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.

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 [52, 43, 29, 27]. However, the algorithms resulting from this line of work usually run in time exponential in the dimension of the data [7]. An influential line of recent work [33, 1, 16, 37, 9, 34, 35, 28, 15, 17, 32] 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 [9], various works [18, 34] 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* [14, 15], separation oracles for inliers [14] or the *sum-of-squares* method [35, 28, 34, 32]). In contrast, progress on algorithms that tolerate the significantly harsher conditions in the list-decodable setting has been slower. The only prior

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

works [9, 18, 34] 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 [32, 19, 13, 47] 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 [11, 30, 24, 55, 2, 10, 56, 49, 40] 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 [28, 34] and obtaining algorithms for moment estimation [35] and linear regression [32] 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 [28, 34] and obtaining algorithms for moment estimation [35] and linear regression [32] 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 [34] 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 [36] who developed such a method to give a simpler proof of the list-decodable mean estimation result of [34]. In Section 2, 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 [41].

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] There's a long line of work on robust regression algorithms (see for e.g. [8, 31]) that can tolerate corruptions only in the *labels*. We are interested in algorithms robust against corruptions in both examples and labels.

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 $\|\ell^*\|_2 \leq 1$, 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 1.6 and 4.4) 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 [20, 51, 48] 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 $\Pr[\langle Y, v\rangle = 0] < \delta$.

In Proposition 2.4, we provide a simple but conceptually illuminating proof that anti-concentration is *sufficient* for list-decodable regression. In Theorem 6.1, 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 2.4 and Theorem 6.1)**.** *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 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. $\forall v$, $\langle Y, v\rangle^2 \leq \delta^2 \mathbb{E}\langle Y, v\rangle^2$ 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 3 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 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 (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 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 [40] 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 [40] (see Page 10) use the relationship to learning mixtures of k gaussians (with an $\exp(k)$ lower bound [44]) 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 5, 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 [35, 34, 32]. 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 2, anti-concentration is information-theoretically *necessary* (see Theorem 6.1) 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 [20] 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 4.5) of such a restricted anti-concentration condition. As a specific Corollary (see Theorem 4.6 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 4.1, 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.

2 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, 42, 35, 28, 34, 32] 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 B.1) 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 6.1) *requires* that the distribution of inliers[2] be sufficiently *anti-concentrated*:

Definition 2.1 (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 2.2. *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 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 2.2 to derive the same conclusion starting from any maximally uniform distribution supported on large soluble sets.

Proposition 2.3. *For a maximally uniform μ on αn size soluble subsets of $\mathcal{S}$, $\sum_{i \in \mathcal{I}} \mathbb{E}_{S \sim \mu}[\mathbf{1}\,(i \in S)] \geq \alpha|\mathcal{I}|$.*

[2] As in the standard robust estimation setting, the outliers are arbitrary and potentially adversarially chosen.

[3] Definition 1.4 differs slightly to handle list-decodable regression with additive noise in the inliers.

The proof proceeds by contradiction (see Lemma 4.3). 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 2.4 (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 2.2 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.

$$\mathcal{A}_{w,\ell}\colon \left\{ \begin{aligned} & & \textstyle\sum_{i=1}^{n} w_i &= \alpha n \\ &\forall i \in [n]. & w_i^2 &= w_i \\ &\forall i \in [n]. & w_i \cdot (y_i - \langle x_i, \ell\rangle) &= 0 \\ & & \|\ell\|^2 &\leq 1 \end{aligned} \right\} \tag{2.1}$$

Our efficient algorithm searches for a maximally uniform *pseudo-distribution* on w satisfying (2.1). Degree k pseudo-distributions (see Section 3 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 (2.1)) 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 2.3 for pseudo-distributions and gives us an efficient replacement for Step 1.

Proposition 2.5. *For any maximally uniform $\tilde{\mu}$ of degree ≥ 2, $\sum_{i\in\mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}}[w_i] \geq \alpha|\mathcal{I}| = \alpha \sum_{i\in[n]} \tilde{\mathbb{E}}_{\tilde{\mu}}[w_i]$.*

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:

[4] See Fact 3.3 for a precise statement.

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

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 (2.1) 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 2.5, $\frac{\sum_{i\in\mathcal{I}}\tilde{\mathbb{E}}[w_i]}{\sum_{i\in[n]}\tilde{\mathbb{E}}[w_i]}\geq\alpha$ 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 (2.1). 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 2.2, 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 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 $\mathbf{1}\left(z=0\right)$. 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 A.1, 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 5.

3 Preliminaries

In this section, we define pseudo-distributions and sum-of-squares proofs. See the lecture notes [6] for more details and the appendix in [42] 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 $p=q_1^2+\cdots+q_r^2$.

3.1 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 $D \geq 0$ 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 $D \geq 0$ 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 $\sum_x D(x) f(x)^2 \geq 0$ 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 $D \geq 0$ 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

$$\tilde{\mathbb{E}}_{D(x)} f(x) = \sum_x D(x) f(x) \ . \tag{3.1}$$

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 3.1 ([50, 46, 45, 38])**.** *For any $n, \ell \in \mathbb{N}$, the following set has a $n^{O(\ell)}$-time weak separation oracle (in the sense of [26]):*

$$\left\{ \tilde{\mathbb{E}}_{D(x)}(1, x_1, x_2, \ldots, x_n)^{\otimes d} \mid \textit{degree-}d \textit{ pseudo-distribution } D \textit{ over } \mathbb{R}^n \right\} \ . \tag{3.2}$$

This fact, together with the equivalence of weak separation and optimization [26] 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 3.2 (Constrained pseudo-distributions)**.** Let D be a level-ℓ pseudo-distribution over $\mathbb{R}^n$. Let $\mathcal{A} = \{f_1 \geq 0, f_2 \geq 0, \ldots, f_m \geq 0\}$ be a system of m polynomial inequality constraints. We say that D *satisfies the system of constraints $\mathcal{A}$ at degree r*, denoted $D \mathrel{\underset{r}{\models}} \mathcal{A}$, if for every $S \subseteq [m]$ and every sum-of-squares polynomial h with $\deg h + \sum_{i \in S} \max\{\deg f_i, r\}$,

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

We write $D \models \mathcal{A}$ (without specifying the degree) if $D \mathrel{\underset{0}{\models}} \mathcal{A}$ holds. Furthermore, we say that $D \mathrel{\underset{r}{\models}} \mathcal{A}$ 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[5] 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 $\{\|x\|^2 \leq M\}$. The following fact is a consequence of Fact 3.1 and [26],

Fact 3.3 (Efficient Optimization over Pseudo-distributions)**.** *There exists an $(n+m)^{O(\ell)}$-time algorithm that, given any explicitly bounded and satisfiable system[6] $\mathcal{A}$ of m polynomial constraints in n variables, outputs a level-ℓ pseudo-distribution that satisfies $\mathcal{A}$ approximately.*

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

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

3.2 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 $\{f_1 \geq 0, \ldots, f_m \geq 0\}$ imply the constraint $\{g \geq 0\}$ consists of polynomials $(p_S)_{S \subseteq [m]}$ such that

$$g = \sum_{S \subseteq [m]} p_S \cdot \Pi_{i \in S} f_i \,. \tag{3.3}$$

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 $\{f_i \geq 0 \mid i \leq r\}$ implies $\{g \geq 0\}$, we write:

$$\{f_i \geq 0 \mid i \leq r\} \vdash_{\ell} \{g \geq 0\} \,. \tag{3.4}$$

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:

$$\frac{\mathcal{A} \vdash_{\ell} \{f \geq 0, g \geq 0\}}{\mathcal{A} \vdash_{\ell} \{f + g \geq 0\}}, \frac{\mathcal{A} \vdash_{\ell} \{f \geq 0\}, \mathcal{A} \vdash_{\ell'} \{g \geq 0\}}{\mathcal{A} \vdash_{\ell + \ell'} \{f \cdot g \geq 0\}} \qquad \text{(addition and multiplication)}$$

$$\frac{\mathcal{A} \vdash_{\ell} \mathcal{B}, \mathcal{B} \vdash_{\ell'} C}{\mathcal{A} \vdash_{\ell \cdot \ell'} C} \qquad \text{(transitivity)}$$

$$\frac{\{F \geq 0\} \vdash_{\ell} \{G \geq 0\}}{\{F(H) \geq 0\} \vdash_{\ell \cdot \deg(H)} \{G(H) \geq 0\}} \,. \qquad \text{(substitution)}$$

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 3.4 (Soundness)**.** *If $D \models_{r} \mathcal{A}$ for a level-ℓ pseudo-distribution D and there exists a sum-of-squares proof $\mathcal{A} \vdash_{r'} \mathcal{B}$, then $D \models_{r \cdot r' + r'} \mathcal{B}$.*

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 $\mathcal{A} \vdash_{r'} B$ (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 3.5 (Completeness)**.** *Suppose $d \geq r' \geq r$ and $\mathcal{A}$ is a collection of polynomial constraints with degree at most r, and $\mathcal{A} \vdash \{\sum_{i=1}^{n} x_i^2 \leq B\}$ for some finite B.*

Let $\{g \geq 0\}$ be a polynomial constraint. If every degree-d pseudo-distribution that satisfies $D \models_{r} \mathcal{A}$ also satisfies $D \models_{r'} \{g \geq 0\}$, then for every $\epsilon > 0$, there is a sum-of-squares proof $\mathcal{A} \vdash_{d} \{g \geq -\epsilon\}$.

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

Fact 3.6 (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 $\tilde{\mu}$, $\tilde{\mathbb{E}}_{\tilde{\mu}}[fg] \leq \sqrt{\tilde{\mathbb{E}}_{\tilde{\mu}}[f^2]}\sqrt{\tilde{\mathbb{E}}_{\tilde{\mu}}[g^2]}$.*

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

Fact 3.7. *For any univariate degree d polynomial $p(x) \geq 0$ for all $x \in \mathbb{R}$, $\vdash^{x}_{d} \{p(x) \geq 0\}$.*

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

Fact 3.8 (Fekete and Markov-Lukács, see [39])**.** *For any univariate degree d polynomial $p(x) \geq 0$ for $x \in [a, b]$, $\{x \geq a, x \leq b\} \mid\!\frac{x}{d} \{p(x) \geq 0\}$.*

4 Algorithm for List-Decodable Robust Regression

In this section, we describe and analyze our algorithm for list-decodable regression and prove our first main result restated here.

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 $n^{O(k)} = (kd)^{O(k^2)}$.*

We will analyze Algorithm 1 to prove Theorem 1.5.

$$\mathcal{A}_{w,\ell}\colon \left\{ \begin{array}{lr} & \sum_{i=1}^{n} w_i = \alpha n \\ \forall i \in [n]. & w_i^2 = w_i \\ \forall i \in [n]. & w_i \cdot (y_i - \langle x_i, \ell \rangle) = 0 \\ & \displaystyle\sum_{i \leq d} \ell_i^2 \leq 1 \end{array} \right\} \tag{4.1}$$

```
Algorithm 1 (List-Decodable Regression).

Given: Sample $\mathcal{S}$ of size $n$ drawn according to $\mathrm{Lin}(\alpha, n, \ell^*)$ with inliers $\mathcal{I}$, $\eta > 0$.

Output: A list $L \subseteq \mathbb{R}^d$ of size $O(1/\alpha)$ such that there exists a $\ell \in L$ satisfying $\|\ell - \ell^*\|_2 < \eta$.

Operation:

    1. Find a degree $O(1/\alpha^4\eta^4)$ pseudo-distribution $\tilde{\mu}$ satisfying $\mathcal{A}_{w,\ell}$ that minimizes
       $\|\tilde{\mathbb{E}}[w]\|_2$.
    2. For each $i \in [n]$ such that $\tilde{\mathbb{E}}_{\tilde{\mu}}[w_i] > 0$, let $v_i = \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[w_i \ell]}{\tilde{\mathbb{E}}_{\tilde{\mu}}[w_i]}$. Otherwise, set $v_i = 0$.
    3. Take $J$ be a random multiset formed by union of $O(1/\alpha)$ independent draws of
       $i \in [n]$ with probability $\frac{\tilde{\mathbb{E}}[w_i]}{\alpha n}$.
    4. Output $L = \{v_i \mid i \in J\}$ where $J \subseteq [n]$.
```

Our analysis follows the discussion in the overview. We start by formally proving ($\star$).

Lemma 4.1. *For any $t \geq k$ and any $\mathcal{S}$ so that $\mathcal{I} \subseteq \mathcal{S}$ is k-certifiably $(C, \alpha^2\eta^2/4C)$-anti-concentrated,*

$$\mathcal{A}_{w,\ell} \mid\!\frac{w,\ell}{t} \left\{ \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}}^{n} w_i \|\ell - \ell^*\|_2^2 \leq \frac{\alpha^2\eta^2}{4} \right\}$$

Proof. We start by observing:

$$\mathcal{A}_{w,\ell} \mid\!\frac{\ell}{2} \|\ell - \ell^*\|_2^2 \leq 2\,.$$

Since $\mathcal{I}$ is $(C, \alpha\eta/2C)$-anti-concentrated, there exists a univariate polynomial p such that $\forall i$:

$$\{w_i \langle x, \ell - \ell^* \rangle = 0\} \mid\!\frac{k}{\ell} \{p(w_i \langle x_i, \ell - \ell^* \rangle) = 1\} \tag{4.2}$$

and

$$\{\|\ell\|^2 \leq 1\} \mid\!\frac{k}{\ell} \left\{ \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p(\langle x_i, \ell - \ell^* \rangle)^2 \leq \frac{\alpha^2\eta^2}{4} \right\} \tag{4.3}$$

Using (4.2), we have:

$$\mathcal{A}_{w,\ell} \mathrel{\Big|\!\frac{w,\ell}{t+2}} \left\{1 - p^2(w_i\langle x_i, \ell - \ell^*\rangle) = 0\right\} \mathrel{\Big|\!\frac{w,\ell}{t+2}} \left\{1 - w_i p^2(\langle x_i, \ell - \ell^*\rangle) = 0\right\}$$

Using (4.3) and $\mathcal{A}_{w,\ell} \mathrel{\Big|\!\frac{2}{w}} \left\{w_i^2 = w_i\right\}$, we thus have:

$$\begin{aligned}\mathcal{A}_{w,\ell} \mathrel{\Big|\!\frac{w,\ell}{t+2}} \Bigg\{ \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i\|\ell-\ell^*\|_2^2 = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i\|\ell-\ell^*\|_2^2 w_i p^2(\langle x_i, \ell-\ell^*\rangle) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i\|\ell-\ell^*\|_2^2 p^2(\langle x_i, \ell-\ell^*\rangle)\\ \leq \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \|\ell-\ell^*\|_2^2 p^2(\langle x_i, \ell-\ell^*\rangle) \leq \frac{\alpha^2\eta^2}{4}\,. \Bigg\}\end{aligned}$$

□

As a consequence of this lemma, we can show that a constant fraction of the v_i for $i \in \mathcal{I}$ constructed in the algorithm are close to ℓ^*.

Lemma 4.2. *For any $\tilde{\mu}$ of degree k satisfying $\mathcal{A}_{w,\ell}$, $\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \tilde{\mathbb{E}}[w_i]\cdot\|v_i - \ell^*\|_2 \leq \frac{\alpha}{2}\eta$.*

Proof. By Lemma 4.1, we have: $\mathcal{A}_{w,\ell} \mathrel{\Big|\!\frac{w,\ell}{k}} \left\{\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}^n w_i\|\ell-\ell^*\|_2^2 \leq \frac{\alpha^2\eta^2}{4}\right\}$.

We also have: $\mathcal{A}_{w,\ell} \mathrel{\Big|\!\frac{w,\ell}{2}} \left\{w_i^2 - w_i = 0\right\}$ for any i. This yields:

$$\mathcal{A}_{w,\ell} \mathrel{\Big|\!\frac{w,\ell}{k}} \left\{\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}^n \|w_i\ell - w_i\ell^*\|_2^2 \leq \frac{\alpha^2\eta^2}{4}\right\}$$

Since $\tilde{\mu}$ satisfies $\mathcal{A}_{w,\ell}$, taking pseudo-expectations yields:$\frac{1}{\mathcal{I}}\sum_{i\in\mathcal{I}} \tilde{\mathbb{E}}\,\|w_i\ell - w_i\ell^*\|_2^2 \leq \frac{\alpha^2\eta^2}{4}$.

By Cauchy-Schwarz for pseudo-distributions (Fact 3.6), we have:

$$\left(\frac{1}{\mathcal{I}}\sum_{i\in\mathcal{I}} \|\tilde{\mathbb{E}}[w_i\ell] - \tilde{\mathbb{E}}[w_i]\ell^*\|_2\right)^2 \leq \frac{1}{\mathcal{I}}\sum_{i\in\mathcal{I}} \|\tilde{\mathbb{E}}[w_i\ell] - \tilde{\mathbb{E}}[w_i]\ell^*\|_2^2 \leq \frac{\alpha^2\eta^2}{4}\,.$$

Using $v_i = \frac{\tilde{\mathbb{E}}[w_i\ell]}{\tilde{\mathbb{E}}[w_i]}$ if $\tilde{\mathbb{E}}[w_i] > 0$ and 0 otherwise, we have: $\frac{1}{\mathcal{I}}\sum_{i\in\mathcal{I},\tilde{\mathbb{E}}[w_i]>0} \tilde{\mathbb{E}}[w_i]\cdot\|v_i - \ell^*\|_2 \leq \frac{\alpha}{2}\eta$.

□

Next, we formally prove that maximally uniform pseudo-distributions satisfy Proposition 2.5.

Lemma 4.3. *For any $\tilde{\mu}$ of degree ≥ 4 satisfying $\mathcal{A}_{w,\ell}$ that minimizes $\|\tilde{\mathbb{E}}[w]\|_2$, $\sum_{i\in\mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}}[w_i] \geq \alpha^2 n$.*

Proof. Let $u = \frac{1}{\alpha n}\tilde{\mathbb{E}}[w]$. Then, u is a non-negative vector satisfying $\sum_{i\sim[n]} u_i = 1$.

Let $\mathsf{wt}(\mathcal{I}) = \sum_{i\in\mathcal{I}} u_i$ and $\mathsf{wt}(\mathcal{O}) = \sum_{i\notin\mathcal{I}} u_i$. Then, $\mathsf{wt}(\mathcal{I}) + \mathsf{wt}(\mathcal{O}) = 1$.

We will show that if $\mathsf{wt}(\mathcal{I}) < \alpha$, then there's a pseudo-distribution $\tilde{\mu}'$ that satisfies $\mathcal{A}_{w,\ell}$ and has a lower value of $\|\tilde{\mathbb{E}}[w]\|_2$. This is enough to complete the proof.

To show this, we will "mix" $\tilde{\mu}$ with another pseudo-distribution satisfying $\mathcal{A}_{w,\ell}$. Let $\tilde{\mu}^*$ be the *actual* distribution supported on single (w,ℓ) - the indicator $\mathbf{1}_{\mathcal{I}}$ and ℓ^*. Thus, $\tilde{\mathbb{E}}_{\tilde{\mu}^*} w_i = 1$ iff $i \in \mathcal{I}$ and 0 otherwise. $\tilde{\mu}^*$ clearly satisfies $\mathcal{A}_{w,\ell}$. Thus, any convex combination (mixture) of $\tilde{\mu}$ and $\tilde{\mu}^*$ also satisfies $\mathcal{A}_{w,\ell}$.

Let $\tilde{\mu}_\lambda = (1-\lambda)\tilde{\mu} + \lambda\tilde{\mu}^*$. We will show that there is a $\lambda > 0$ such that $\|\tilde{\mathbb{E}}_{\tilde{\mu}_\lambda}[w]\|_2 < \|\tilde{\mathbb{E}}[w]\|_2$.

We first lower bound $\|u\|_2^2$ in terms of $\mathsf{wt}(\mathcal{I})$ and $\mathsf{wt}(\mathcal{O})$. Observe that for any fixed values of $\mathsf{wt}(\mathcal{I})$ and $\mathsf{wt}(\mathcal{O})$, the minimum is attained by the vector u that ensures $u_i = \frac{1}{\alpha n}\mathsf{wt}(\mathcal{I})$ for each $i \in \mathcal{I}$ and $u_i = \frac{1}{(1-\alpha)n}\mathsf{wt}(\mathcal{O})$.

$$\text{This gives } \|u\|^2 \geq \left(\frac{\mathsf{wt}(\mathcal{I})}{\alpha n}\right)^2 \alpha n + \left(\frac{1-\mathsf{wt}(\mathcal{I})}{(1-\alpha)n}\right)^2 (1-\alpha)n = \frac{1}{\alpha n} \cdot \left(\mathsf{wt}(\mathcal{I}) + (1-\mathsf{wt}(\mathcal{I}))^2 \left(\frac{\alpha}{1-\alpha}\right)\right).$$

Next, we compute the the ℓ_2 norm of $u' = \frac{1}{\alpha n}\tilde{\mathbb{E}}_{\tilde{\mu}_\lambda} w$ as:

$$\|u'\|_2^2 = (1-\lambda)^2\|u\|^2 + \frac{\lambda^2}{\alpha n} + 2\lambda(1-\lambda)\frac{\mathsf{wt}(\mathcal{I})}{\alpha n}.$$

$$\begin{aligned}\text{Thus, } \|u'\|^2 - \|u\|^2 &= (-2\lambda+\lambda^2)\|u\|^2 + \frac{\lambda^2}{\alpha n} + 2\lambda(1-\lambda)\frac{\mathsf{wt}(\mathcal{I})}{\alpha n} \\ &\leq \frac{-2\lambda+\lambda^2}{\alpha n} \cdot \left(\mathsf{wt}(\mathcal{I})^2 + (1-\mathsf{wt}(\mathcal{I}))^2 \frac{\alpha}{1-\alpha}\right) + \frac{\lambda^2}{\alpha n} + 2\lambda(1-\lambda)\frac{\mathsf{wt}(\mathcal{I})}{\alpha n}\end{aligned}$$

$$\begin{aligned}\text{Rearranging, } \|u\|^2 - \|u'\|^2 &\geq \frac{\lambda}{\alpha n}\left((2-\lambda) \cdot \left(\mathsf{wt}(\mathcal{I})^2 + (1-\mathsf{wt}(\mathcal{I}))^2\left(\frac{\alpha}{1-\alpha}\right)\right) - \lambda - 2(1-\lambda)\mathsf{wt}(\mathcal{I})\right) \\ &\geq \frac{\lambda(2-\lambda)}{\alpha n}\left(\mathsf{wt}(\mathcal{I})^2 + (1-\mathsf{wt}(\mathcal{I}))^2 \frac{\alpha}{1-\alpha} - \mathsf{wt}(\mathcal{I})\right)\end{aligned}$$

Now, whenever $\mathsf{wt}(\mathcal{I}) < \alpha$, $\mathsf{wt}(\mathcal{I})^2 + (1-\mathsf{wt}(\mathcal{I}))^2 \frac{\alpha}{1-\alpha} - \mathsf{wt}(\mathcal{I}) > 0$. Thus, we can choose a small enough $\lambda > 0$ so that $\|u\|^2 - \|u'\|^2 > 0$.

□

Lemma 4.3 and Lemma 4.2 immediately imply the correctness of our algorithm.

Proof of Main Theorem 1.5. First, since D is k-certifiably $(C, \alpha\eta/4C)$-anti-concentrated, Lemma 5.5 implies taking $\geq n = (kd)^{O(k)}$ samples ensures that $\mathcal{I}$ is k-certifiably $(C, \alpha\eta/2C)$-anti-concentrated with probability at least $1 - 1/d$. Let's condition on this event in the following.

Let $\tilde{\mu}$ be a pseudo-distribution of degree t satisfying $\mathcal{A}_{w,\ell}$ and minimizing $\|\tilde{\mathbb{E}}[w]\|_2$. Such a pseudo-distribution exists as can be seen by just taking the distribution with a single-point support w where $w_i = 1$ iff $i \in \mathcal{I}$.

From Lemma 4.2, we have: $\frac{1}{|\mathcal{I}|}\sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}[w_i] \cdot \|v_i - \ell^*\|_2 \leq \frac{\alpha}{2}\eta$. Let $Z = \frac{1}{\alpha n}\sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}[w_i]$. By a rescaling, we obtain:

$$\frac{1}{|\mathcal{I}|}\sum_{i \in \mathcal{I}} \frac{\tilde{\mathbb{E}}[w_i]}{Z} \cdot \|v_i - \ell^*\|_2 \leq \frac{1}{Z}\frac{\alpha}{2}\eta. \tag{4.4}$$

Using Lemma 4.3, $Z \geq \alpha$. Thus,

$$\frac{1}{|\mathcal{I}|}\sum_{i \in \mathcal{I}} \frac{\tilde{\mathbb{E}}[w_i]}{Z} \cdot \|v_i - \ell^*\|_2 \leq \eta/2. \tag{4.5}$$

Let $i \in [n]$ be chosen with probability $\frac{\tilde{\mathbb{E}}[w_i]}{\alpha n}$. Then, $i \in \mathcal{I}$ with probability $Z \geq \alpha$. By Markov's inequality applied to (4.5), with $\frac{1}{2}$ conditioned on $i \in \mathcal{I}$, $\|v_i - \ell^*\|_2 < \eta$. Thus, in total, with probability at least $\alpha/2$, $\|v_i - \ell^*\|_2 \leq \eta$. Thus, the with probability at least 0.99 over the draw of the random set J, the list constructed by the algorithm contains an ℓ such that $\|\ell - \ell^*\|_2 \leq \eta$.

Let us now account for the running time and sample complexity of the algorithm. The sample size for the algorithm is dictated by Lemma 5.5 and is $(kd)^{O(k)}$, which for our choice of p goes as $(kd)^{O(k)}$. A pseudo-distribution satisfying $\mathcal{A}_{w,\ell}$ and minimizing $\|\tilde{\mathbb{E}}[w]\|_2$ can be found in time $n^{O(k)} = (kd)^{O(k^2)}$. The rounding procedure runs in time at most $O(nd)$. □

Remark 4.4 (Tolerating Additive Noise). To tolerate independent additive noise, our algorithm and analysis change minimally. For an additive noise of variance $\zeta^2 \ll \alpha^2\eta^2$ in the inliers, we modify $\mathcal{A}_{w,\ell}$ by replacing the constraint $\forall i,\, w_i \cdot (y_i - \langle x_i, \ell\rangle) = 0$ by $\forall i,\, \pm w_i \cdot (y_i - \langle x_i, \ell\rangle) \leq 4\zeta$. And $\sum_{i=1}^n w_i = \alpha n$ to $\sum_{i=1}^n w_i = (\alpha/2)n$.

This means that instead of searching for a subsample of size αn that has a exact solution ℓ, we search for a subsample of size $\alpha/2n$ where there's a solution ℓ with an additive error of at most 2ζ. With additive noise of variance ζ^2, it is easy to check that there's a subset of $1/2$ fraction of inliers that satisfies this property. Thus, $\mathcal{A}_{w,\ell}$ is feasible.

Our analysis remains exactly the same except for one change in the proof of Lemma 4.1. We start from a distribution that is $(C, \alpha\eta\zeta/100C)$-certifiably anti-concentrated. And instead of inferring that $p(w_i(y_i - \langle x_i, \ell\rangle)) = 1$, we use that whenever $\pm(y_i - \langle x_i, \ell\rangle) \leq 4\zeta$, $p^2((y_i - \langle x_i, \ell\rangle)) \geq 1 - 4\zeta$.

4.1 List-Decodable Regression for Boolean Vectors

In this section, we show algorithms for list-decodable regression when the distribution on the inliers satisfies a weaker anti-concentration condition. This allows us to handle more general inlier distributions including the product distributions on $\{\pm 1\}^d$, $[0,1]^d$ and more generally any product domain. We however require that the unknown linear function be "Boolean", that is, all its coordinates be of equal magnitude.

We start by defining the weaker anti-concentration inequality. Observe that if $v \in \mathbb{R}^d$ satisfies $v_i^3 = \frac{1}{d} v_i$ for every i, then the coordinates of v are in $\{0, \pm\frac{1}{\sqrt{d}}\}$.

Definition 4.5 (Certifiable Anti-Concentration for Boolean Vectors)**.** A $\mathbb{R}^d$ valued random variable Y is *k-certifiably* (C,δ)-anti-concentrated in *Boolean directions* 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: for all $x^2 \leq \delta^2$, $(p(x)-1)^2 \leq \delta^2$ and for all v such that $v_i^3 = \frac{4}{d} v_i$ for all i, $\|v\|^2 \mathbb{E}_Y p(\langle Y, v\rangle)^2 \leq C\delta$.

We can now state the main result of this section.

Theorem 4.6 (List-Decodable Regression in Boolean Directions)**.** *For every α, η, there's a algorithm that takes input a sample generated according to $\mathrm{Lin}_D(\alpha, n, \ell^*)$ in $\mathbb{R}^d$ for D that is k-certifiably $(C, \alpha\eta/10C)$-anti-concentrated in Boolean directions and $\ell^* \in \left\{\pm\frac{1}{\sqrt{d}}\right\}^d$ and outputs a list L of size $O(1/\alpha)$ such that there's an $\ell \in L$ satisfying $\|\ell - \ell^*\| < \eta$ with probability at least 0.99 over the draw of the sample. The algorithm requires a sample of size $n \geq (d/\alpha\eta)^{O(\frac{1}{\alpha^2\eta^2})}$ and runs in time $n^{O(k)} = (d/\alpha\eta)^{O(k^2)}$.*

The only difference in our algorithm and rounding is that instead of the constraint set $\mathcal{A}_{w,\ell}$, we will work with $\mathcal{B}_{w,\ell}$ that has an additional constraint $\ell_i^2 = \frac{1}{d}$ for every i. Our algorithm is exactly the same as Algorithm 1 replacing $\mathcal{A}_{w,\ell}$ by $\mathcal{B}_{w,\ell}$.

$$\mathcal{B}_{w,\ell}\colon \left\{ \begin{aligned} & & \textstyle\sum_{i=1}^n w_i &= \alpha n \\ &\forall i \in [n], & w_i^2 &= w_i \\ &\forall i \in [n], & w_i \cdot (y_i - \langle x_i, \ell\rangle) &= 0 \\ &\forall in \in [d], & \ell_i^2 &= \frac{1}{d} \end{aligned} \right\} \tag{4.6}$$

We will use the following fact in our proof of Theorem 4.6.

Lemma 4.7. *If a, b satisfy $a^2 = b^2 = \frac{2}{d}$, then, $(a-b)^3 = \frac{1}{d}(a-b)$*

Proof. $(a-b)^3 = a^3 - b^3 - 3a^2b + 3ab^2 = \frac{1}{d}(a - b - 3b + 3a) = \frac{4}{d}(a-b)$. □

Proof of Theorem 4.6. The proof remains the same as in the previous section with one additional step. First, we can obtain the analog of Lemma 4.1 with a few quick modifications to the proof. Then, Lemma 4.2 follows from modified Lemma 4.1 as in the previous section. And the proof of Lemma 4.3 remains exactly the same. We can then put the above lemmas together just as in the proof of Theorem 1.5.

We now describe the modifications to obtain the analog of Lemma 4.1. The key additional step in the proof of the analog of Lemma 4.1 which follows immediately from Lemma 4.7.

$$\left\{\forall i\ \ell_i^2 = \frac{1}{d}\right\} \vdash_{4}^{\ell} \left\{(\ell_i - \ell_i^*)^3 = \frac{4}{d}(\ell_i - \ell_i^*)\right\}$$

This allows us to replace the usage of certifiable anti-concentration by certifiable anti-concentration for Boolean vectors and derive:

$$\left\{\forall i\ \ell_i^2 = \frac{2}{d}\right\} \vdash_{4}^{\ell} \left\{\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} p(\langle x_i, \ell - \ell^*\rangle)^2 \le \frac{\alpha^2\eta^2}{4}\right\}$$

The rest of the proof of Lemma 4.1 remains the same.

□

5 Certifiably Anti-Concentrated Distributions

In this section, we prove certifiable anti-concentration inequalities for some basic families of distributions. We first formally state the definition of certified-anti-concentration.

Definition 5.1 (Certifiable Anti-Concentration)**.** A $\mathbb{R}^d$-valued zero-mean random variable Y has a (C,δ)-*anti-concentrated* distribution if $\Pr[|\langle Y, v\rangle| \le \delta\sqrt{\mathbb{E}\langle Y, v\rangle^2}] \le C\delta$.

Y has a k-*certifiably* (C,δ)-anti-concentrated distribution if there is a univariate polynomial p satisfying $p(0) = 1$ such that

1. $\left\{\langle Y, v\rangle^2 \le \delta^2 \mathbb{E}\langle Y, v\rangle^2\right\} \vdash_{k}^{v} \left\{(p(\langle Y, v\rangle) - 1)^2 \le \delta^2\right\}$.
2. $\left\{\|v\|_2^2 \le 1\right\} \vdash_{k}^{v} \left\{\|v\|_2^2 \mathbb{E} p^2(\langle Y, v\rangle) \le C\delta\right\}$.

We will say that such a polynomial p "witnesses the certifiable anti-concentration of Y". We will use the phrases "Y has a certifiably anti-concentrated distribution" and "Y is a certifiably anti-concentrated random variable" interchangeably.

Before proceeding to prove certifiable anti-concentration of some important families of distributions, we observe the invariance of the definition under scaling and shifting.

Lemma 5.2 (Scale invariance)**.** *Let Y be a k-certifiably (C,δ)-anti-concentrated random variable. Then, so is cY for any $c \neq 0$.*

Proof. Let p be the polynomial that witnesses the certifiable anti-concentration of Y. Then, observe that $q(z) = p(z/c)$ satisfies the requirements of the definition for cY. □

Lemma 5.3 (Certified anti-concentration of gaussians)**.** *For every $0.1 > \delta > 0$, there is a $k = O\left(\frac{\log^2(1/\delta)}{\delta^2}\right)$ such that $\mathcal{N}(0, I)$ is k-certifiably $(2, 2\delta)$-anti-concentrated.*

Proof. Lemma A.1 yields that there exists an univariate even polynomial p of degree k as above such that for all v, whenever $|\langle x, v\rangle| \le \delta$, $p(\langle x, v\rangle) \le 2\delta$, and whenever $\|v\|^2 \le 1$, $\mathbb{E}_{x\sim\mathcal{N}(0,I)} p(\langle x, v\rangle)^2 \le 2\delta$. Since p is even, $p(z) = \frac{1}{2}(p(z) + p(-z))$ and thus, any monomial in $p(z)$ with non-zero coefficient must be of even degree. Thus, $p(z) = q(z^2)$ for some polynomial q of degree $k/2$.

The first property above for p implies that whenever $z \in [0, \delta]$, $p(z) \le 2\delta$. By Fact 3.8, we obtain that:

$$\left\{\langle x, v\rangle^2 \le \delta^2\right\} \vdash_{k}^{v} \left\{p(\langle x, v\rangle)^2 \le \delta\right\}$$

Next, observe that for any j, $\mathbb{E}_{x\sim\mathcal{N}(0,I)}\langle x, v\rangle^{2j} = (2j)!! \cdot \|v\|_2^{2j}$. Thus, $\|v\|_2^2 \mathbb{E}_{x\sim\mathcal{N}(0,I)} p^2(\langle x, v\rangle)$ is a univariate polynomial F in $\|v\|_2^2$. The second property above thus implies that $F(\|v\|_2^2) \le C\delta$ whenever $\|v\|_2^2 \le 1$. By another application of Fact 3.8, we obtain:

$$\left\{\|v\|_2^2 \le 1\right\} \vdash_{k}^{v} \left\{\mathbb{E}_{x\sim\mathcal{N}(0,I)} p(\langle x, v\rangle)^2 \le 2\delta\right\}$$

□

We say that Y is a *spherically symmetric* random variable over $\mathbb{R}^d$ if for every orthogonal matrix R, RY has the same distribution as Y. Examples include the standard gaussian random variable and uniform (Haar) distribution on $\mathbb{S}^{d-1}$. Our argument above for the case of standard gaussian extends to any distribution that is spherically symmetric and has sufficiently light tails.

Lemma 5.4 (Certified anti-concentration of spherically symmetric, light-tail distributions)**.** *Suppose Y is a $\mathbb{R}^d$-valued, spherically symmetric random variable such that for any $k \in (0,2)$, for all t and for all v, $\Pr[\langle v, Y\rangle \geq t\sqrt{\mathbb{E}\langle Y, v\rangle^2}] \leq Ce^{-t^{2/k}/C}$ and for all $\eta > 0$, $\Pr_{x \sim D}[|x| < \eta\sigma] \leq C\eta$, for some absolute constant $C > 0$. Then, for $d = O\left(\frac{\log^{(4+k)/(2-k)}(1/\delta)}{\delta^{2/(2-k)}}\right)$, Y is d-certifiably $(10C, \delta)$-anti-concentrated.*

Lemma 5.5 (Certified anti-concentration under sampling)**.** *Let D be k-certifiably (C, δ)-anti-concentrated, subexponential and unit covariance distribution. Let S be a collection of n independent samples from D. Then, for $n \geq \Omega\left((kd\log(d))^{O(k)}\right)$, with probability at least $1 - 1/d$, the uniform distribution on S is $(2C, \delta)$-anti-concentrated.*

Proof. Let p be the degree k polynomial that witnesses the certifiable anti-concentration of D. Let Y be the random variable with distribution D', the uniform distribution on n i.i.d. samples from D. We will show that p also witnesses that k-certifiable $(4C, \delta/2)$-anti-concentration of Y. To this end it is sufficient to take enough samples such that the following holds.

$$\Pr\left(\left|\mathbb{E}_D[p^2(\langle Y, v\rangle)] - \mathbb{E}_{D'}[p^2(\langle Y, v\rangle)]\right| > \mathbb{E}_D[p^2(\langle Y, v\rangle)]/2\right) < 1/d$$

Observe that $p^2(\langle Y, v\rangle)$ may be written as $\langle c(Y)c(Y)^T, m(v)m(v)^T\rangle$ where $c(Y)$ are the coefficients of $p(\langle Y, v\rangle)$ and $m(v)$ is the vector containing monomials. The dot product above is the usual trace inner product between matrices. Now, it is sufficient to show that

$$\Pr\left(\|\mathbb{E}_{D'}c(Y)c(Y)^T - \mathbb{E}_D c(Y)c(Y)^T\|_F^2 > \|\mathbb{E}_D c(Y)c(Y)^T\|_F^2/4\right) < 1/d$$

Since p was a univariate polynomial of degree k in d dimensional variables, there are at most d^{2k} entries in total, and each entry is at most a degree $2k$ polynomial of subexponential random variables in d variables. Using standard concentration results for polynomials of subexponential random variables (for instance Theorem 1.2 from [25] and the references therein). We see that each entry satisfies

$$\Pr\left(|\mathbb{E}_D c(Y)_i c(Y)_j - \mathbb{E}_{D'} c(Y)_i c(Y)_j| > \epsilon\right) \leq \exp\left(-\Omega\left(\frac{n\epsilon}{\mathbb{E}(c(Y)_i c(Y)_j)^2}\right)^{1/2k}\right)$$

An application of a union bound, squaring the term inside and replacing ϵ^2 by $\mathbb{E}(c(Y)_i c(Y)_j)^2/4$ gives us

$$\Pr\left(\sum_{i,j=1}^{d^{2k}} \left(\mathbb{E}_D c(Y)_i c(Y)_j - \mathbb{E}_{D'} c(Y)_i c(Y)_j\right)^2 > \|\mathbb{E}c(Y)c(Y)^T\|_F^2/4\right) \leq d^{2k}\exp\left(-\Omega\left(\frac{n}{d^{O(k)}}\right)^{1/2k}\right)$$

Hence, setting $n = O((kd\log(d))^{O(k)})$ ensures that with probability at least $1 - 1/d$, the distribution D' is $(2C, \delta)$-anti-concentrated.

□

We say that a $d \times d$ matrix A is C'-well-conditioned if all singular values of A are within a factor of C' of each other.

Lemma 5.6 (Certified anti-concentration under linear transformations)**.** *Let Y be k-certifiably (C, δ)-anti-concentrated random variable over $\mathbb{R}^d$. Let A be any C'-well-conditioned linear transformation. Then, AY is k-certifiably $(C, C'^2\delta)$-anti-concentrated.*

Proof. Let $\|A\|$ be the largest singular value of A. Let p be a polynomial that witnesses the certifiable anti-concentration of Y. Let $q(z) = p(z/\|A\|)$. We will prove that q witnesses the k-certifiable $(C, C'^2\delta)$-anti-concentration of AY.

Towards this, observe that:

$$\left\{\langle Y,v\rangle^2 \leq \delta^2 \mathbb{E}\langle Y,v\rangle^2\right\} \Big|\frac{v}{2} \left\{\langle AY,v\rangle^2 \leq \delta^2 \mathbb{E}\langle AY,v\rangle^2\right\}\,.$$

$$\left\{\langle Y,(A^T v)/\|A\|\rangle^2 \leq \delta^2 \mathbb{E}\langle Y,(A^T v)/\|A\|\rangle^2\right\} \Big|\frac{v}{k} \left\{(p(\langle Y,(A^T v)/\|A\|\rangle)-1)^2 \leq \delta^2\right\},$$

this is the same as

$$\left\{\langle AY,v\rangle^2 \leq \delta^2 \mathbb{E}\langle AY,v\rangle^2\right\} \Big|\frac{v}{k} \left\{(q(\langle AY,v\rangle)-1)^2 \leq \delta^2\right\}.$$

Where $q = p(x/\|A\|)$. Now, for $w = (A^T v)/\|A\|$ and any unit vector v,

$$\left\{\|w\|_2^2 \leq 1\right\} \Big|\frac{v}{k} \left\{\|A^T v\|_2^2/\|A\|_2^2 \mathbb{E}p^2(\langle AY,v\rangle/\|A\|) \leq C\delta\right\}\,,$$

Thus,

$$\left\{\|A^T v\|_2^2 \leq \|A\|^2\right\} \Big|\frac{v}{k} \left\{\|A^T v\|_2^2 \mathbb{E}q^2(\langle AY,v\rangle) \leq C\|A\|_2^2\delta\right\}\,.$$

However,

$$\left\{\|v\|_2^2 \leq 1\right\} \Big|\frac{v}{2} \left\{\|A^T v\|_2^2 \leq \|A\|^2\right\}\,,$$

and thus,

$$\left\{\|v\|_2^2 \leq 1\right\} \Big|\frac{v}{k} \left\{\|v\|_2^2 \mathbb{E}q^2(\langle AY,v\rangle) \leq CC'^2\delta\right\}\,.$$

□

Lemma 5.7 (Certifiable Anti-Concentration in Boolean Directions)**.** *Fix $C > 0$. Let Y be a $\mathbb{R}^d$ valued* product *random variable satisfying:*

1. ***Identical Coordinates**: Y_i are identically distributed for every $1 \leq i \leq d$.*
2. ***Anti-Concentration** For every $v \in \left\{0, \pm\frac{1}{\sqrt{d}}\right\}^d$, $\Pr[|\langle Y,v\rangle| \leq \delta\sqrt{\mathbb{E}\langle Y,v\rangle^2}] \leq C\delta$.*
3. ***Light tails** For every $v \in \mathbb{S}^{d-1}$, $\Pr[|\langle Y,v\rangle| > t\sqrt{\mathbb{E}\langle Y,v\rangle^2}] \leq \exp(-t^2/C)$.*

Then, Y is k-certifiably (C,δ)-anti-concentrated for $k = O\left(\frac{\log^2(1/\delta)}{\delta^2}\right)$.

Proof. We use the p from Lemma A.1. To see that p witnesses the anti-concentration of Y, once again observe that Lemma A.1 applies to give us a real life proof of the required statements. We now exhibit a sum of squares proof. Observe that every monomial of even degree $2k$ for any $k \in \mathbb{N}$, $\mathbb{E}_{Y\sim D}\langle Y,v\rangle^{2k}$ is a *symmetric* polynomial in v with non-zero coefficients only on even-degree monomials in v. This follows by noting that the coordinates of D are independent and identically distributed and x^2 is an even function. It is a fact that all symmetric polynomials in v can be expressed as polynomials in the "power-sum" polynomials $\|v\|_{2i}^{2i}$ for $i \leq 2t$. However, since $v_i^2 \in \left\{0, \frac{1}{d}\right\}$ for $i \geq 1$, $\|v\|_{2i}^{2i} = \frac{1}{d^{i-1}}\|v\|_2^2$. Hence a polynomial in $\|v\|_{2i}^{2i}$ is also a univariate polynomial in $\|v\|_2^2$. Since these are polynomial inequalities, they are also sum-of-squares proofs of these inequalities.

The observation above implies $\|v\|_2^2 \mathbb{E}_Y p(\langle Y,v\rangle)^2 = \|v\|_2^2 \cdot F(\|v\|_2^2)$ for some degree k univariate polynomial F. Since Since F is a univariate polynomial and $\|v\|_2^2 \leq 1$ is an "interval constraint" by applying Fact 3.8, we get: $\Big|\frac{\|v\|_2^2}{2t} \left\{\|v\|_2^2 F(\|v\|_2^2) \leq C\delta\right\}$. Recalling the fact that $\|v\|_2^2 \mathbb{E}_Y p(\langle Y,v\rangle)^2 = \|v\|_2^2 \cdot F(\|v\|_2^2)$, this completes the proof. □

6 Information-Theoretic Lower Bounds for List-Decodable Regression

In this section, we show that list-decodable regression on $\mathrm{Lin}_D(\alpha,\ell^*)$ information-theoretically requires that D satisfy α-anti-concentration: $\Pr_{x\sim D}[\langle x,v\rangle = 0] < \alpha$ for any non-zero v.

Theorem 6.1 (Main Lower Bound)**.** *For every q, there is a distribution D on $\mathbb{R}^d$ satisfying $\Pr_{x\sim D}[\langle x,v\rangle = 0] \leq \frac{1}{q}$ such that there's no $\frac{1}{2q}$-approximate list-decodable regression algorithm for $\mathrm{Lin}_D(\frac{1}{q},\ell^*)$ that can output a list of size $< d$.*

Remark 6.2 (Impossibility of Mixed Linear Regression on the Hypercube). Our construction for the case of $q = 2$ actually shows the impossibility of the well-studied and potentially easier problem of noiseless *mixed linear regression* on the uniform distribution on $\{0,1\}^n$. This is because $\mathcal{R}_i$ is, by construction, obtained by using one of e_i or $\mathbf{1} - e_i$ to label each example point with equal probability.

Theorem 6.1 is tight in a precise way. In Proposition 2.4, we proved that whenever D satisfies $\Pr_{x\sim D}[\langle x, v\rangle = 0] < \frac{1}{q}$, there is an (inefficient) algorithm for *exact* list-decodable regression algorithm for $\mathrm{Lin}_D(\frac{1}{q}, \ell^*)$. Note that our lower bound holds even in the setting where there is no additive noise in the inliers.

Somewhat surprisingly, our lower bound holds for extremely natural and well-studied distributions - uniform distribution on $\{0,1\}^n$ and more generally, uniform distribution on $\{0,1,\ldots,q-1\}^d = [q]^d$ for any q. We can easily determine a tight bound on the anti-concentration of both these distributions.

Lemma 6.3. *For any non-zero $v \in \mathbb{R}^d$, $\Pr_{x\sim\{0,1\}^n}\langle x, v\rangle = 0 \leq \frac{1}{2}$ and $\Pr_{x\sim[q]^d}[\langle x, v\rangle = 0] \leq \frac{1}{q}$.*

Note that this is tight for any $v = e_i$, the vector with 1 in the ith coordinates and 0s in all others.

Proof. Fix any v. Without loss of generality, assume that all coordinates of v are non-zero. If not, we can simply work with the uniform distribution on the sub-hypercube corresponding to the non-zero coordinates of v.

Let $S \subseteq \{0,1\}^n$ ($[q]^d$, respectively) be the set of all $x \in \{0,1\}^n$ ($[q]^d$, respectively) such that $\langle x, v\rangle = 0$. Then, observe that for any $x \in S$, and any i, $x^{(i)}$ obtained by flipping the ith bit (changing the ith coordinate to any other value) of x cannot be in S. Thus, S is an independent set in the graph on $\{0,1\}^n$ (in $[q]^d$, respectively) with edges between pairs of points with hamming distance 1.

It is a standard fact [54] that the maximum independent set in the d-hypercube is of size exactly 2^{d-1} and in the q-ary Hamming graph $[q]^d$ is of size q^{d-1}. Thus, $\Pr_{x\sim\{0,1\}^d}[\langle x, v\rangle = 0] \leq \frac{1}{2}$ and $\Pr_{x\sim[q]^d}[\langle x, v\rangle = 0] \leq \frac{1}{q}$.

□

To prove our lower bound, we give a family of d distributions on labeled linear equations, $\mathcal{R}_i$ for $1 \leq i \leq d$ that satisfy the following:

1. The examples in each are chosen from uniform distribution on $[q]^d$,
2. $\frac{1}{q}$ fraction of the samples are labeled by e_i in $\mathcal{R}_i$, and,
3. for any i, j, $\mathcal{R}_i$ and $\mathcal{R}_j$ are statistically indistinguishable.

Thus, given samples from $\mathcal{R}_i$, any $\frac{1}{2q}$-approximate list-decoding algorithm must produce a list of size at least d.

Our construction and analysis of $\mathcal{R}_i$ is simple and exactly the same in both the cases. However it is somewhat easier to understand for the case of the hypercube ($q = 2$). The following simple observation is the key to our construction.

Lemma 6.4. *For $1 \leq i \leq d$, let $\mathcal{R}_i$ be the distribution on linear equations induced by the following sampling method: Sample $x \sim \{0,1\}^d$, choose $a \sim \{0,1\}$ uniformly at random and output: $(x, \langle x, (1-a)e_i\rangle)$. Then, $\mathcal{R}_i = \mathcal{R}_j$ for any $i, j \leq d$.*

Proof. The proof follows by observing that $\mathcal{R}_i$ when viewed as a distribution on $\mathbb{R}^{d+1}$ is same as the uniform distribution on $\{0,1\}^{d+1}$ and thus independent of i. □

The argument immediately generalizes to $[q]^d$ and yields:

Lemma 6.5. *For $1 \leq i \leq d$, let $\mathcal{R}_i$ be the distribution on linear equations induced by the following sampling method: Sample $x \sim [q]^d$, choose $a \sim \{0,1\}$ uniformly at random and output: $(x, (\langle x, e_i\rangle + a)$ mod $q)$. Then, $\mathcal{R}_i = \mathcal{R}_j$ for any $i, j \leq d$.*

In this case, we interpret the $1/q$ fraction of the samples where $a = 0$ as the inliers. Observe that these are labeled by a single linear function e_i in any $\mathcal{R}_i$. Thus, they form a valid model in $\mathrm{Lin}_D(\alpha, \ell^*)$ for $\alpha = 1/q$.

Since the linear functions defined by e_i on $[q]^d$, when normalized to have unit norm, have a pairwise Euclidean distance of at least $1/q$, we immediately obtain a proof of Theorem 6.1.

A Polynomial Approximation for Core-Indicator

The main result of this section is a low-degree polynomial approximator for the function $\mathbf{1}(|x| < \delta)$ with respect to all distributions that have asymptotically lighter-than-exponential tails.

Lemma A.1. *Let D be a distribution on $\mathbb{R}$ with mean 0, variance $\sigma^2 \leq 1$ and satisfying:*

1. ***Anti-Concentration:*** *For all $\eta > 0$, $\Pr_{x \sim D}[|x| < \eta\sigma] \leq C\eta$, and,*
2. ***Tail bound:*** $\Pr[|x| \geq t\sigma] \leq e^{-\frac{t^{2/k}}{C}}$ *for $k < 2$ and all t,*

for some $C > 1$. Then, for any $\delta > 0$, there is a $d = O\left(\frac{\log^{(4+k)/(2-k)}(1/\delta)}{\delta^{2/(2-k)}}\right) = \tilde{O}\left(\frac{1}{\delta^{2/(2-k)}}\right)$ and an even polynomial $q(x)$ of degree d such that $q(0) = 1$, $q(x) = 1 \pm \delta$ for all $|x| \leq \delta$ and $\sigma^2 \cdot \mathbb{E}_{x \sim D}\left[q^2(x)\right] \leq 10C\delta$.

Before proceeding to the proof, we note that the bounds on the degree above are tight up to poly logarithmic factors for the gaussian distribution.

Lemma A.2. *For every polynomial p of degree d such that $p(0) = 1$, $\mathbb{E}_{x \sim \mathcal{N}(0,1)}[p^2(x)] = \Omega\left(\frac{1}{\sqrt{d}}\right)$. Further, there is a polynomial p_* of degree d such that $p_*(0) = 1$ and $\mathbb{E}_{x \sim \mathcal{N}(0,1)} p_*^2(x) = \Theta\left(\frac{1}{\sqrt{d}}\right)$.*

Our construction of the polynomial is based on standard techniques in approximation theory for constructing polynomial approximators for continuous functions over an interval. Most relevant for us are various works of Eremenko and Yuditskii [22, 23, 21] and Diakonikolas, Gopalan, Jaiswal, Servedio and Viola [12] on such constructions for the sign function on the interval $[-1, a] \cup [a, 1]$ for $a > 0$. We point the reader to the excellent survey of this beautiful line of work by Lubinsky [41].

Fact A.3 (Theorem 3.5 in [12]). *Let $0 < \eta < 0.1$, then there exist constants C, c such that for*

$$a := \eta^2/C \log(1/\eta) \text{ and } K = 4c\log(1/\eta)/a + 2 < O(\log^2(1/\eta)/\eta^2)$$

there is a polynomial $p(t)$ of degree K satisfying

1. *$p(t) > \mathrm{sign}(t) > -p(-t)$ for all $t \in \mathbb{R}$.*
2. *$p(t) \in [\mathrm{sign}(t), \mathrm{sign}(t) + \eta]$ for $t \in [-1/2, -2a] \cup [0, 1/2]$.*
3. *$p(t) \in [-1, 1 + \eta]$ for $t \in (-2a, 0)$*
4. *$|p(t)| \leq 2 \cdot (4t)^K$ for all $t > \frac{1}{2}$.*

We will also rely on the following elementary integral estimate.

Lemma A.4 (Tail Integral).

$$\int_{[L,\infty]} \exp\left(-\frac{x^{2/k}}{C}\right) x^{2d} dx < \exp\left(-\frac{L^{2/k}}{C}\right) \left((L)^{4d} + (16kd)^{kd}\right).$$

Proof. We first prove the claim for $k = 1$. Let $y = x - L$. The, $\int_L^\infty e^{-x^2} x^{2d} dx = \int_0^\infty e^{-(y+L)^2} (y + L)^{2d} dy$. We now use that $y^2 + L^2 \leq (y + L)^2$ for all $y \geq 0$ and $(y + L)^{2d} \leq 2^{2d}(y^{2d} + L^{2d})$ to upper bound the integral above by: $e^{-L^2} L^{2d} + 2^{2d} e^{-L^2} \int_0^\infty e^{-y^2} y^{2d}$. Using $\int_0^\infty e^{-y^2} y^{2d} < (4d)^d$ gives a bound of $e^{-L^2}(L^{2d} + (8d)^d)$.

For larger k, we substitute $y = x^{1/k}$ and write the integral in question as $\int_{L^{1/k}}^\infty e^{-y^2} y^{2kd-(k-1)} dy$. Applying the calculation from the above special case, this integral is upper bounded by: $e^{-L^{2/k}}(L^{4d} + (16kd)^{kd})$. □

Proof of Lemma A.1. Let $p(x)$ be the degree $d < O\left(\frac{L\log^2(1/\delta)}{\delta}\right)$ polynomial from Fact A.3. We then construct a polynomial $q(x)$ that will be close to 0 in the range $[\delta, L]$ and $[-L, -\delta]$ and close to 1 in the range $[-\delta, \delta]$. Our polynomial q is obtained by shifting and appropriately scaling two copies of p.

$$q(x) = \frac{p\left(a + \frac{x}{4L}\right) + p\left(-(a + \frac{x}{4L})\right) - 1}{p\left(a\right) + p\left(-a\right) - 1}$$

Then, $q(0) = 1$. It further satisfies:

1. $q(x) \in [0, C\sqrt{\delta/L}]$ for $x \in [\delta, L] \cup [-L, \delta]$.
2. $q(x) \in [1 - C\sqrt{\delta/L}, 1 + \sqrt{\delta/L}]$ for $x \in [-\delta, \delta]$.
3. $q(x) \in [0, 1 + \sqrt{\delta/L}]$ for $x \in [-3\delta, -\delta] \cup [\delta, 3\delta]$.
4. $|q(x)| < 4 \cdot (4x)^t$ for $|x| > L$

We now prove the bound the $\mathbb{E}p^2$. We do this by providing upper bounds on the contributions to $\sigma^2 \cdot \mathbb{E}_{x\sim\mathcal{D}}\left[q^2(\sigma x)\right]$ from the disjoint sets with different guarantees below. Since we are going to evaluate $q(\sigma x)$ the intervals will be scaled by σ.

The contributions from the regions $\frac{1}{\sigma}[\delta, L]$ and $\frac{1}{\sigma}[-\delta, \delta]$ can be naively upper bounded by the maximum value that the polynomial can take here times the probability of landing in these regions. The first of these contributes $\sigma \cdot \frac{\delta}{L} \cdot (L - \delta) \leq \delta$, and using anticoncentration, the second region contributes $\left(1 + \sqrt{\frac{\delta}{L}}\right)^2 \cdot 2C\delta \leq 4C\delta$. The region $\frac{1}{\sigma}[\delta, 3\delta]$ can be bounded similarly to get an upper bound of $2\left(1 + \sqrt{\frac{\delta}{L}}\right)^2 \sigma^2\delta \leq 4\delta$. To finish, we use Lemma A.4 to upper bound the contribution to $\mathbb{E}p^2$ from the tail:

$$\begin{aligned}\sigma^2 C' \int_{\frac{1}{\sigma}[L,\infty]} q^2(\sigma x) \exp\left(-\frac{x^{2/k}}{C}\right) dx &\lesssim \sigma^{2+d} 4^d \exp\left(-\frac{1}{C} \cdot \left(\frac{L}{\sigma}\right)^{2/k}\right) ((L/\sigma)^{4d} + (16kd)^{kd}) \\ &\lesssim \exp\left(2d + 4d\log\left(\frac{L}{\sigma}\right) - \frac{1}{C} \cdot \left(\frac{L}{\sigma}\right)^{2/k} + kd\log(16kd)\right)\end{aligned}$$

We choose L satisfying $10d\log(d) + 4d\log(\frac{L}{\sigma}) - \frac{1}{C} \cdot (\frac{L}{\sigma})^{2/k} < 2\log(1/\delta)$.

Since $d = O\left(\frac{L\log^2(1/\delta)}{\delta}\right)$, $k < 2$, and $\sigma < 1$ we can now choose $L = \left(\frac{C100\log^3(1/\delta)}{\delta}\right)^{k/(2-k)}$ to satisfy the inequality above and to get $d \lesssim \frac{\log^{2+3k/(2-k)}(1/\delta)}{\delta^{1+k/(2-k)}}$. When $k = 1$ we get $d = \tilde{O}(1/\delta^2)$. Since $\sigma < 1$ in all the above calculations, we get our result by re-scaling δ.

□

We now complete the proof of Lemma A.2.

Proof of Lemma A.2. Any polynomial p of degree d can be written as $p(x) = \sum_{i=1}^{d} \alpha_i h_i(x)$ where h_i denote the hermite polynomials of degree i, satisfying $\mathbb{E}_{x\sim\mathcal{N}(0,1)} h_i = 0$ and $\mathbb{E}_{x\sim N(0,1)}[h_i^2(x)] = 1$. Since $p(0) = 1$, using Cauchy-Schwartz inequality, we obtain:

$$\mathbb{E}_{x\sim N(0,1)}[p^2(x)] \cdot \sum_{i=1}^{d} h_i^2(0) = \left(\sum_{i=1}^{d} \alpha_i^2\right) \cdot \left(\sum_{i=1}^{d} h_i^2(0)\right) \geq \left(\sum_{i=1}^{d} \alpha_i h_i(0)\right)^2 \geq 1$$

Further, observe that for the polynomial $p_*(x) = \frac{1}{\sum_i h_i^2(0)} \sum_i h_i(0) h_i(x)$, the above inequality is tight. Using that $h_{2i}(0) = \frac{(2i-1)!!}{\sqrt{(2i)!}}$ and $h_i(0) = 0$ if i is odd, (see, for e.g., [53]), we have:

$$\mathbb{E}_{x\sim N(0,1)}[p^2(x)] \geq \mathbb{E}_{x\sim\mathcal{N}(0,1)} p_*^2(x) = \left(\sum_{i=1}^{d} h_i^2(0)\right)^{-1} = \left(\sum_{i=1}^{d/2} \left(\frac{(2i-1)!!}{\sqrt{(2i)!}}\right)^2\right)^{-1}$$

$$= \left(\sum_{i=1}^{d/2} \frac{(2i)!}{2^{2i} i!^2}\right)^{-1} = \left(\sum_{i=1}^{d/2} \binom{2i}{i} \cdot \frac{1}{2^{2i}}\right)^{-1} = \Theta\left(\sum_{i=1}^{d/2} \frac{1}{\sqrt{i}}\right)^{-1} = \Theta\left(\sqrt{d}\right)^{-1}.$$

□

B Brute-force search can generate a $\exp(d)$ size list

In the following, we write e_i to denote the vector with 1 in the ith coordinate and 0s in all others.

Proposition B.1. *There exists a distribution D on $\mathbb{R}^d$ and a model $\mathrm{Lin}_D(\alpha, \ell^*)$ such that for every $\alpha < 1/2$, with probability at least $1 - 1/d$ over the draw of a n-size sample $\mathcal{S}$ from $\mathrm{Lin}_D(\alpha, \ell^*)$, there exists a collection $\mathrm{Sol} \subseteq \{S \subseteq \mathcal{S} \mid |S| = \alpha n\}$ of size $\exp(d)$ and unit length vectors ℓ_S for every $S \in \mathrm{Sol}$ such that ℓ_S satisfies all equations in S and for every $S \neq S' \in \mathrm{Sol}$, $\|\ell_S - \ell_{S'}\|_2 \geq 0.1$.*

Proof. Let D be the uniform distribution on $e_1, e_2, \ldots, e_d \in \mathbb{R}^d$. Let $\ell^* := \vec{1}/\sqrt{d}$ be the all-ones vector in $\mathbb{R}^d$ scaled by $1/\sqrt{d}$ and let d samples be drawn from the uncorrupted distribution. These give us our inliers, $\mathcal{I} = \{(x_i, y_i)\}_{i=1}^{\alpha n}$. For the outliers, choose the following multiset $\mathcal{O} := 1/\alpha - 1$ copies of $\{(e_i, j) \mid i \in [d], j \in \{\pm 1/\sqrt{d}\}\}$. This is a sample set of size $2d/\alpha$. Any $a \in \{\pm 1/\sqrt{d}\}^d$ is a valid candidate for a solution for this data. This is because for any such a, $\mathcal{I}_a := \{(e_i, a_i) \mid i \in [d]\} \subset S$ satisfies the following

1. $\mathcal{I}_a \subset S$, $|\mathcal{I}_a| = d = \frac{\alpha}{2}|S|$ and
2. for any $(x, y) \in \mathcal{I}_a$, $y = \langle x, a \rangle$.

The Gilbert–Varshamov bound from coding theory now tells us that there are at least $\Omega(\exp(\Omega(d)))$ $\{0, 1\}$ vectors in d dimensions that pairwise have a hamming distance of $0.1 \cdot d$. This transfers to the set $\{\pm 1/\sqrt{d}\}$ to give us that there are $\Omega(\exp(\Omega(d)))$ vectors in $\{\pm 1/\sqrt{d}\}$ that are pairwise 0.1 apart in 2-norm.

□

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. 4

[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. 4

[6] Boaz Barak and David Steurer. Proofs, beliefs, and algorithms through the lens of sum-of-squares, 2016. Lecture notes in preparation, available on `http://sumofsquares.org`. 7

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

[8] 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

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

[10] 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

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

[12] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. Bounded independence fools halfspaces. In *50th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2009, October 25-27, 2009, Atlanta, Georgia, USA*, pages 171–180, 2009. 18

[13] 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

[14] 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. 1

[15] 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

[16] 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

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

[18] 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

[19] 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

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

[21] Alexandre Eremenko and Peter Yuditskii. Uniform approximation of $\operatorname{sgn} x$ by polynomials and entire functions. *J. Anal. Math.*, 101:313–324, 2007. 18

[22] Alexandre Eremenko and Peter Yuditskii. An extremal problem for a class of entire functions. *C. R. Math. Acad. Sci. Paris*, 346(15-16):825–828, 2008. 18

[23] Alexandre Eremenko and Peter Yuditskii. Polynomials of the best uniform approximation to $\mathrm{sgn}(x)$ on two intervals. *J. Anal. Math.*, 114:285–315, 2011. 18

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

[25] Friedrich Götze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for polynomials in α-sub-exponential random variables. *arXiv e-prints*, page arXiv:1903.05964, Mar 2019. 15

[26] M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981. 8

[27] 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

[28] Sam B. Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. 2017. 1, 2, 4

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

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

[31] 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

[32] 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

[33] 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

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

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

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

[37] 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

[38] Jean B. Lasserre. New positive semidefinite relaxations for nonconvex quadratic programs. In *Advances in convex analysis and global optimization (Pythagorion, 2000)*, volume 54 of *Nonconvex Optim. Appl.*, pages 319–331. Kluwer Acad. Publ., Dordrecht, 2001. 8

[39] Monique Laurent. Sums of squares, moment matrices and optimization over polynomials. In *Emerging applications of algebraic geometry*, pages 157–270. Springer, 2009. 10

[40] 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

[41] Doron S Lubinsky. A Survey of Weighted Approximation for Exponential Weights. *arXiv Mathematics e-prints*, page math/0701099, Jan 2007. 2, 18

[42] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *FOCS*, pages 438–446. IEEE Computer Society, 2016. 4, 7

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

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

[45] Yurii Nesterov. Squared functional systems and optimization problems. In *High performance optimization*, volume 33 of *Appl. Optim.*, pages 405–440. Kluwer Acad. Publ., Dordrecht, 2000. 8

[46] Pablo A Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, 2000. 8

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

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

[49] 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

[50] N. Z. Shor. Quadratic optimization problems. *Izv. Akad. Nauk SSSR Tekhn. Kibernet.*, (1):128–139, 222, 1987. 8

[51] 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

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

[53] Eric W. Weisstein. Hermite number from mathworld. `http://mathworld.wolfram.com/HermiteNumber.html`. 20

[54] Wikipedia. Singleton bound. `https://en.wikipedia.org/wiki/Singleton_bound`. 17

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

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