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 $\{(x_i, y_i)\}_{i \leq n}$ 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 [37] have obtained a similar result for list-decodable regression also using the sum-of-squares method.

1 Introduction

2 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 [41, 34, 22, 20]. However, the algorithms resulting from this line of work usually run in time exponential in the dimension of the data [6]. An influential line of recent work [26, 1, 14, 30, 8, 27, 28, 21, 13, 15, 25] 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 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 [3] to give a discriminative framework for clustering. More recently, beginning with [8], various works [16, 27] have considered this as a model of *untrusted* data.

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

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* [12, 13], separation oracles for inliers [12] or the *sum-of-squares* method [28, 21, 27, 25]). In contrast, progress on algorithms that tolerate the significantly harsher conditions in the list-decodable setting has been slower. The only prior works [8, 16, 27] 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). A key conceptual insight in this result is the observation that list-decodable regression information-theoretically requires the inlier-distribution to be *anti-concentrated*. Our algorithm succeeds whenever the distribution satisfies a stronger "algorithmically usable" *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 [25, 17, 11, 36] 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 [10, 23, 19, 42, 2, 9, 43, 39, 31] 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 [21, 27] and obtaining algorithms for moment estimation [28] and linear regression [25] 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 [21, 27] and obtaining algorithms for moment estimation [28] and linear regression [25] 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 [27] for 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 [29] who developed such a method to give a simpler proof of the list-decodable mean estimation result of [27]. In Section 3, 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 [32].

[1] There's a long line of work on robust regression algorithms (see for e.g. [7, 24]) that can tolerate corruptions only in the *labels*. We are interested in algorithms robust against corruptions in both examples and labels.

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.

2.1 Our Results

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

Model 2.1 (Robust Linear Regression)**.** For $0 < \alpha < 1$ and $\ell^* \in \mathbb{R}^d$ with $\|\ell^*\|_2 \leq 1$, let $\mathrm{Lin}_D(\alpha, \ell^*)$ denote the following probabilistic process to generate n noisy linear equations $\mathcal{S} = \{\langle x_i, a\rangle = y_i \mid 1 \leq i \leq n\}$ 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 noise ζ,
2. Construct $\mathcal{O}$ by choosing the remaining $(1-\alpha)n$ equations arbitrarily and potentially adversarially w.r.t the inliers $\mathcal{I}$.

Note that α measures the "signal" (fraction of inliers) and can be $\ll 1/2$. The bound on the norm of ℓ^* is without any loss of generality. For the sake of exposition, we will restrict to $\zeta = 0$ for most of this paper and discuss (see Remarks 2.6 and **??**) how our algorithms can tolerate additive noise.

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 [18, 40, 38] 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 2.2 (Anti-Concentration)**.** A $\mathbb{R}^d$-valued zero-mean random variable Y has a δ-*anti-concentrated* distribution if $\Pr[\langle Y, v\rangle = 0] < \delta$.

In Proposition 3.4, 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 2.3 (See Proposition 3.4 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* $(\alpha + \epsilon)$*-anti-concentrated for every* $\epsilon > 0$ *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 $\Pr[|\langle x, v\rangle| \leq \zeta]$. Thus, we extend our notion of anti-concentration and then define a *certified* analog of it:

Definition 2.4 (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. $\forall v$, $\langle Y, v\rangle^2 \leq \delta^2 \mathbb{E}\langle Y, v\rangle^2$ implies $(p(\langle Y, v\rangle) - 1)^2 \leq \delta^2$.
2. $\forall v$, $\|v\|_2^2 \leq 1$ implies $\mathbb{E}p^2(\langle Y, v\rangle) \leq C\delta$.

Intuitively, certified anti-concentration asks for a *certificate* of the anti-concentration property of Y in the "sum-of-squares" proof system (see Section **??** for precise definitions). SoS is a proof system that reasons about polynomial inequalities. Since the "core indicator" $\mathbf{1}(|\langle x, v\rangle| \leq \delta)$ 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 2.5 (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 2.6 (Tolerating Additive Noise). For additive noise (not necessarily independent) 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 $\|\ell - \ell^*\|_2 \leq \frac{\zeta}{\alpha} + \eta$. Since we normalize ℓ^* to have unit norm, this guarantee is meaningful only when $\zeta \ll \alpha$.

Remark 2.7 (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 [31] 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 [31] (see Page 10) use the relationship to learning mixtures of k gaussians (with an $\exp(k)$ lower bound [35]) 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 2.8 (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 [28, 27, 25]. 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 3, 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 [18] 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 2.9 (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.

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 [37].

3 Overview of our Technique

In this section, we illustrate the important ideas in our algorithm for list-decodable regression. Thus, given a sample $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ from $\mathrm{Lin}_D(\alpha, \ell^*)$, we must construct a constant-size list L of linear functions containing an ℓ close to ℓ^*.

Our algorithm is based on the sum-of-squares method. We build on the "identifiability to algorithms" paradigm developed in several prior works [5, 4, 33, 28, 21, 27, 25] with some important conceptual differences.

An *inefficient* algorithm Let's start by designing an inefficient algorithm for the problem. This may seem simple at the outset. But as we'll see, solving this relaxed problem will rely on some important conceptual ideas that will serve as a starting point for our efficient algorithm.

Without computational constraints, it is natural to just return the list L of all linear functions ℓ that correctly labels all examples in some $S \subseteq \mathcal{S}$ of size αn. We call such an S, a large, *soluble* set. True inliers $\mathcal{I}$ satisfy our search criteria so $\ell^* \in L$. However, it's not hard to show (Proposition **??**) that one can choose outliers so that the list so generated has size $\exp(d)$ (far from a fixed constant!).

A potential fix is to search instead for a *coarse soluble partition* of $\mathcal{S}$, if it exists, into disjoint $S_1, S_2, \ldots, S_k$ and linear functions $\ell_1, \ell_2, \ldots, \ell_k$ so that every $|S_i| \geq \alpha n$ and ℓ_i correctly computes the labels in S_i. In this setting, our list is small ($k \leq 1/\alpha$). But it is easy to construct samples $\mathcal{S}$ for which this fails because there are coarse soluble partitions of $\mathcal{S}$ where every ℓ_i is far from ℓ^*.

Anti-Concentration It turns out that any (even inefficient) algorithm for list-decodable regression provably (see Theorem **??**) *requires* that the distribution of inliers[2] be sufficiently *anti-concentrated*:

Definition 3.1 (Anti-Concentration)**.** A $\mathbb{R}^d$-valued random variable Y with mean 0 is δ-anti-concentrated[3] if for all non-zero v, $\Pr[\langle Y, v\rangle = 0] < \delta$. A set $T \subseteq \mathbb{R}^d$ is δ-anti-concentrated if the uniform distribution on T is δ-anti-concentrated.

As we discuss next, anti-concentration is also *sufficient* for list-decodable regression. Intuitively, this is because anti-concentration of the inliers prevents the existence of a soluble set that intersects significantly with $\mathcal{I}$ and yet can be labeled correctly by $\ell \neq \ell^*$. This is simple to prove in the special case when $\mathcal{S}$ admits a coarse soluble partition.

Proposition 3.2. *Suppose $\mathcal{I}$ is α-anti-concentrated. Suppose there exists a partition $S_1, S_2, \ldots, S_k \subseteq \mathcal{S}$ such that each $|S_i| \geq \alpha n$ and there exist $\ell_1, \ell_2, \ldots, \ell_k$ such that $y_j = \langle \ell_i, x_j\rangle$ for every $j \in S_i$. Then, there is an i such that $\ell_i = \ell^*$.*

Proof. Since $k \leq 1/\alpha$, there is a j such that $|\mathcal{I} \cap S_j| \geq \alpha|\mathcal{I}|$. Then, $\langle x_i, \ell_j\rangle = \langle x_i, \ell^*\rangle$ for every $i \in \mathcal{I} \cap S_j$. Thus, $\Pr_{i \sim \mathcal{I}}[\langle x_i, \ell_j - \ell^*\rangle = 0] \geq \alpha$. This contradicts anti-concentration of $\mathcal{I}$ unless $\ell_j - \ell^* = 0$. □

The above proposition allows us to use *any* soluble partition as a *certificate* of correctness for the associated list L. Two aspects of this certificate were crucial in the above argument: 1) *largeness*: each S_i is of size αn - so the generated list is small, and, 2) *uniformity*: every sample is used in exactly one of the sets so $\mathcal{I}$ must intersect one of the S_is in at least α-fraction of the points.

Identifiability via anti-concentration For arbitrary $\mathcal{S}$, a coarse soluble partition might not exist. So we will generalize coarse soluble partitions to obtain certificates that exist for every sample $\mathcal{S}$ and guarantee largeness and a relaxation of uniformity (formalized below). For this purpose, it is convenient to view such certificates as distributions μ on $\geq \alpha n$ size soluble subsets of $\mathcal{S}$ so any collection $\mathcal{C} \subseteq 2^{\mathcal{S}}$ of αn size sets corresponds to the uniform distribution μ on $\mathcal{C}$.

To precisely define uniformity, let $W_i(\mu) = \mathbb{E}_{S \sim \mu}[\mathbf{1}(i \in S)]$ be the "frequency of i", that is, probability that the ith sample is chosen to be in a set drawn according to μ. Then, the uniform distribution μ on any coarse soluble k-partition satisfies $W_i = \frac{1}{k}$ for every i. That is, all samples $i \in \mathcal{S}$ are *uniformly* used in such a μ. To generalize this idea, we define $\sum_i W_i(\mu)^2$ as the *distance*

[2] As in the standard robust estimation setting, the outliers are arbitrary and potentially adversarially chosen.
[3] Definition 2.4 differs slightly to handle list-decodable regression with additive noise in the inliers.

to uniformity of μ. Up to a shift, this is simply the variance in the frequencies of the points in $\mathcal{S}$ used in draws from μ. Our generalization of a coarse soluble partition of $\mathcal{S}$ is any μ that minimizes $\sum_i W_i(\mu)^2$, the distance to uniformity, and is thus *maximally uniform* among all distributions supported on large soluble sets. Such a μ can be found by convex programming.

The following claim generalizes Proposition 3.2 to derive the same conclusion starting from any maximally uniform distribution supported on large soluble sets.

Proposition 3.3. *For a maximally uniform μ on αn size soluble subsets of $\mathcal{S}$, $\sum_{i \in \mathcal{I}} \mathbb{E}_{S \sim \mu}[\mathbf{1}(i \in S)] \geq \alpha |\mathcal{I}|$.*

The proof proceeds by contradiction (see Lemma **??**). We show that if $\sum_{i \in \mathcal{I}} W_i(\mu) \leq \alpha |\mathcal{I}|$, then we can strictly reduce the distance to uniformity by taking a mixture of μ with the distribution that places all its probability mass on $\mathcal{I}$. This allow us to obtain an (inefficient) algorithm for list-decodable regression establishing identifiability.

Proposition 3.4 (Identifiability for List-Decodable Regression)**.** *Let $\mathcal{S}$ be sample from* $\mathrm{Lin}(\alpha, \ell^*)$ *such that $\mathcal{I}$ is δ-anti-concentrated for $\delta < \alpha$. Then, there's an (inefficient) algorithm that finds a list L of size $\frac{20}{\alpha - \delta}$ such that $\ell^* \in L$ with probability at least* 0.99.

Proof. Let μ be *any* maximally uniform distribution over αn size soluble subsets of $\mathcal{S}$. For $k = \frac{20}{\alpha - \delta}$, let $S_1, S_2, \ldots, S_k$ be independent samples from μ. Output the list L of k linear functions that correctly compute the labels in each S_i.

To see why $\ell^* \in L$, observe that $\mathbb{E}|S_j \cap \mathcal{I}| = \sum_{i \in \mathcal{I}} \mathbb{E} \mathbf{1}(i \in S_j) \geq \alpha |\mathcal{I}|$. By averaging, $\Pr[|S_j \cap \mathcal{I}| \geq \frac{\alpha + \delta}{2} |\mathcal{I}|] \geq \frac{\alpha - \delta}{2}$. Thus, there's a $j \leq k$ so that $|S_j \cap \mathcal{I}| \geq \frac{\alpha + \delta}{2} |\mathcal{I}|$ with probability at least $1 - (1 - \frac{\alpha - \delta}{2})^{\frac{20}{\alpha - \delta}} \geq 0.99$. We can now repeat the argument in the proof of Proposition 3.2 to conclude that any linear function that correctly labels S_j must equal ℓ^*. □

An efficient algorithm Our identifiability proof suggests the following simple algorithm: 1) find *any* maximally uniform distribution μ on soluble subsets of size αn of $\mathcal{S}$, 2) take $O(1/\alpha)$ samples S_i from μ and 3) return the list of linear functions that correctly label the equations in S_is. This is inefficient because searching over distributions is NP-hard in general.

To make this into an efficient algorithm, we start by observing that soluble subsets $S \subseteq \mathcal{S}$ of size αn can be described by the following set of quadratic equations where w stands for the indicator of S and ℓ, the linear function that correctly labels the examples in S.

$$\mathcal{A}_{w,\ell} \colon \left\{ \begin{array}{lr} & \sum_{i=1}^{n} w_i = \alpha n \\ \forall i \in [n]. & w_i^2 = w_i \\ \forall i \in [n]. & w_i \cdot (y_i - \langle x_i, \ell \rangle) = 0 \\ & \|\ell\|^2 \leq 1 \end{array} \right\} \tag{3.1}$$

Our efficient algorithm searches for a maximally uniform *pseudo-distribution* on w satisfying (3.1). Degree k pseudo-distributions (see Section **??** for precise definitions) are generalization of distributions that nevertheless "behave" just as distributions whenever we take (pseudo)-expectations (denoted by $\tilde{\mathbb{E}}$) of a class of degree k polynomials. And unlike distributions, degree k pseudo-distributions satisfying[4] polynomial constraints (such as (3.1)) can be computed in time $n^{O(k)}$.

For the sake of intuition, it might be helpful to (falsely) think of pseudo-distributions $\tilde{\mu}$ as simply distributions where we only get access to moments of degree $\leq k$. Thus, we are allowed to compute expectations of all degree $\leq k$ polynomials with respect to $\tilde{\mu}$. Since $W_i(\tilde{\mu}) = \tilde{\mathbb{E}}_{\tilde{\mu}} w_i$ are just first moments of $\tilde{\mu}$, our notion of maximally uniform distributions extends naturally to pseudo-distributions. This allows us to prove an analog of Proposition 3.3 for pseudo-distributions and gives us an efficient replacement for Step 1.

Proposition 3.5. *For any maximally uniform $\tilde{\mu}$ of degree ≥ 2, $\sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}}[w_i] \geq \alpha |\mathcal{I}| = \alpha \sum_{i \in [n]} \tilde{\mathbb{E}}_{\tilde{\mu}}[w_i]$.*

[4] See Fact **??** for a precise statement.

For Step 2, however, we hit a wall: it's not possible to obtain independent samples from $\tilde{\mu}$ given only low-degree moments. Our algorithm relies on an alternative strategy instead.

Consider the vector $v_i = \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[w_i \ell]}{\tilde{\mathbb{E}}_{\tilde{\mu}}[w_i]}$ whenever $\tilde{\mathbb{E}}_{\tilde{\mu}}[w_i] \neq 0$ (set v_i to zero, otherwise). This is simply the (scaled) average, according to $\tilde{\mu}$, of all the linear functions ℓ that are used to label the sets S of size αn in the support of $\tilde{\mu}$ whenever $i \in S$. Further, v_i depends only on the first two moments of $\tilde{\mu}$.

We think of v_is as "guesses"made by the ith sample for the unknown linear function. Let us focus our attention on the guesses v_i of $i \in \mathcal{I}$ - the inliers. We will show that according to the distribution proportional to $\tilde{\mathbb{E}}[w]$, the average squared distance of v_i from ℓ^* is at max η:

$$\frac{1}{\sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}[w_i]} \sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}[w_i] \|v_i - \ell^*\|_2 < \eta \,. \tag{$\star$}$$

Before diving into ($\star$), let's see how it gives us our efficient list-decodable regression algorithm:

1. Find a pseudo-distribution $\tilde{\mu}$ satisfying (3.1) that minimizes distance to uniformity $\sum_i \tilde{\mathbb{E}}_{\tilde{\mu}}[w_i]^2$.
2. For $O(\frac{1}{\alpha})$ times, independently choose a random index $i \in [n]$ with probability proportional to $\tilde{\mathbb{E}}_{\tilde{\mu}}[w_i]$ and return the list of corresponding v_is.

Step 1 above is a convex program and can be solved in polynomial time. Let's analyze step 2 to see why the algorithm works. Using ($\star$) and Markov's inequality, conditioned on $i \in \mathcal{I}$, $\|v_i - \ell^*\|_2 \leq 2\eta$ with probability $\geq 1/2$. By Proposition 3.5, $\frac{\sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}[w_i]}{\sum_{i \in [n]} \tilde{\mathbb{E}}[w_i]} \geq \alpha$ so $i \in \mathcal{I}$ with probability at least α. Thus in each iteration of step 2, with probability at least $\alpha/2$, we choose an i such that v_i is 2η-close to ℓ^*. Repeating $O(1/\alpha)$ times gives us the 0.99 chance of success.

($\star$) via anti-concentration As in the information-theoretic argument, ($\star$) relies on the anti-concentration of $\mathcal{I}$. Let's do a quick proof for the case when $\tilde{\mu}$ is an actual distribution μ.

Proof of ($\star$) for actual distributions μ. Observe that μ is a distribution over (w, ℓ) satisfying (3.1). Recall that w indicates a subset $S \subseteq \mathcal{S}$ of size αn and $w_i = 1$ iff $i \in S$. And $\ell \in \mathbb{R}^d$ satisfies all the equations in S.

By Cauchy-Schwarz, $\sum_i \|\mathbb{E}[w_i \ell] - \mathbb{E}[w_i]\ell^*\| \leq \mathbb{E}_\mu[\sum_{i \in \mathcal{I}} w_i \|\ell - \ell^*\|]$. Next, as in Proposition 3.2, since $\mathcal{I}$ is η-anti-concentrated, and for all S such that $|\mathcal{I} \cap S| \geq \eta|\mathcal{I}|$, $\ell - \ell^* = 0$. Thus, any such S in the support of μ contributes 0 to the expectation above. We will now show that the contribution from the remaining terms is upper bounded by η. Observe that since $\|\ell - \ell^*\| \leq 2$, $\mathbb{E}[\sum_{i \in \mathcal{I}} w_i \|\ell - \ell^*\|] = \mathbb{E}[\mathbf{1}\left(|S \cap \mathcal{I}| < \eta|\mathcal{I}|\right) w_i \|\ell - \ell^*\|] = \mathbb{E}[\sum_{i \in S \cap \mathcal{I}} \|\ell - \ell^*\|] \leq 2\eta|\mathcal{I}|$. □

SoSizing Anti-Concentration The key to proving ($\star$) for pseudo-distributions is a *sum-of-squares* (SoS) proof of anti-concentration inequality: $\Pr_{x \sim \mathcal{I}}[\langle x, v \rangle = 0] \leq \eta$ in variable v. SoS (see Section **??**) is a restricted system for proving polynomial inequalities subject to polynomial inequality constraints. Thus, to even ask for a SoS proof we must phrase anti-concentration as a polynomial inequality.

To do this, let $p(z)$ be a low-degree polynomial approximator for the function $\mathbf{1}\,(z = 0)$. Then, we can hope to "replace" the use of the inequality $\Pr_{x \sim \mathcal{I}}[\langle x, v \rangle = 0] \leq \eta \equiv \mathbb{E}_{x \sim \mathcal{I}}[\mathbf{1}(\langle x, v \rangle = 0)] \leq \eta$ in the argument above by $\mathbb{E}_{x \sim \mathcal{I}}[p(\langle x, v \rangle)^2] \leq \eta$. Since polynomials grow unboundedly for large enough inputs, it is *necessary* for the uniform distribution on $\mathcal{I}$ to have sufficiently light-tails to ensure that $\mathbb{E}_{x \sim \mathcal{I}} p(\langle x, v \rangle)^2$ is small. In Lemma **??**, we show that anti-concentration and light-tails are *sufficient* to construct such a polynomial.

We can finally ask for a SoS proof for $\mathbb{E}_{x \sim \mathcal{I}} p(\langle x, v \rangle) \leq \eta$ in variable v. We prove such *certified* anti-concentration inequalities for broad families of inlier distributions in Section **??**.

References

[1] 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

[2] 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

[3] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In *STOC*, pages 671–680. ACM, 2008. 1

[4] Boaz Barak, Jonathan A. Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method [extended abstract]. In *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing*, pages 143–151. ACM, New York, 2015. 5

[5] Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *COLT*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 417–445. JMLR.org, 2016. 5

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

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

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

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

[10] Richard D. De Veaux. Mixtures of linear regressions. *Comput. Statist. Data Anal.*, 8(3):227–245, 1989. 2

[11] 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

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

[13] 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

[14] 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

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

[16] 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, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1047–1060, 2018. 1, 2

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

[18] P. Erdös. On a lemma of littlewood and offord. *Bull. Amer. Math. Soc.*, 51(12):898–902, 12 1945. 3, 4

[19] Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *J. Stat. Comput. Simul.*, 80(1-2):201–225, 2010. 2

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

[21] Sam B. Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. 2017. 1, 2, 5

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

[23] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6(2):181–214, 1994. 2

[24] Sushrut Karmalkar and Eric Price. Compressed sensing with adversarial sparse noise via l1 regression. In Jeremy T. Fineman and Michael Mitzenmacher, editors, *SOSA@SODA*, volume 69 of *OASICS*, pages 19:1–19:19. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2019. 2

[25] Adam R. Klivans, Pravesh K. Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 1420–1430, 2018. 1, 2, 4, 5

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

[27] Pravesh K. Kothari and Jacob Steinhardt. Better agnostic clustering via relaxed tensor norms. 2017. 1, 2, 4, 5

[28] Pravesh K. Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. *CoRR*, abs/1711.11581, 2017. 1, 2, 4, 5

[29] Pravesh K. Kothari and David Steurer. List-decodable mean estimation made simple. In *Manuscript*, 2019. 2

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

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

[32] Doron S Lubinsky. A Survey of Weighted Approximation for Exponential Weights. *arXiv Mathematics e-prints*, page math/0701099, Jan 2007. 2

[33] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *FOCS*, pages 438–446. IEEE Computer Society, 2016. 5

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

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

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

[37] Prasad Raghavendra and Morris Yau. List decodable learning via sum of squares. In *Manuscript*, 2019. 1, 4

[38] Mark Rudelson and Roman Vershynin. The Littlewood-Offord problem and invertibility of random matrices. *Adv. Math.*, 218(2):600–633, 2008. 3

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

[40] Terence Tao and Van Vu. The Littlewood-Offord problem in high dimensions and a conjecture of Frankl and Füredi. *Combinatorica*, 32(3):363–372, 2012. 3

[41] John W. Tukey. Mathematics and the picturing of data. pages 523–531, 1975. 1

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

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