldots, f_r$ and g be multivariate polynomials in x. A *sum-of-squares proof* that the constraints $\{f_1 \geqslant 0, \ldots, f_m \geqslant 0\}$ imply the constraint $\{g \geqslant 0\}$ consists of polynomials $(p_S)_{S \subseteq [m]}$ such that

$$g = \sum_{S \subseteq [m]} p_S \cdot \Pi_{i \in S} f_i \, . \tag{2.3}$$

We say that this proof has *degree* ℓ if for every set $S \subseteq [m]$, the polynomial $p_S \Pi_{i \in S} f_i$ has degree at most ℓ. If there is a degree ℓ SoS proof that $\{f_i \geqslant 0 \mid i \leqslant r\}$ implies $\{g \geqslant 0\}$, we

[2]The choice of norm is not important here because the factor 2^{-n^ℓ} swamps the effects of choosing another norm.

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

write:

$$\{f_i \geqslant 0 \mid i \leqslant r\} \big|\!\frac{}{}_{\ell}\ \{g \geqslant 0\}\,. \tag{2.4}$$

Sum-of-squares proofs satisfy the following inference rules. For all polynomials $f, g\colon \mathbb{R}^n \to \mathbb{R}$ and for all functions $F\colon \mathbb{R}^n \to \mathbb{R}^m$, $G\colon \mathbb{R}^n \to \mathbb{R}^k$, $H\colon \mathbb{R}^p \to \mathbb{R}^n$ such that each of the coordinates of the outputs are polynomials of the inputs, we have:

$$\frac{\mathcal{A} \big|\!\frac{}{}_{\ell}\ \{f \geqslant 0, g \geqslant 0\}}{\mathcal{A} \big|\!\frac{}{}_{\ell}\ \{f+g \geqslant 0\}}, \frac{\mathcal{A} \big|\!\frac{}{}_{\ell}\ \{f \geqslant 0\}, \mathcal{A} \big|\!\frac{}{}_{\ell'}\ \{g \geqslant 0\}}{\mathcal{A} \big|\!\frac{}{}_{\ell+\ell'}\ \{f \cdot g \geqslant 0\}} \qquad \text{(addition and multiplication)}$$

$$\frac{\mathcal{A} \big|\!\frac{}{}_{\ell}\ \mathcal{B}, \mathcal{B} \big|\!\frac{}{}_{\ell'}\ C}{\mathcal{A} \big|\!\frac{}{}_{\ell\cdot\ell'}\ C} \qquad \text{(transitivity)}$$

$$\frac{\{F \geqslant 0\} \big|\!\frac{}{}_{\ell}\ \{G \geqslant 0\}}{\{F(H) \geqslant 0\} \big|\!\frac{}{}_{\ell\cdot\deg(H)}\ \{G(H) \geqslant 0\}}\,. \qquad \text{(substitution)}$$

Low-degree sum-of-squares proofs are sound and complete if we take low-level pseudo-distributions as models.

Concretely, sum-of-squares proofs allow us to deduce properties of pseudo-distributions that satisfy some constraints.

Fact 2.4 (Soundness)**.** *If $D \big|\!\models_{r}\ \mathcal{A}$ for a level-ℓ pseudo-distribution D and there exists a sum-of-squares proof $\mathcal{A} \big|\!\frac{}{}_{r'}\ \mathcal{B}$, then $D \big|\!\models_{r\cdot r'+r'}\ \mathcal{B}$.*

If the pseudo-distribution D satisfies $\mathcal{A}$ only approximately, soundness continues to hold if we require an upper bound on the bit-complexity of the sum-of-squares $\mathcal{A} \big|\!\frac{}{}_{r'}\ B$ (number of bits required to write down the proof).

In our applications, the bit complexity of all sum of squares proofs will be $n^{O(\ell)}$ (assuming that all numbers in the input have bit complexity $n^{O(1)}$). This bound suffices in order to argue about pseudo-distributions that satisfy polynomial constraints approximately.

The following fact shows that every property of low-level pseudo-distributions can be derived by low-degree sum-of-squares proofs.

Fact 2.5 (Completeness)**.** *Suppose $d \geqslant r' \geqslant r$ and $\mathcal{A}$ is a collection of polynomial constraints with degree at most r, and $\mathcal{A} \vdash \{\sum_{i=1}^n x_i^2 \leqslant B\}$ for some finite B.*

Let $\{g \geqslant 0\}$ be a polynomial constraint. If every degree-d pseudo-distribution that satisfies $D \big|\!\models_{r}\ \mathcal{A}$ also satisfies $D \big|\!\models_{r'}\ \{g \geqslant 0\}$, then for every $\varepsilon > 0$, there is a sum-of-squares proof $\mathcal{A} \big|\!\frac{}{}_{d}\ \{g \geqslant -\varepsilon\}$.

We will use the following Cauchy-Schwarz inequality for pseudo-distributions:

Fact 2.6 (Cauchy-Schwarz for Pseudo-distributions)**.** *Let f, g be polynomials of degree at most d in indeterminate $x \in \mathbb{R}^d$. Then, for any degree d pseudo-distribution $\tilde{\mu}$, $\tilde{\mathbb{E}}_{\tilde{\mu}}[fg] \leqslant \sqrt{\tilde{\mathbb{E}}_{\tilde{\mu}}[f^2]}\sqrt{\tilde{\mathbb{E}}_{\tilde{\mu}}[g^2]}$.*

The following fact is a simple corollary of the fundamental theorem of algebra:

Fact 2.7. *For any univariate degree d polynomial $p(x) \geqslant 0$ for all $x \in \mathbb{R}$, $\big|\!\frac{x}{d}\ \{p(x) \geqslant 0\}$.*

This can be extended to univariate polynomial inequalities over intervals of $\mathbb{R}$.

Fact 2.8 (Fekete and Markov-Lukács, see [**?**])**.** *For any univariate degree d polynomial $p(x) \geqslant 0$ for $x \in [a, b]$, $\{x \geqslant a, x \leqslant b\} \big|\!\frac{x}{d}\ \{p(x) \geqslant 0\}$.*

References

[ABL13] Pranjal Awasthi, Maria-Florina Balcan, and Philip M. Long, *The power of localization for efficiently learning linear separators with malicious noise*, CoRR **abs/1307.8371** (2013). 1

[BBV08] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala, *A discriminative framework for clustering via similarity functions*, STOC, ACM, 2008, pp. 671–680. 2

[Ber06] Thorsten Bernholt, *Robust estimators are hard to compute*, Tech. report, Technical Report/Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen, 2006. 1

[BJKK17] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar, *Consistent robust regression*, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2017, pp. 2107–2116. 2

[BWY14] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu, *Statistical guarantees for the EM algorithm: From population to sample-based analysis*, CoRR **abs/1408.2156** (2014). 2

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant, *Learning from untrusted data*, STOC, ACM, 2017, pp. 47–60. 1, 2

[CYC14] Yudong Chen, Xinyang Yi, and Constantine Caramanis, *A convex formulation for mixed regression with two components: Minimax optimal rates*, Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014, 2014, pp. 560–604. 2

[DKK+16a] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart, *Robust estimators in high dimensions without the computational intractability*, FOCS, IEEE Computer Society, 2016, pp. 655–664. 2

[DKK+16b] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Zheng Li, Ankur Moitra, and Alistair Stewart, *Robust estimators in high dimensions without the computational intractability*, CoRR **abs/1604.06443** (2016). 1

[DKK+17] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart, *Robustly learning a gaussian: Getting optimal error, efficiently*, CoRR **abs/1704.03866** (2017). 1, 2

[DKK+18] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li 0001, Jacob Steinhardt, and Alistair Stewart, *Sever: A robust meta-algorithm for stochastic optimization*, CoRR **abs/1803.02815** (2018). 2

[DKS17] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart, *Learning geometric concepts with nasty noise*, CoRR **abs/1707.01242** (2017). 1

[DKS18] _____, *List-decodable robust mean estimation and learning mixtures of spherical gaussians*, Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018, 2018, pp. 1047–1060. 2

[DKS19] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart, *Efficient algorithms and lower bounds for robust linear regression*, Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019 (Timothy M. Chan, ed.), SIAM, 2019, pp. 2745–2754. 2

[DV89] Richard D. De Veaux, *Mixtures of linear regressions*, Comput. Statist. Data Anal. **8** (1989), no. 3, 227–245. MR 1028403 2

[Erd45] P. Erdös, *On a lemma of littlewood and offord*, Bull. Amer. Math. Soc. **51** (1945), no. 12, 898–902. 3, 5

[FS10] Susana Faria and Gilda Soromenho, *Fitting mixtures of linear regressions*, J. Stat. Comput. Simul. **80** (2010), no. 1-2, 201–225. MR 2757044 2

[HL17] Sam B. Hopkins and Jerry Li, *Mixture models, robustness, and sum of squares proofs*, 2017. 1, 2

[HRRS11] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel, *Robust statistics: the approach based on influence functions*, vol. 114, John Wiley & Sons, 2011. 1

[Hub11] Peter J Huber, *Robust statistics*, International Encyclopedia of Statistical Science, Springer, 2011, pp. 1248–1251. 1

[JJ94] Michael I. Jordan and Robert A. Jacobs, *Hierarchical mixtures of experts and the em algorithm*, Neural Computation **6** (1994), no. 2, 181–214. 2

[KKM18] Adam R. Klivans, Pravesh K. Kothari, and Raghu Meka, *Efficient algorithms for outlier-robust regression*, Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018., 2018, pp. 1420–1430. 1, 2, 5

[KLS09] Adam R. Klivans, Philip M. Long, and Rocco A. Servedio, *Learning halfspaces with malicious noise*, Journal of Machine Learning Research **10** (2009), 2715–2740. 1

[KP19] Sushrut Karmalkar and Eric Price, *Compressed sensing with adversarial sparse noise via l1 regression*, SOSA@SODA (Jeremy T. Fineman and Michael Mitzenmacher, eds.), OASICS, vol. 69, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2019, pp. 19:1–19:19. 2

[KS17a] Pravesh K. Kothari and Jacob Steinhardt, *Better agnostic clustering via relaxed tensor norms*, 2017. 1, 2, 3, 5

[KS17b] Pravesh K. Kothari and David Steurer, *Outlier-robust moment-estimation via sum-of-squares*, CoRR **abs/1711.11581** (2017). 1, 2, 5

[KS19] _____, *List-decodable mean estimation made simple*, Manuscript, 2019. 3

[LL18] Yuanzhi Li and Yingyu Liang, *Learning mixtures of linear regressions with nearly optimal complexity*, Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018., 2018, pp. 1125–1144. 2, 4

[LRV16] Kevin A. Lai, Anup B. Rao, and Santosh Vempala, *Agnostic estimation of mean and covariance*, FOCS, IEEE Computer Society, 2016, pp. 665–674. 1

[Lub07] Doron S Lubinsky, *A Survey of Weighted Approximation for Exponential Weights*, arXiv Mathematics e-prints (2007), math/0701099. 3

[MMY06] RARD Maronna, R Douglas Martin, and Victor Yohai, *Robust statistics*, John Wiley & Sons, Chichester. ISBN, 2006. 1

[MV10] Ankur Moitra and Gregory Valiant, *Settling the polynomial learnability of mixtures of gaussians*, FOCS, IEEE Computer Society, 2010, pp. 93–102. 4

[PSBR18] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar, *Robust estimation via robust gradient estimation*, CoRR **abs/1802.06485** (2018). 2

[RV08] Mark Rudelson and Roman Vershynin, *The Littlewood-Offord problem and invertibility of random matrices*, Adv. Math. **218** (2008), no. 2, 600–633. MR 2407948 3

[RY19] Prasad Raghavendra and Morris Yau, *List decodable learning via sum of squares*, Manuscript, 2019. 1, 5

[SJA16] Hanie Sedghi, Majid Janzamin, and Anima Anandkumar, *Provable tensor methods for learning mixtures of generalized linear models*, AISTATS, JMLR Workshop and Conference Proceedings, vol. 51, JMLR.org, 2016, pp. 1223–1231. 2

[Tuk75] John W. Tukey, *Mathematics and the picturing of data*, 523–531. MR 0426989 1

[TV12] Terence Tao and Van Vu, *The Littlewood-Offord problem in high dimensions and a conjecture of Frankl and Füredi*, Combinatorica **32** (2012), no. 3, 363–372. MR 2965282 3

[YCS13] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi, *Alternating Minimization for Mixed Linear Regression*, arXiv e-prints (2013), arXiv:1310.3745. 2

[ZJD16] Kai Zhong, Prateek Jain, and Inderjit S. Dhillon, *Mixed linear regression with multiple components*, NIPS, 2016, pp. 2190–2198. 2