# Efficient Algorithms for Outlier-Robust Regression

**Adam Klivans** KLIVANS@CS.UTEXAS.EDU
*University of Texas at Austin*

**Pravesh K. Kothari** KOTHARI@CS.PRINCETON.EDU
*Princeton University and IAS*

**Raghu Meka** RAGHUM@CS.UCLA.EDU
*University of California, Los Angeles*

**Editors:** Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

## Abstract

We give the first polynomial-time algorithm for performing linear or polynomial regression resilient to adversarial corruptions in both examples and labels.

Given a sufficiently large (polynomial-size) training set drawn i.i.d. from distribution $\mathcal{D}$ and subsequently corrupted on some fraction of points, our algorithm outputs a linear function whose squared error is close to the squared error of the best-fitting linear function with respect to $\mathcal{D}$, assuming that the marginal distribution of $\mathcal{D}$ over the input space is *certifiably hypercontractive*. This natural property is satisfied by many well-studied distributions such as Gaussian, strongly log-concave distributions and, uniform distribution on the hypercube among others. We also give a simple statistical lower bound showing that some distributional assumption is necessary to succeed in this setting.

These results are the first of their kind and were not known to be even information-theoretically possible prior to our work.

Our approach is based on the sum-of-squares (SoS) method and is inspired by the recent applications of the method for parameter recovery problems in unsupervised learning. Our algorithm can be seen as a natural convex relaxation of the following conceptually simple non-convex optimization problem: find a linear function and a large subset of the input corrupted sample such that the least squares loss of the function over the subset is minimized over all possible large subsets.

**Keywords:** sum-of-squares, regression, robust learning

## 1. Introduction

An influential recent line of work has focused on developing *robust* learning algorithms–algorithms that succeed on a data set that has been contaminated with adversarially corrupted outliers. It has led to important achievements such as efficient algorithms for robust clustering and estimation of moments (Lai et al., 2016; Diakonikolas et al., 2016; Charikar et al., 2017; Kothari and Steurer, 2017; Kothari and Steinhardt, 2017a) in

. Extended abstract. Full version appears as [arXiv 1803.03241 v3]

unsupervised learning and efficient learning of halfspaces (Klivans et al., 2009; Diakonikolas et al., 2017) with respect to malicious or "nasty noise" in classification. In this paper, we continue this line of work and give the first efficient algorithms for performing outlier-robust least-squares *regression*. That is, given a training set drawn from distribution $\mathcal{D}$ and arbitrarily corrupting an $\eta$ fraction of its points (by changing both labels and/or locations), our goal is to efficiently find a linear function (or polynomial in the case of polynomial regression) whose least squares loss is competitive with the best fitting linear function for $\mathcal{D}$.

We give simple examples showing that unlike classical regression, achieving any non-trivial guarantee for robust regression is information-theoretically impossible without making assumptions on the distribution $\mathcal{D}$. In this paper, we study the case where the marginal of $\mathcal{D}$ on examples in the well-studied class of *hypercontractive* distributions. Many natural distributions such as Gaussians, strongly log-concave distributions, and product distributions on the hypercube with bounded marginals fall into this category. To complement our algorithmic results, we also show that for the class of hypercontractive distributions, the bounds on the loss of the linear function output by our algorithm is optimal in its dependence on the fraction of corruptions $\eta$ up to multiplicative constants.

### 1.1. Outlier-Robust Regression

We formally define the problem next. In the following, we will use the following notations for brevity: For a distribution $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$ and for a vector $\ell \in \mathbb{R}^d$, let $\mathsf{err}_{\mathcal{D}}(\ell) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\langle \ell, x\rangle - y)^2]$ and let $\mathrm{opt}(\mathcal{D}) = \min_{\ell\in\mathbb{R}^d} \mathsf{err}_{\mathcal{D}}(\ell)$ be the least error achievable.

In the classical least-squares linear regression problem, we are given access to i.i.d. samples from a distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathbb{R}$ and our goal is to find a linear function $\ell : \mathbb{R}^d \to \mathbb{R}$ whose squared-error—$\mathsf{err}_{\mathcal{D}}(\ell)$—is close to the best possible, $\mathrm{opt}(\mathcal{D})$.

In outlier-robust regression, our goal is similar with the added twist that we only get access to a sample from the distribution $\mathcal{D}$ where up to an $\eta$ fraction of the samples have been arbitrarily corrupted.

**Definition 1 ($\eta$-Corrupted Samples)** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \mathbb{R}$. We say that a set $U \subseteq \mathbb{R}^d \times \mathbb{R}$ is an $\eta$-corrupted training set drawn from $\mathcal{D}$ if it is formed in the following fashion: generate a set $X$ of i.i.d samples from $\mathcal{D}$ and arbitrarily modify any $\eta$ fraction to produce $U$.*

Observe that the corruptions can be *adaptive*, that is, they can depend on the original uncorrupted sample $X$ in an arbitrary way as long as $|U \cap X|/|X| \geqslant 1 - \eta$.[1]

Our goal—which we term *outlier-robust regression*—now is as follows: Given access to an $\eta$-corrupted training set $U$ drawn from $\mathcal{D}$, find a linear function $\ell$ whose error $\mathsf{err}_{\mathcal{D}}(\ell)$ under the true distribution $\mathcal{D}$ is small.

1. In unsupervised learning, this has been called the *strong adversary* model of corruptions and is the strongest notion of robustness studied in the context.

### 1.2. Statement of Results

Our main results give outlier-robust least-squares regression algorithms for hypercontractive distributions.

**Definition 2 (4-Hypercontractivity)** *A distribution $D$ on $\mathbb{R}^d$ is $(C, 4)$-hypercontractive if for all $\ell \in \mathbb{R}^d$, $\mathbb{E}_{x\sim D}[\langle x, \ell\rangle^4] \leqslant C^2 \cdot \mathbb{E}_{x\sim D}[\langle x, \ell\rangle^2]^2$.*

*In addition, we say that $D$ is* certifiably *$(C, 4)$-hypercontractive if there is a degree* 4 sum-of-squares proof *of the above inequality.*

Observe that 4-hypercontractivity is invariant under arbitrary affine transformation, and in particular, doesn't depend on the condition number of the covariance of the distribution.

We will elaborate on the notion of *certifiability* later on (once we have the appropriate preliminaries). For the time being, we note that many well-studied distributions including (potentially non-spherical) Gaussians, affine transformations of isotropic strongly log-concave distributions, the uniform distribution on the Boolean hypercube, and more generally, product distributions on bounded domains are known to satisfy this condition with $C$ a fixed constant.

**Theorem 3** *[Informal] Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times [-M, M]$ and let $\mathcal{D}_X$ be its marginal distribution on $\mathbb{R}^d$ which is certifiably $(C, 4)$-hypercontractive. Let $\ell^* = \arg\min_\ell \mathsf{err}_{\mathcal{D}}(\ell)$ have polynomial bit-complexity. Then for all $\varepsilon > 0$ and $\eta < c/C^2$ for a universal constant $c > 0$, there exists an algorithm $\mathcal{A}$ with run-time $\mathrm{poly}(d, 1/\eta, 1/\varepsilon, M)$ that given a polynomial-size $\eta$-corrupted training set $U$, outputs a linear function $\ell$ such that with probability at least $1 - \varepsilon$,*

$$\mathsf{err}_{\mathcal{D}}(\ell) \leqslant (1 + O(\sqrt{\eta})) \cdot \mathrm{opt}(\mathcal{D}) + O(\sqrt{\eta}) \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - \langle \ell^*, x\rangle)^4] + \varepsilon.$$

The above statement assumes that the marginal distribution is (certifiably) hypercontractive with respect to its fourth moments. Our results improve for higher-order certifiably hypercontractive distributions $\mathcal{D}_X$.In the *realizable case* where $(x, y) \sim \mathcal{D}$ satisfies $y = \langle \ell^*, x\rangle$ for some $\ell^*$, the guarantee of Theorem 3 becomes $\mathsf{err}_{\mathcal{D}}(\ell) \leqslant \varepsilon$; in particular, the error approaches zero at a polynomial rate.

We also get analogous results for outlier-robust polynomial regression.

We also give a simple argument to show that the above guarantee is optimal in its dependence on $\eta$ up to the $O(1)$ factors: even for distributions supported on $\mathbb{R}^d \times [-1, 1]$, it is statistically impossible to achieve an error bound of $(1 + o(\sqrt{\eta}))\mathsf{opt} + o(\sqrt{\eta})$ under the same assumptions.

Our result is a outlier-robust analog of *agnostic* regression problem - that is, the *non-realizable* setting. In addition, our guarantees makes no assumption about the condition number of the covariance of $\mathcal{D}_X$ and thus, in particular, holds for $\mathcal{D}_X$ with poorly conditioned covariances. Alternately, we give a similar guarantee for $\ell_1$ regression when the condition number of covariance of $\mathcal{D}_X$ is bounded without any need for hypercontractivity. We show that in the absence of distributional assumptions (such as hypercontractivity) it is statistically impossible to obtain any meaningful bounds on robust regression.

**Application to Learning Boolean Functions under Nasty Noise.** Our work has immediate applications for learning Boolean functions in the *nasty noise* model, where the learner is presented with an $\eta$-corrupted training set that is derived from an uncorrupted training set of the form $(x, f(x))$ with $x$ drawn from $\mathcal{D}$ on $\{0,1\}^n$ and $f$ is an unknown Boolean function. The goal is to output a hypothesis $h$ with $\mathbb{P}_x[h(x) \neq f(x)]$ as small as possible. The nasty noise model is considered the most challenging noise model for classification problems in computational learning theory.

Applying a result due to Kalai et al. (2008) (c.f. Theorem 5) for learning with respect to adversarial *label noise only* (standard agnostic learning) and a generalization of Theorem 3 to higher degree polynomials we obtain the following:

**Corollary 4** *Let $C$ be a class of Boolean functions on $n$ variables such that for every $c \in C$ there exists a (multivariate) polynomial $p$ of degree $d(\varepsilon)$ with $\mathbb{E}_{x\sim D}[(p(x) - c(x))^2] \leqslant \varepsilon$. Assume that $d(\varepsilon)$ is a constant for any $\varepsilon = O(1)$ and that $\mathcal{D}$ is $(C, 4)$ hypercontractive for polynomials of degree $d(\varepsilon^2)$. Then $C$ can be learned in the nasty noise model in time $n^{O(d(\varepsilon^2))}$ via an output hypothesis $h$ such that $\mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq c(x)] \leqslant O(\sqrt{\eta})\,\mathbb{E}_{x\sim D}[(p(x) - c(x))^4] + \varepsilon$.*

One of the main conclusions of work due to Kalai et al. (2008) is that the existence of low-degree polynomial approximators for a concept class $C$ implies learnability for $C$ in the agnostic setting. Corollary 4 shows that existence of low-degree polynomial approximators *and* hypercontractivity of $D$ imply learnability in the harsher nasty noise model.

We note that Corollary 4 gives an incomparable set of results in comparison to recent work of Diakonikolas et al. (2017) for learning polynomial threshold functions in the nasty noise model.

**Concurrent Works.** Using a set of different techniques, Diakonikolas, Kamath, Kane, Li, Steinhardt and Stewart Diakonikolas et al. (2018a) and Prasad, Suggala, Balakrishnan and Ravikumar Prasad et al. (2018) also obtained robust algorithms for regression in the setting where data $(x, y)$ is generated via the process: $y = \langle w, x\rangle + e$ for an fixed unknown vector $w$ and zero mean noise $e$. For improved bounds for the case when $x$ is distributed according to a Gaussian, see recent (independent and concurrent) work due to Diakonikolas, Kong, and Stewart Diakonikolas et al. (2018b).

### 1.3. Our Approach

In this section, we give an outline of Theorem 3. At a high level, our approach resembles several recent works (Ma et al., 2016; Barak and Moitra, 2016; Potechin and Steurer, 2017; Kothari and Steurer, 2017; Hopkins and Li, 2017) starting with the pioneering work of Barak et al. (2015) that use the Sum-of-Squares method for designing efficient algorithms for learning problems. An important conceptual difference, however, is that previous works have focused on *parameter recovery* problems. For such problems, the paradigm involves showing that there's a simple (in the "SoS proof system") proof that a small sample *uniquely*

identifies the underlying hidden parameters (referred to as "identifiability") up to a small error.

In contrast, in our setting, samples do not uniquely determine a good hypothesis as there can be multiple hypotheses (linear functions) that all have low-error on the true distribution. Our approach thus involves establishing that there's a "simple" proof that *any* low-error hypotheses that is inferred from the observed (corrupted) sample has low-error on the true distribution (we call this *certifiability* of a good hypothesis). To output a good solution in our approach (unlike in cases where there are uniqueness results), we have to crucially rely on the convexity (captured in the SoS proof system) of the empirical loss function.

**Part One: *Certifying* that a linear function has low loss.** Let $X$ be an uncorrupted sample from the underlying distribution $\mathcal{D}$ and suppose we are given an $\eta$-corruption $U = \{(u_1, v_1), (u_2, v_2), \ldots, (u_n, v_n)\}$ of $X$. Let $\hat{\mathcal{D}}$[2] be the uniform distribution on $X$. Our goal is to come up with a linear function $\ell$ that has low error on $\hat{\mathcal{D}}$ given access only to $U$. By standard generalization bounds, this will also imply that $\ell$ has low error on $\mathcal{D}$ with high probability.

It is important to observe that even without computational constraints, that is, *information theoretically*, it is unclear why this should at all be possible. To see why, let's consider the following natural strategy: brute-force search over all subsets $T$ of $U$ of size $(1-\eta)|U|$ and perform least-squares regression to obtain linear function $\ell_T$ with empirical loss $\varepsilon_T$. Then, output $\ell_T$ with minimal empirical loss $\varepsilon_T$ over all subsets $T$.

Since some subset $T^*$ of size $(1-\eta)|U|$ will be a proper subset of the uncorrupted sample, the empirical loss of $\ell_{T^*}$ will clearly be small. However, a priori, there's nothing to rule out the existence of another subset $R$ of size $(1-\eta)|U|$ such that the optimal regression hypothesis $\ell_R$ on $R$ has loss smaller than that of $\ell_{T^*}$ while $\ell_R$ has a large error on the $\hat{\mathcal{D}}$.

This leads to the following interesting question on *certifying a good hypothesis*: given a linear function $\ell$ that has small empirical loss with respect to some subset $T$ of $(1-\eta)$ fraction of the corrupted training set $U$, can we *certify* that its *true* loss with respect to $X$ is small?

We can phrase this as a more abstract "robust certification" question: given two distributions $\mathcal{D}_1$ (=uniform distribution on $X$ above) and $\mathcal{D}_2$ (=uniform distribution on $T$ above) on $\mathbb{R}^d \times \mathbb{R}$ that are $\eta$ close in total variation distance, and a linear function $\ell$ that has small error on $\mathcal{D}_2$, when can we certify a good upper bound on the error of $\ell$ on $\mathcal{D}_1$?

Without making any assumptions on $\mathcal{D}_1$, it is not hard to construct examples where we can give no meaningful bound on the error of a good hypothesis $\ell$ on $\mathcal{D}_1$. More excitingly, we show an elementary proof of a "robust certifiability lemma" that proves a statement as above whenever $\mathcal{D}_1$ has *hypercontractive* one dimensional marginals. The loss with respect to $\mathcal{D}_1$ increases as a function of the statistical distance and the degree of hypercontractivity.

2. We use superscript $\hat{}$ to denote empirical quantities and superscript $'$ to denote quantities on corrupted samples.

Applying our certification lemma, it thus suffices to find a subset $T$ of $U$ of size $\geqslant (1-\eta)|U|$ and a linear function $\ell$ such that the least squares error of $\ell$ over $T$ is small.

**Part Two: Inefficient Algorithm via Polynomial Optimization.** Coming back to the question of efficient algorithms, the above approach can appear hopeless in general since simultaneously finding $\ell$ and a subset $T$ of size $(1-\eta)|U|$ that minimizes the error of $\ell$ w.r.t. uniform distribution on $T$ is a non-convex quadratic optimization problem. At a high-level, we will be able to get around this intractability by observing that the *proof* of our robust certifiability lemma is "simple" in a precise technical sense. This simplicity allows us to convert such a certifiability proof into an efficient algorithm in a principled manner. To describe this connection, we will first translate the naive idea for an algorithm above into a polynomial optimization problem.

For concreteness in this high-level description, we suppose that for $(x, y) \sim \mathcal{D}$, the distribution on $x$ is $(C, 4)$-hypercontractive for a fixed constant $C$ and $\mathbb{E}[y^4] = O(1)$. Further, it can also be shown that, with high probability, $\hat{\mathcal{D}}$ is also $(O(1), 4)$-hypercontractive as long as the size of the original uncorrupted sample $X$ is large enough.

Following the certification lemma, our goal is to use $U$ to find a distribution $\mathcal{D}'$ and a linear function $\ell$ such that 1) the loss of $\ell$ with respect to $\mathcal{D}'$ is small and 2) $\mathcal{D}'$ is close to $\widehat{\mathcal{D}}$. It is easy to phrase this as a polynomial optimization problem.

To do so we will look for $X' = \{(x'_1, y'_1), \ldots, (x'_n, y'_n)\}$ and *weights* $w_1, w_2, \ldots, w_n \in \{0, 1\}$ with $\sum_i w_i \geqslant (1-\eta)n$ and $(x'_i, y'_i) = (u_i, v_i)$ if $w_i = 1$. Let $\mathcal{D}'$ be the uniform distribution on $X'$. Clearly, the condition on weights $w$ ensures that the statistical distance between $\hat{\mathcal{D}}, \mathcal{D}'$ is at most $\eta$. Ideally, we intend $w_i$'s to be the indicators of whether or not the $i$'th sample is corrupted. We now try to find $\ell$ that minimizes the least squares error on $\mathcal{D}'$. This can be captured by the following optimization program: $\min_{w,\ell,X'}(1/n)\sum_i (y'_i - \langle \ell, x'_i \rangle)^2$ where $(w, \ell, X')$ satisfy the polynomial system of constraints:

$$\mathcal{P} = \left\{ \begin{array}{ll} \sum_{i=1}^n w_i = (1-\eta)\cdot n & \\ w_i^2 = w_i & \forall i \in [n]. \\ w_i \cdot (u_i - x'_i) = 0 & \forall i \in [n]. \\ w_i \cdot (v_i - y'_i) = 0 & \forall i \in [n]. \end{array} \right\} \tag{1.1}$$

In this notation, our robust certifiability lemma implies that for any $(w, \ell, X')$ satisfying $\mathcal{P}$,

$$\mathsf{err}_{\hat{\mathcal{D}}}(\ell) \leqslant (1 + O(\sqrt{\eta})) \cdot \mathsf{err}_{\mathcal{D}'}(\ell) + O(\sqrt{\eta}). \tag{1.2}$$

It is easy to show that the minimum of the optimization program $\mathrm{opt}(\widehat{\mathcal{D}}) \lessapprox \mathrm{opt}(\mathcal{D})$ (up to standard generalization error) by setting $X' = X$ and $w_i = 1$ if and only if $i$'th sample is uncorrupted. By the above arguments, solutions to the above program satisfy the bound stated in Theorem 3. Unfortunately, this is a quadratic optimization problem and is NP-hard in general.

We are now ready to describe the key idea that allows us to essentially turn this hopelessly inefficient algorithm into an efficient one. This exploits a close relationship between the

simplicity of the proof of robust certifiability and the success of a canonical semi-definite relaxation of (1.1).

**Part Three: From Simple Proofs to Efficient Algorithms.** Suppose that instead of finding a single solution to the program in (1.1), we attempt to find a distribution $\mu$ supported on $(w, \ell, X')$ that satisfy $\mathcal{P}$ and minimizes $\mathbb{E}_\mu[(1/n)\sum_i (y'_i - \langle \ell, x'_i \rangle)^2]$. Let $\mathrm{opt}_\mu$ be the minimum value. Then, as Equation 1.2 holds for all $(w, \ell, X')$ satisfying $\mathcal{P}$, it also follows that

$$\mathop{\mathbb{E}}_{(w,\ell,X')\sim\mu} [\mathsf{err}_{\hat{\mathcal{D}}}(\ell)] \leqslant (1 + O(\sqrt{\eta}))\,\mathrm{opt}_\mu + O(\sqrt{\eta}). \tag{1.3}$$

A priori, we appear to have made our job harder. While computing a distribution on solutions is no easier than computing a single solution, even describing a distribution on solutions appears to require exponential resources in general. However, by utilizing the convexity of the square loss, we can show that having access to just the first moments of $\mu$ is enough to recover a good solution.

Formally, by the convexity of the square loss, the above inequality yields:

$$\mathsf{err}_{\hat{\mathcal{D}}}\left(\mathop{\mathbb{E}}_{\mu}[\ell]\right) \leqslant \mathop{\mathbb{E}}_{(w,\ell,X')\sim\mu} [\mathsf{err}_{\hat{\mathcal{D}}}(\ell)] \leqslant (1 + O(\sqrt{\eta}))\,\mathrm{opt}_\mu + O(\sqrt{\eta}). \tag{1.4}$$

All of the above still doesn't help us in solving program 1.1 as even finding first moments of distributions supported on solutions to a polynomial optimization program is NP-Hard.

The key algorithmic insight is to observe that we can replace distributions $\mu$ by an efficiently computable (via the SoS algorithm) proxy called as *pseudo-distributions* without changing any of the conclusions above.

In what way is a pseudo-distribution a proxy for an actual distribution $\mu$ satisfying $\mathcal{P}$? It turns out that if a polynomial inequality (such as the one in (1.2)) can be derived from $\mathcal{P}$ via a *low-degree sum-of-squares* proof, then (1.3) remains valid even if we replace $\mu$ in (1.3) by a pseudo-distribution $\tilde{\mu}$ of large enough degree. Roughly speaking, the SoS degree of a proof measures the "simplicity" of the proof (in the "SoS proof system"). In other words, facts with simple proofs holds not just for distributions but also for pseudo-distributions.

Thus, the important remaining steps are to show that 1) the inequality (1.2) (which is essentially the conclusion of our robust certifiability lemma) and 2) the convexity argument in (1.4) has a low-degree SoS proof. We establish both these claims by relying on standard tools such as the SoS versions of the Cauchy-Schwarz and Hölder's inequalities.

We give a brief primer to the SoS method in the full version that includes rigorous definitions of concepts appearing in this high-level overview.

### 1.4. Related Work

The literature on grappling with outliers in the context of regression is vast, and we do not attempt a survey here[3]. Many heuristics have been developed modifying the ordinary least

3. Even the term "robust" is very overloaded and can now refer to a variety of different concepts.

squares objective with the intent of minimizing the effect of outliers (see Rousseeuw and Leroy (1987)). Another active line research is concerned with *parameter recovery*, where each label $y$ in the training set is assumed to be from a generative model of the form $\theta^T x + e$ for some (usually independent) noise parameter $e$ and unknown weight vector $\theta \in \mathbb{R}^d$. For example, the recovery properties of LASSO and related algorithms in this context have been intensely studied (see e.g., Xu et al. (2010), Loh and Wainwright (2011)). For more challenging noise models, recent work due to Du, Balakrishnan, and Singh (Du et al., 2017) studies sparse recovery in the Gaussian generative setting in Huber's $\varepsilon$-contamination model, which is similar but formally weaker than the noise model we consider here.

It is common for "robust regression" to refer to a scenario where only the labels are allowed to be corrupted adversarially (for example, see Bhatia et al. (2017) and the references therein), or where the noise obeys some special structure (e.g., Herman and Strohmer (2010)) (although there are some contexts where both the covariates (the $x$'s) and labels may be subject to a small adversarial corruption (Chen et al., 2013)).

What distinguishes our setting is 1) we do not assume the labels come from a generative model; each $(x, y)$ element of the training set is drawn iid from $\mathcal{D}$ and 2) we make no assumptions on the structure or type of noise that can affect a training set (other than that at most an $\eta$ fraction of points may be affected). In contrast to the parameter recovery setting, our goal is similar to that of *agnostic learning*: we will output a linear function whose squared error with respect to $\mathcal{D}$ is close to optimal.

From a technical standpoint, as discussed before our work follows the recent paradigm of converting certifiability proofs to algorithms. Previous applications in machine learning have focused on various parameter-recovery problems in unsupervised learnings. Our work is most closely related to the recent works on robust unsupervised learning (moment estimation and clustering) (Kothari and Steurer, 2017; Hopkins and Li, 2017; Kothari and Steinhardt, 2017b).

## 2. Preliminaries and Notation

### 2.1. Notation

We will use the following notations and conventions throughout: For a distribution $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$ and function $f : \mathbb{R}^d \to \mathbb{R}$, we define $\mathsf{err}_{\mathcal{D}}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2]$. For a vector $\ell \in \mathbb{R}^d$, we abuse notation and write $\mathsf{err}_{\mathcal{D}}(\ell)$ for $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(\langle \ell, x\rangle - y)^2]$. For a real-valued random variable $X$, and integer $k \geqslant 0$, we let $\|X\|_k = \mathbb{E}[X^k]^{1/k}$.

### 2.2. Distribution Families

Our algorithmic results for a wide class of distributions that include Gaussian distributions and others such as log-concave and other product distributions. We next define the properties we need for the marginal distribution on examples to satisfy.

**Definition 5 (Certifiable hypercontractivity)** *For a function $C : [k] \to \mathbb{R}_+$, we say a distribution $D$ on $\mathbb{R}^d$ is $k$-certifiably $C$-hypercontractive if for every $r \leqslant k/2$, there's a degree $k$ sum of squares proof of the following inequality in variable $v$:*

$$\mathop{\mathbb{E}}_{D} \langle x, v\rangle^{2r} \leqslant \left(C(r) \mathop{\mathbb{E}}_{D} \langle x, v\rangle^2\right)^r .$$

Many natural distribution families satisfy certifiable hypercontractivity with reasonably growing functions $C$. For instance, Gaussian distributions, uniform distribution on Boolean hypercube satisfy the definitions with $C(r) = cr$ for a fixed constant $c$. More generally, all distributions that are affine transformations of isotropic distributions satisfying the Poincaré inequality (Kothari and Steinhardt, 2017a), are also certifiably hypercontractive. In particular, this includes all strongly log-concave distributions. Certifiable hypercontractivity also satisfies natural closure properties under simple operations such as affine transformations, taking bounded weight mixtures and taking products. We refer the reader to Kothari and Steurer (2017) for a more detailed overview where certifiable hypercontractivity is referred to as certifiable subgaussianity.

## References


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.

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.

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. URL http://papers.nips.cc/paper/6806-consistent-robust-regression.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 47–60, 2017. doi: 10.1145/3055399.3055491. URL http://doi.acm.org/10.1145/3055399.3055491.

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *ICML*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 774–782. JMLR.org, 2013. URL http://jmlr.org/proceedings/papers/v28/.

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.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. *CoRR*, abs/1707.01242, 2017. URL http://arxiv.org/abs/1707.01242.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *Preprint*, 2018a. URL https://arxiv.org/abs/1803.02815.

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. *Preprint*, 2018b. URL https://arxiv.org/abs/1806.00040.

Simon S. Du, Sivaraman Balakrishnan, and Aarti Singh. Computationally efficient robust estimation of sparse functionals. *CoRR*, abs/1702.07709, 2017. URL http://arxiv.org/abs/1702.07709.

Matthew A. Herman and Thomas Strohmer. General deviants: An analysis of perturbations in compressed sensing. *J. Sel. Topics Signal Processing*, 4(2):342–349, 2010.

Sam B. Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. 2017.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM J. Comput.*, 37(6):1777–1805, 2008. doi: 10.1137/060649057. URL https://doi.org/10.1137/060649057.

Adam R. Klivans, Philip M. Long, and Rocco A. Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10:2715–2740, 2009. doi: 10.1145/1577069.1755877. URL http://doi.acm.org/10.1145/1577069.1755877.

Pravesh K. Kothari and Jacob Steinhardt. Better agnostic clustering via relaxed tensor norms. *CoRR*, abs/1711.07465, 2017a. URL http://arxiv.org/abs/1711.07465.

Pravesh K. Kothari and Jacob Steinhardt. Better agnostic clustering via relaxed tensor norms. 2017b.

Pravesh K. Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. *CoRR*, abs/1711.11581, 2017. URL http://arxiv.org/abs/1711.11581.

Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. *CoRR*, abs/1604.06968, 2016.

Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *CoRR*, abs/1109.3714, 2011. URL http://arxiv.org/abs/1109.3714.

Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. *CoRR*, abs/1610.01980, 2016.

Aaron Potechin and David Steurer. Exact tensor completion with sum-of-squares. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 1619–1673, 2017. URL http://proceedings.mlr.press/v65/potechin17a.html.

A. Prasad, A. Sai Suggala, S. Balakrishnan, and P. Ravikumar. Robust Estimation via Robust Gradient Estimation. *ArXiv e-prints*, 2018.

Peter J. Rousseeuw and Annick M. Leroy. *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1987. ISBN 0-471-85233-3. URL https://doi.org/10.1002/0471725382.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. *IEEE Trans. Information Theory*, 56(7):3561–3574, 2010. doi: 10.1109/TIT.2010.2048503. URL https://doi.org/10.1109/TIT.2010.2048503.