Robust Linear Regression: Optimal Rates in Polynomial Time

Ainesh Bakshi*
abakshi@cs.cmu.edu
CMU

Adarsh Prasad
adarshp@andrew.cmu.edu
CMU

Abstract

We obtain robust and computationally efficient estimators for learning several linear models that achieve statistically optimal convergence rate under minimal distributional assumptions. Concretely, we assume our data is drawn from a k-hypercontractive distribution and an ϵ-fraction is adversarially corrupted. We then describe an estimator that converges to the optimal least-squares minimizer for the true distribution at a rate proportional to $\epsilon^{2-2/k}$, when the noise is independent of the covariates. We note that no such estimator was known prior to our work, even with access to unbounded computation. The rate we achieve is information-theoretically optimal and thus we resolve the main open question in Klivans, Kothari and Meka [COLT'18].

Our key insight is to identify an analytic condition that serves as a polynomial relaxation of independence of random variables. In particular, we show that when the moments of the noise and covariates are negatively-correlated, we obtain the same rate as independent noise. Further, when the condition is not satisfied, we obtain a rate proportional to $\epsilon^{2-4/k}$, and again match the information-theoretic lower bound. Our central technical contribution is to algorithmically exploit independence of random variables in the "sum-of-squares" framework by formulating it as the aforementioned polynomial inequality.

*AB would like to thank the partial support from the Office of Naval Research (ONR) grant N00014-18-1-2562, and the National Science Foundation (NSF) under Grant No. CCF-1815840.

Contents

1 Introduction

While classical statistical theory has focused on designing statistical estimators assuming access to i.i.d. samples from a nice distribution, estimation in the presence of adversarial outliers has been a challenging problem since it was formalized by Huber [Hub64]. A long and influential line of work in high-dimensional robust statistics has since focused on studying the trade-off between sample complexity, accuracy and more recently, computational complexity for basic tasks such as estimating mean, covariance [LRV16, DKK+16, CSV17, KS17b, SCV17, CDG19, DKK+17, DKK+18a, CDGW19], moment tensors of distributions [KS17b] and regression [DKS17, KKM18b, DKK+19, PSBR20, KKK19, RY20a].

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

Seminal works on robust regression focused on designing non-convex loss functions, including M-estimators [Hub11], Theil-Sen estimators[The92, Sen68], R-estimators[Jae72], Least-Median-Squares [Rou84] and S-estimators[RY84]. These estimators have desirable statistical properties under disparate assumptions, yet remain computationally intractable in high dimensions. Further, recent works show that it is information-theoretically impossible to design robust estimators for linear regression without distributional assumptions [KKM18b].

An influential recent line of work showed that when the data is drawn from the well studied and highly general class of *hypercontractive* distributions (see Definition 1.3), there exist robust and computationally efficient estimators for regression [KKM18b, PSBR20, DKS19]. Several families of natural distributions fall into this category, including Gaussians, strongly log-concave distributions and product distributions on the hypercube. However, both estimators converge to the the true hyperplane (in ℓ_2-norm) at a sub-optimal rate, as a function of the fraction of corrupted points.

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

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

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

We note that, despite decades of progress, prior to our work, estimators achieving optimal convergence rate in terms of the fraction of corrupted points were not known, even with independent noise and access to unbounded computation.

1.1 Our Results

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

In contrast to the classical setting, we work in the *strong contamination model*. Here, an adversary has access to the input samples and is allowed to corrupt an ϵ-fraction arbitrarily. Note, the adversary has access to unbounded computation and has knowledge of the estimators we design. We note that this is the most stringent corrupt model and captures Huber contamination, additive corruption, label noise, agnostic learning etc (see [DK19]). Formally,

Model 1.1 (Robust Regression Model)**.** *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d \times R$ such that the marginal distribution over $\mathbb{R}^d$ is centered and has covariance Σ^* and let $\Theta^* = \arg\min_\Theta \mathbb{E}_{x,y\sim\mathcal{D}}\left[(y - \langle\Theta, x\rangle)^2\right]$ be the optimal hyperplane for $\mathcal{D}$. Let $\{(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots (x_n^*, y_n^*)\}$ be n i.i.d. random variables drawn from $\mathcal{D}$. Given $\epsilon > 0$, the robust regression model $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$ outputs a set of n samples $\{(x_1, y_1), \ldots (x_n, y_n)\}$ such that for at least $(1-\epsilon)n$ points $x_i = x_i^*$ and $y_i = y_i^*$. The remaining ϵn points are arbitrary, and potentially adversarial w.r.t. the input and estimator.*

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

Theorem 1.2 (Bounded Covariance does not suffice, Theorem 7.1 informal)**.** *For all $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathbb{R}^d \times \mathbb{R}$ such that the marginal distribution over the covariates has bounded covariance, denoted by $\Sigma^2 = \Theta(1)$, yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega(1)$, where Θ_1 and Θ_2 are the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$.*

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

Definition 1.3 ((C,k)-Hypercontractivity)**.** A distribution $\mathcal{D}$ over $\mathbb{R}^d$ is (C,k)-hypercontractive for an even integer $k \geqslant 4$, if for all $r \in [k/2]$, for all $v \in \mathbb{R}^d$,

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

Remark 1.4. Hypercontractivity captures a broad class of distributions, including Gaussian distributions, uniform distributions over the hypercube and sphere, affine transformations of isotropic distributions satisfying Poincare inequalities [KS17a] and strongly log-concave distributions. Further, hypercontractivity is preserved under natural closure properties like affine transformations, products and weighted mixtures [KS17c]. Further, efficiently computable estimators appearing in this work require *certifiable*-hypercontractivity (Definition 3.5), a strengthening that continues to capture aforementioned distribution classes.

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

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

Definition 1.5 (Negatively Correlated Moments)**.** Given a distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathbb{R}$, such that the marginal distribution on $\mathbb{R}^d$ is (c_k, k)-hypercontractive, the corresponding regression instance has negatively correlated moments if for all $r \leqslant k$, and for all v,

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

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

Remark 1.6. We show that when this condition is satisfied by the true distribution, $\mathcal{D}$, we obtain rates that match the information theoretically optimal rate in a *true linear model*, where the noise (marginal distribution over $y - x^\top \Theta^*$) is independent of the covariates (marginal distribution over x). Further, when this condition is not satisfied, we show that there exist distributions for which obtaining rates matching the *true linear model* is impossible.

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

Theorem 1.7 (Robust Regresssion with Negatively Correlated Noise, Theorem 5.1 informal)**.** *Given $\epsilon > 0, k \geqslant 4$, and $n \geqslant (d\log(d))^{\mathcal{O}(k)}$ samples from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$, such that $\mathcal{D}$ is (c,k)-certifiably hypercontractive and has negatively correlated moments, there exists an algorithm that runs in $n^{\mathcal{O}(k)}$ time and outputs an estimator $\tilde{\Theta}$ such that with high probability,*

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

and,

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

Remark 1.8. We note that prior work does not draw a distinction between the independent and dependent noise models. In comparison (see Table 1), Klivans, Kothari and Meka [KKM18b] obtained a sub-optimal least-squares error scales proportional to $\epsilon^{1-2/k}$. For the special case of $k = 4$, Prasad et. al. [PSBR20] obtain least squares error proportional to $O(\epsilon\kappa^2(\Sigma))$, where κ is the condition number. In very recent independent work Zhu, Jiao and Steinhardt [ZJS20] obtained a sub-optimal least-squares error scales proportional to $\epsilon^{2-4/k}$.

Further, we show that the rate we obtained in Theorem 1.7 is information-theoretically optimal, even when the noise and covariates are independent:

Estimator	Independent Noise	Arbitrary Noise
Prasad et. al. [PSBR20], Diakonikolas et. al. [DKK+18b]	$\epsilon\,\kappa^2$ (only $k=4$)	$\epsilon\,\kappa^2$ (only $k=4$)
Klivans, Kothari and Meka [KKM18b]	$\epsilon^{1-2/k}$	$\epsilon^{1-2/k}$
Zhu, Jiao and Steinhardt [ZJS20]	$\epsilon^{2-4/k}$	$\epsilon^{2-4/k}$
Our Work Thm 1.7, Cor 1.10	$\epsilon^{2-2/k}$	$\epsilon^{2-4/k}$
Lower Bounds Thm 1.9, Thm 1.11	$\epsilon^{2-2/k}$	$\epsilon^{2-4/k}$

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

Theorem 1.9 (Lower Bound for Independent Noise, Theorem 6.1 informal). *For any $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathbb{R}^2 \times \mathbb{R}$ such that the marginal distribution over $\mathbb{R}^2$ has covariance Σ and is (c,k)-hypercontractive for both distributions, and yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega\left(\epsilon^{1-1/k}\sigma\right)$, where Θ_1, Θ_2 are the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, $\sigma = \max(err_{\mathcal{D}_1}(\Theta_1), err_{\mathcal{D}_2}(\Theta_2))$ and the noise is uniform over $[-\sigma, \sigma]$. Further, $|err_{\mathcal{D}_1}(\Theta_2) - err_{\mathcal{D}_1}(\Theta_1)| = \Omega\left(\epsilon^{2-2/k}\sigma^2\right)$.*

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

Corollary 1.10 (Robust Regresssion with Dependent Noise, Corollary 4.2 informal). *Given $\epsilon > 0, k \geqslant 4$ and $n \geqslant (d\log(d))^{\mathcal{O}(k)}$ samples from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$, such that $\mathcal{D}$ is (c,k)-certifiably hypercontractive, there exists an algorithm that runs in $n^{\mathcal{O}(k)}$ time and outputs an estimator $\tilde{\Theta}$ such that with probability 9/10,*

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

and,

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

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

Theorem 1.11 (Lower Bound for Dependent Noise, Theorem 6.2 informal). *For any $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathbb{R}^2 \times \mathbb{R}$ such that the marginal distribution over $\mathbb{R}^2$ has covariance Σ and is (c,k)-hypercontractive for both distributions, and yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega\left(\epsilon^{1-2/k}\sigma\right)$, where Θ_1, Θ_2 be the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively and $\sigma = \max(err_{\mathcal{D}_1}(\Theta_1), err_{\mathcal{D}_2}(\Theta_2))$. Further, $|err_{\mathcal{D}_1}(\Theta_2) - err_{\mathcal{D}_1}(\Theta_1)| = \Omega\left(\epsilon^{2-4/k}\sigma^2\right)$.*

Applications for Gaussian Covariates. The special case where the marginal distribution over x is Gaussian has received considerable interest recently [DKS18, DKK+18b]. We note that our

estimators extend to the setting of Gaussian covariates, since the uniform distribution over samples from $\mathcal{N}(0,\Sigma)$ are $(\mathcal{O}(k),\mathcal{O}(k))$-certifiably hypercontractive for all k (see Section 5 in Kothari and Steurer [KS17c]). As a consequence, instantiating Corollary 1.10 with $k = \log(1/\epsilon)$ yields the following:

Corollary 1.12 (Robust Regression with Gaussian Covariates). *Given $\epsilon > 0$ and $n \geqslant (d\log(d))^{\mathcal{O}(log(1/\epsilon))}$ samples from $\mathcal{R}_{\mathcal{N}}(\epsilon,\Sigma^*,\Theta^*)$, such that the marginal distribution over the x's is $\mathcal{N}(0,\Sigma^*)$, there exists an algorithm that runs in $n^{\mathcal{O}(\log(1/\epsilon))}$ time and outputs an estimator $\tilde{\Theta}$ such that with high probability,*

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

and,

$$err_{\mathcal{N}}(\tilde{\Theta}) \leqslant \left(1 + \mathcal{O}\Big((\epsilon\log(1/\epsilon))^2\Big)\right) err_{\mathcal{N}}(\Theta^*)$$

We note that our estimators obtain the rate matching recent work for Gaussians, albeit in quasi-polynomial time. In comparison, Diakonikolas, Kong and Stewart [DKS18] obtain the same rate in polynomial time, when the noise is independent of the covariates. We note that obtaining the optimal rate for Gaussian covariates (shaving the additional $\log(1/\epsilon)$ factor) remains an outstanding open question.

1.2 Related Work.

Robust Statistics. Computationally efficient estimators for robust statistics in high dimension have been extensively studied, following the initial work on robust mean estimation [DKK+16, LRV16]. We focus on literature regarding robust regression and sum-of-squares. We refer the reader to recent surveys and theses for an extensive discussion of the literature on robust statistics [RSS18, Li18, Ste18, DK19].

Robust Regression. Computationally efficient estimators for robust linear regression were proposed by [PSBR20, KKM18b, DKK+19, DKS19]. While [PSBR20] and [DKK+19] obtained estimators for the more general case of distributions with bounded 4th moment. However, their estimators suffer an error of $\mathcal{O}\big(\mathrm{err}_D(\Theta^*)\epsilon\kappa^2(\Sigma)\big)$, where $\kappa(\Sigma)$ is the condition number of the covariance matrix of X. Hence, these estimators don't obtain the optimal dependence on ϵ in the negatively correlated noise setting, and also suffer an additional condition number dependence in the dependent noise setting. [DKS19] obtained improved bounds under the restrictive assumption that X is distributed according to a Gaussian. [KKM18b] obtained polynomial-time estimators for distributions with certifiably bounded distributions, however, their estimators obtain a sub-optimal error of $\mathcal{O}\big(\mathrm{err}_D(\Theta^*)\epsilon^{1-2/k}\big)$. In very recent and independent work, Zhu, Jiao and Steinhardt [ZJS20] obtained polynomial time estimators for the dependent noise setting, but their estimators are sub-optimal for the negatively correlated setting.

There has been significant work in more restrictive noise models as well. For instance, a series of works [BJK15, BJKK17, SBRJ19] consider a noise model where the adversary is only allowed to corrupt the labels and obtain *consistent* estimators in this regime (error goes to zero with more samples). In comparison, our estimators do not obtain For a comprehensive overview we refer the reader to references in the aforementioned papers.

Sum-of-Squares Algorithms. In recent years, there has been a significant progress in applying the Sum-of-Squares framework to design efficient algorithms for several fundamental computational problems. Starting with the work of Barak, Kelner and Steurer [BKS15], sum-of-squares algorithms have the best known running time for dictionary learning and tensor decomposition [MSS16a, SS17, HSS19], optimizing random tensors over the sphere [BGL17] and refuting CSPs below the spectral threshold [RRS17].

In the context of high-dimensional estimation, sum-of-squares algorithms achieved state-of-the-art performance for robust moment estimation [KS17c], robust regression[KKM18b], robustly learning mixtures of spherical [KS17a, HL18] and arbitrary Gaussians [BK20b, DHKK20], heavy-tailed estimation [Hop18, CHK$^+$20] and list-decodable variants of these problems [KKK19, RY20a, RY20b, BK20a, CMY20].

Concurrent Work. We note that a statistical estimator achieving rate proportional to $\epsilon^{1-1/k}$ can be obtained from combining ideas in [ZJS19] and [ZJS20][1]. However, this approach remains computationally intractable. Finally, Cherapanamjeri et al. [CHK$^+$20] consider the special case of $k = 4$ and obtain nearly linear sample complexity and running time. However, their running time and rate incurs a condition number dependence. Further, their rate scales proportional to $\epsilon^{1/2}$, even when the noise is indpendent of the covariates (as opposed to $\epsilon^{3/4}$).

We emphasize that the bottleneck in all prior and concurrent work remains algorithmically exploiting the independence of the noise and covariates, which we achieve via the *negatively correlated moments* condition (Definition 1.5).

2 Technical Overview

In this section, we provide an overview of our approach, the new algorithmic ideas we introduce and the corresponding technical challenges. At a high level, we build on several recent works that study Sum-of-Squares relaxations for solving algorithmic problems arising in robust statistics Following the *proofs-to-algorithms* paradigm arising from the aforementioned works, we show that given two distributions over regression instances that are close in *total variation distance* (definition 3.1), *any* hyperplane minimizing the least-squares loss on one distribution must be close (in ℓ_2 distance) to any other hyperplane minimizing the loss on the second distribution.

This information-theoretic statement immediately yields a robust estimator achieving optimal rate, albeit given access to unbounded computation. To see this, let $\mathcal{D}_1$ be the uniform distribution over n i.i.d samples from the true distribution, and $\mathcal{D}_2$ be the uniform distribution over n corrupted samples from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$, denoted by $\mathcal{D}_2$. It is easy to check that the total variation distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ is at most ϵ. Therefore, the two hyperplanes must be close in ℓ_2 norm. In order to make this strategy algorithmic, we show that we can distilled a set of polynomial constraints from the information theoretic proof and can efficiently optimize over them using the Sum-of-Squares framework. We note that we crucially require several new constraints, including the *gradient* condition and the *negatively-correlated moments* condition (see Section 2.1), which did not appear in any prior works. We describe each step in more detail subsequently.

[1] We thank Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt for communicating their observation to us.

2.1 Total Variation Closeness implies Hyperplane Closeness

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

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

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

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

We refer the reader to Theorem 4.1 for a precise statement. Further, we show that the $\epsilon^{1-1/k}$ dependence can be achieved even when the noise is not completely independent of the covariates but satisfies an analytic condition which we refer to as *negatively correlated moments* (see Definition 1.5). We provide an outline of the proof as it illustrates where the techniques we introduced in this work.

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

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

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

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

With a bit of effort, we can combine Equations (1) and (2), and upper bound them as follows (see Theorem 4.1 for details):

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

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

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

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

Gradient Conditions. To bound terms (i) and (ii) in Equation (3), we crucially rely on *gradient information* provided by the least-squares objective. Concretely, a key observation in our information-theoretic proof is that the candidate hyperplanes are locally optimal: given least-squares loss, for $i \in \{1, 2\}$ for all vectors v,

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

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

Independence and Negatively Correlated Moments. We observe that term (iii) can be interpreted as the k-th order correlation between the distribution of the covariates projected along v and the distribution of the noise in the linear model. Here, we observe that if the linear model satisfies the *negatively correlated moments* condition (Definition 1.5), we can decouple the expectation and bound each term independently:

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

Observe, when the underlying linear model has independent noise, Equation (5) follows for any k. We thus crucially exploit the structure of the noise and require a considerably weaker notion than independence. Further, if the *negatively correlated moments* property is not satisfied, we can use Cauchy-Schwarz to decouple the expectation in Equation (5) and incur a $\epsilon^{1-2/k}$ dependence (see Corollary 4.2). Conceptually, we emphasize that the *negatively correlated moments* condition may be of independent interest to design estimators that exploit independence in various statistics problems.

Hypercontractivity. To bound the RHS in Equation (5), we use our central distributional assumption of hypercontractive k-th moments (Definition 1.3) of the covariates :

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

We can bound the noise similarly, by assuming that the noise is hypercontractive and this considerably simplifies our statements. However, hypercontractivity of the noise is not a necessary assumption and prior work indeed incurs a term proportional to the k-th moment of the noise. Assuming boundedness of the regression vectors, Klivans, Kothari and Meka [KKM18b] obtained a uniform upper bound on k-th moment of the noise by truncating large samples. We note that the same holds for our estimators and we refer the reader to Section 5.2.3 in their paper. Finally, substituting $v = \Theta_1 - \Theta_2$ and rearranging, completes the information-theoretic proof.

We note that our approach already differs from prior work [KKM18b, PSBR20, ZJS19] and to our knowledge, we obtain the first information theoretic proof that being ϵ-close in TV distance implies that the optimal hyperplanes are $\mathcal{O}(\epsilon^{1-1/k})$ close in ℓ_2 distance. Next, we use this proof as motivation to construct a robust estimator. We do this by explicitly enforcing the gradient condition, negatively correlated moments, and hypercontractivity as constraints in the description of the robust estimator. In contrast, prior work only uses hypercontractivity or the gradient information in the analysis of their robust estimators. We describe these details below.

2.2 Proofs to Inefficient Algorithms: Distilling Constraints

Given an ϵ-corrupted sample generated by Model 1.1, a natural approach to design a robust (albeit inefficient) estimator is to search over all subsets of size $(1-\epsilon)n$, minimize the least squares objective on each one and output the hyperplane that achieves the minimal least-squares objective. Klivans, Kothari and Meka [KKM18b] implicitly analyze this estimator and show that it obtains rate $\epsilon^{1/2-1/k}$, which is sub-optimal. Instead, we search over all subsets of size $(1-\epsilon)n$ and in addition to minimizing the least-squares objective, we check if the uniform distribution over the samples is hypercontractive, the corresponding hyperplane satisfies the gradient condition and that the distribution over the covariates and noise satisfies the negatively correlated moments condition. Then, picking the hyperplane that achieves minimal least-squares cost and satisfies the aforementioned constraints obtains the optimal rate.

Since we work in the strong contamination model, where the adversary is allowed to both add and delete samples, there may not be any i.i.d. subset of $(1-\epsilon)n$ samples and thus enforcing constraints directly on the samples does not suffice. Instead, we create variables $2n$ variables denoted by $\{(x'_1, y'_1), \ldots (x'_n, y'_n)\}$ that serve as a proxy for the uncorrupted samples. We can now enforce constraints on the variables with impunity since a there exists a feasible assignment, namely the

uncorrupted samples. With the goal of obtaining an efficiently computable estimator, we restrict to enforcing only polynomial constraints, instead of arbitrary ones.

Intersection Constraints. The discrete analogue of the coupling argument is to ensure high intersection between the variables of our polynomial program and the uncorrupted samples. We know that at least a $(1-\epsilon)$-fraction of the samples we observe agree with the uncorrupted samples. To this end, we create indicator variables w_i, for $i \in [n]$ such that :

$$\begin{Bmatrix} & \sum_{i\in[n]} w_i = (1-\epsilon)n \\ \forall i \in [n]. & w_i^2 = w_i \\ \forall i \in [n] & w_i(x_i' - x_i) = 0 \\ \forall i \in [n] & w_i(y_i' - y_i) = 0 \end{Bmatrix}$$

The intersection constraints ensure that our polynomial system variables agree with the observed samples on $(1-\epsilon)n$ points. We note that such constraints are now standard in the literature, and indeed are the only constraints explicitly enforced by [KKM18b].

Independence as a Polynomial Inequality. The central challenge in obtaining optimal rates for robust regression is to leverage the independence of the noise and covariates. Since independence is a property of the marginals of $\mathcal{D}$, it is not immediately clear how to leverage it while designing a robust estimator.

However, recall that we do not require independence in full generality and use *negatively correlated moments* as a proxy for independence. Ideally, we would want to enforce the polynomial inequality corresponding to *negatively correlated moments* directly on the variables of our polynomial program as follows:

$$\left\{\forall r \leqslant k/2, \quad \frac{1}{n}\sum_{i\in[n]} \left(v^\top x_i' \left(y_i' - (x_i')^\top \Theta\right)\right)^{2r} \leqslant \mathcal{O}(1) \left(\frac{1}{n}\sum_{i\in[n]} (v^\top x_i')^{2r}\right) \left(\frac{1}{n}\sum_{i\in[n]} \left(y_i' - (x_i')^\top \Theta\right)^{2r}\right)\right\}$$

where Θ is a variable corresponding to the true hyperplane. To demonstrate feasibility of this constraint, we would require a finite sample analysis, showing that uncorrupted samples from a hypercontractive distribution satisfy the above inequality. Observe, when $r = k/2$, the LHS is a degree-k polynomial and our distribution may be too heavy-tailed to achieve any concentration.

Instead, we observe that since hypercontractivity is preserved under sampling, we can relax our polynomial constraint by applying hypercontractivity to the terms in the RHS above :

$$\left\{\forall r \leqslant k/2, \quad \frac{1}{n}\sum_{i\in[n]} \left(v^\top x_i' \left(y_i' - (x_i')^\top \Theta\right)\right)^{2r} \leqslant \mathcal{O}(1) \left(\frac{1}{n}\sum_{i\in[n]} (v^\top x_i')^{2}\right)^r \left(\frac{1}{n}\sum_{i\in[n]} \left(y_i' - (x_i')^\top \Theta\right)^{2}\right)^r\right\}$$

In Lemma 5.4 we show that the above inequality is feasible for the uncorrupted samples. In particular, given at least, $d^{\Omega(k)}$ i.i.d. samples from $\mathcal{D}$, the above inequality holds on the samples with high probability.

Perhaps surprisingly, the dependence on ϵ achieved when $\mathcal{D}$ has *negatively correlated moments* matches the information theoretically optimal rate for independent noise. We thus expect the notion of *negatively correlated moments* to lead to new estimators for problems where independence of random variables requires to be formulated as a polynomial inequality.

Hypercontractivity Constraints. Since hypercontractivity is a already a polynomial identity relating the k-th moment to the variance of a distribution, it can easily enforced as a constraint. Conveniently, if the underlying distribution $\mathcal{D}$ is hypercontractive, then the uniform distribution over a sufficiently large sample is also sampling also hypercontractive (see Lemma 5.7 in [KKM18b]). Therefore, the following constraints are feasible:

$$\left\{ \begin{array}{ll} \forall r \leqslant k/2 & \frac{1}{n}\sum_{i\in[n]} \langle x_i', v\rangle^{2r} \leqslant \left(\frac{c_r}{n}\sum_{i\in[n]} \langle x_i', v\rangle^2\right)^r \\ \forall r \leqslant k/2 & \frac{1}{n}\sum_{i\in[n]} \left(y_i' - \langle \Theta, x_i'\rangle\right)^{2r} \leqslant \left(\frac{\eta_r}{n}\sum_{i\in[n]} \left(y_i' - \langle \Theta, x_i'\rangle\right)^2\right)^r \end{array} \right\}$$

We note that Klivans, Kothari and Meka [KKM18b] use hypercontractivity of the uncorrupted samples in their analysis but do not explicitly enforce this as a constraint. Enforcing hypercontractivity explicitly on the samples was used by Kothari and Steurer [KS17c] in the context of robust moment estimation and Kothari and Steinhardt [KS17a] in the context of robustly clustering a mixture of spherical Gaussians.

Gradient Constraints. Finally, it is crucial in our analysis to enforce that the minimizer we are searching for, Θ, has gradient 0. For the least-squares loss, the gradient has a simple analytic form : for all $v \in \mathbb{R}^d$,

$$\left\{ \left\langle v, \frac{1}{n}\sum_{i\in[n]} x_i' \left(\langle x_i', \Theta\rangle - y_i'\right)\right\rangle^k = 0 \right\}$$

While such optimality conditions are often used in the analysis of estimators (as done in [PSBR20]), we emphasize that we hardcode the gradient condition into the description of our robust estimator. To the best of our knowledge, no estimator for robust/high-dimensional statistics includes explicit optimality constraints as a part of a polynomial system.

Solving the least-squares objective on the samples subject to the polynomial system described by the aforementioned constraints results in an estimator for robust regression that achieves optimal rate. Recall, this follows immidiately from our robust certifiability proof. However, optimizing polynomial systems can be NP-Hard in the worse case, and thus we briefly describe how to avoid this computational intractability next.

2.3 Efficient Algorithms via Sum-of-Squares

We use the sum-of-squares method to make the aforementioned polynomial system efficiently computable and provide an outline of this approach (see Section 3 for a formal treatment of sum-of-squares proofs). Instead of directly solving the polynomial program, let us instead consider finding a distribution, μ, over feasible solutions w, x', y' and Θ that minimizes

$$\mathop{\mathbb{E}}_{w,x',y',\Theta\sim\mu}\left[\frac{1}{n}\sum_{i\in[n]}\left(w_i y_i' - \langle w_i x_i', \Theta\rangle\right)^2\right]$$

and satisfies the constraints above. Then, it follows from our information-theoretic proof (Theorem 4.1) that

$$\mathop{\mathbb{E}}_{\mu}\left[\left\|\Sigma^{1/2}\left(\Theta^* - \Theta\right)\right\|_2\right] \leqslant \mathcal{O}\left(\epsilon^{1-1/k}\right)\mathrm{err}_{\mathcal{D}}\left(\Theta^*\right)^{1/2} \tag{6}$$

where Θ^* is the optimal hyperplane.

We now face two challenges: finding a distribution over solutions is at least as hard as the original problem and we no longer recover a unique hyperplane. The latter is easy to address by observing that the hyperplane obtained by averaging over the distribution, μ, suffices:

$$\left\|\Sigma^{1/2}\left(\Theta^* - \mathop{\mathbb{E}}_{\mu}[\Theta]\right)\right\|_2 \leqslant \mathop{\mathbb{E}}_{\mu}\left[\left\|\Sigma^{1/2}\left(\Theta^* - \Theta\right)\right\|_2\right]$$

where the inequality follows from convexity of the loss.

Following prior works, it is now natural to instead consider searching for a "pseudo-distribution", ζ, over feasible solutions. A pseudo-distribution is an object similar to a real distribution, but relaxed to allow negative mass on its support (see Subsection 3.1 for a formal treatment). Crucially, a pseudo-distribution over the polynomial program can be computed efficiently by formulating it as a large SDP. To see why this helps, note any polynomial inequality that can be derived using "sum-of-squares" proofs from a set of polynomial constraints using a *low-degree sum-of-squares proof* remains valid if we replace distributions by "pseudo-distribution".

For instance, if Equation 6 were a polynomial inequality in w, x', y' and Θ, obtained by applying simple transformations that admit sum-of-squares proofs, we could replace μ by ζ, and obtain an efficient estimator. However, Equation 6 is not a polynomial inequality and the proof outlined in Subsection 2.1 is not a low-degree sum-of-squares proof. Therefore, a central technical contribution of our work is to formulate the right polynomial identity bounding the distance between Θ^* and Θ in terms of the least-squares error incurred by by Θ^*, and deriving this bound from the polynomial constraints using a low-degree sum-of-squares proof. We do this in Section 5.

2.4 Distribution Families

We note that our statistical estimator applies to all distributions, $\mathcal{D}$, that are (c_k, k)-hypercontractive and the rate is completely determined by whether $\mathcal{D}$ has *negatively correlated moments*. In particular, for the important special case of heavy-tailed regression with independent noise, we obtain rate proportional to $\epsilon^{1-1/k}$ for parameter recovery.

However, similar to prior work on hypercontractive distributions, our efficient estimators apply to a more restrictive class, i.e. *certifiably* hypercontractive distributions (Definition 3.5). Intuitively, this condition captures the criteria that information about degree-k moment upper bounds is "algorithmically accessible". Certifiably hypercontractive distributions are a broad class and include affine transformations of isotropic distributions satisfying Poincaré inequalities and all strongly log-concave distributions. For a detailed discussion of distributions satisfying Poincaré inequalities and their closure properties, we refer the reader to [KS17a, KS17c].

Surprisingly, while we enforce a constraint corresponding to *negatively correlated moments*, we do not require a certifiable variant of this condition. Therefore, our efficient estimators hold for regression instances where the true distribution satisfies this condition, including the special case where the noise is independent from the covariates. Finally, our estimators unify various noise models and imply that even in the agnostic setting, the rate degrades only when the noise is positively correlated with the covariates.

3 Preliminaries

Throughout this paper, for a vector v, we use $\|v\|_2$ to denote the Euclidean norm of v. For a $n \times m$ matrix M, we use $\|M\|_2 = \max_{\|x\|_2=1} \|Mx\|_2$ to denote the spectral norm of M and $\|M\|_F = \sqrt{\sum_{i,j} M_{i,j}^2}$ to denote the Frobenius norm of M. For symmetric matrices we use $\succeq$ to denote the PSD/Loewner ordering over eigenvalues of M. Recall, the definition of total variation distance between probability measures:

Definition 3.1 (Total Variation Distance)**.** The TV distance between distributions with PDFs p, q is defined as $\frac{1}{2}\int_{-\infty}^{\infty} |p(x) - q(x)|dx$.

Given a distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathbb{R}$, we consider the least squares error of a vector Θ w.r.t. $\mathcal{D}$ to be $\mathrm{err}_{\mathcal{D}}(\Theta) = \mathbb{E}_{x,y\sim\mathcal{D}}\left[(y - \langle x, \Theta\rangle)^2\right]$. The linear regression problem minimizes the error over all Θ. The minimizer, $\Theta_{\mathcal{D}}$ of the aformentioned error satisfies the following "gradient condition" : for all $v \in \mathbb{R}^d$,

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

Fact 3.2 (Convergence of Empirical Moments, implicit in Lemma 5.5 [KS17c])**.** *Let $\mathcal{D}$ be a (c_k, k)-hypercontractive distribution with covariance Σ and let $\mathcal{X} = \{x_1, \dots x_n\}$ be $n = \Omega((d\log(d)/\delta)^{k/2})$ i.i.d. samples from $\mathcal{D}$. Then, with probability at least $1-\delta$,*

$$(1 - 0.1)\Sigma \preceq \frac{1}{n}\sum_i^n x_i x_i^\top \preceq (1 + 0.1)\Sigma$$

Fact 3.3 (TV Closeness to Covariance Closeness, Lemma 2.2 [KS17c])**.** *Let $\mathcal{D}_1, \mathcal{D}_2$ be (c_k, k)-hypercontractive distributions over $\mathbb{R}^d$ such that $\|\mathcal{D} - \mathcal{D}'\|_{TV} \leqslant \epsilon$, where $0 < \epsilon < \mathcal{O}\Big((1/c_k)^{\frac{k}{k-1}}\Big)$. Let Σ_1, Σ_2 be the corresponding covariance matrices. Then, for $\delta \leqslant \mathcal{O}\big(c_k\,\epsilon^{1-1/k}\big) < 1$,*

$$(1-\delta)\Sigma_2 \preceq \Sigma_1 \preceq (1+\delta)\Sigma_2$$

Lemma 3.4 (Löwner Ordering for Hypercontractive Samples)**.** *Let $\mathcal{D}$ be a (c_k, k)-hypercontractive distribution with covariance Σ and and let $\mathcal{U}$ be the uniform distribution over n samples. Then, with probability $1-\delta$,*

$$\left\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I\right\|_F \leqslant \frac{C_4 d^2}{\sqrt{n}\sqrt{\delta}},$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{i\in[n]} x_i x_i^\top$.

Next, we define the technical conditions required for efficient estimators. Formally,

Definition 3.5 (Certifiable Hypercontractivity)**.** A distribution $\mathcal{D}$ on $\mathbb{R}^d$ is (c_k, k)-certifiably hypercontractive if for all $r \leqslant k/2$, there exists a degree $\mathcal{O}(k)$ sum-of-squares proof (defined below) of the following inequality in the variable v

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

such that $c_r \leqslant c_k$.

Next, we note that if a distribution $\mathcal{D}$ is certifiably hypercontractive, the uniform distribution over n i.i.d. samples from $\mathcal{D}$ is also certifiably hypercontractive.

Fact 3.6 (Sampling Preserves Certifiable Hypercontractivity, Lemma 5.5 [KS17c])**.** *Let $\mathcal{D}$ be a (c_k, k)-certifiably hypercontractive distribution on $\mathbb{R}^d$. Let $\mathcal{X}$ be a set of $n = \Omega\left((d\log(d/\delta))^{k/2}/\gamma^2\right)$ i.i.d. samples from $\mathcal{D}$. Then, with probability $1-\delta$, the uniform distribution over $\mathcal{X}$ is $(c_k+\gamma, k)$-certifiably hypercontractive.*

We also note that certifiably hypercontractivity is preserved under Affine transformations of the distribution.

Fact 3.7 (Certifiable Hypercontractivity under Affine Transformations, Lemma 5.1, 5.2 [KS17c])**.** *Let $x \in \mathbb{R}^d$ be a random variable drawn from a (c_k, k)-certifiably hypercontractive distribution. Then, for matrix A and vector b, the distribution over the random variable $Ax+b$ is also (c_k, k)-certifiably hypercontractive.*

Next, we formally define the condition on the moments and noise that we require to obtain efficient algorithms. We note that for technical reasons it is not simply a polynomial identity encoding Definition 1.5.

Definition 3.8 (Certifiable Negatively Correlated Moments)**.** A distribution $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$ has $\mathcal{O}(1)$-certifiable negatively correlated moments if for all $r \leqslant k/2$ there exists a degree $\mathcal{O}(k)$ sum-of-squares proof of the following inequality

$$\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(v^\top x\left(y-x^\top\Theta\right)\right)^{2r}\right] \leqslant \mathcal{O}(\lambda_r^r)\left(\mathbb{E}\left[(v^\top x)^2\right]^r\right)\left(\mathbb{E}\left[\left(y-x^\top\Theta\right)^2\right]^r\right)$$

for a fixed vector Θ.

3.1 SoS Background.

In this subsection, we provide the necessary background for the sum-of-squares proof system. We follow the exposition as it appears in lecture notes by Barak [Bar], the Appendix of Ma, Shi and Steurer [MSS16b], and the preliminary sections of several recent works [KS17c, KKK19, KKM18a, BK20a, BK20b].

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

Definition 3.9 (Pseudo-distribution)**.** A level-ℓ pseudo-distribution is a finitely-supported function $D : \mathbb{R}^n \to \mathbb{R}$ such that $\sum_x D(x) = 1$ and $\sum_x D(x)f(x)^2 \geqslant 0$ for every polynomial f of degree at most $\ell/2$, where the summation is over all x in the support of D.

Next, we define the notion of pseudo-expectation.

Definition 3.10 (Pseudo-expectation)**.** The pseudo-expectation of a function f on $\mathbb{R}^d$ with respect to a pseudo-distribution D, denoted by $\tilde{\mathbb{E}}_{D(x)}[f(x)]$, is defined as

$$\tilde{\mathbb{E}}_{D(x)}[f(x)] = \sum_x D(x)f(x) \tag{7}$$

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

Fact 3.11 ([Sho87, Par00, Nes00, Las01])**.** *For any $n, \ell \in \mathbb{N}$, the following set has a $n^{O(\ell)}$-time weak separation oracle (in the sense of [GLS81]):*

$$\left\{ \tilde{\mathbb{E}}_{D(x)}(1, x_1, x_2, \ldots, x_n)^{\otimes \ell} \mid \text{degree-}\ell \text{ pseudo-distribution } D \text{ over } \mathbb{R}^n \right\} \tag{8}$$

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

Given a system of polynomial constraints, denoted by $\mathcal{A}$, we say that it is *explicitly bounded* if it contains a constraint of the form $\{\|x\|^2 \leqslant M\}$. Then, the following fact follows from Fact 3.11 and [GLS81]:

Fact 3.12 (Efficient Optimization over Pseudo-distributions)**.** *There exists an $(n+m)^{O(\ell)}$-time algorithm that, given any explicitly bounded and satisfiable system[2] $\mathcal{A}$ of m polynomial constraints in n variables, outputs a level-ℓ pseudo-distribution that satisfies $\mathcal{A}$ approximately.*

We now define basic facts for pseudo-distributions, which extend facts that hold for standard probability distributions, which can be found in several prior works listed above.

Fact 3.13 (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 $\tilde{\zeta}$, $\tilde{\mathbb{E}}_{\tilde{\zeta}}[fg] \leqslant \sqrt{\tilde{\mathbb{E}}_{\tilde{\zeta}}[f^2]}\sqrt{\tilde{\mathbb{E}}_{\tilde{\zeta}}[g^2]}$.*

Fact 3.14 (Hölder's Inequality for Pseudo-Distributions)**.** *Let f, g be polynomials of degree at most d in indeterminate $x \in \mathbb{R}^d$. Fix $t \in \mathbb{N}$. Then, for any degree dt pseudo-distribution $\tilde{\zeta}$, $\tilde{\mathbb{E}}_{\tilde{\zeta}}[f^{t-1}g] \leqslant \left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[f^t]\right)^{\frac{t-1}{t}} \left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[g^t]\right)^{1/t}$. In particular, for all even integers k, $\tilde{\mathbb{E}}_{\tilde{\zeta}}[f]^k \leqslant \tilde{\mathbb{E}}_{\tilde{\zeta}}[f^k]$.*

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

$$g = \sum_{S \subseteq [m]} p_S^2 \cdot \Pi_{i \in S} f_i \tag{9}$$

We say that this proof has *degree* ℓ if for every set $S \subseteq [m]$, the polynomial $p_S^2 \Pi_{i \in S} f_i$ has degree at most ℓ. If there is a degree ℓ SoS proof that $\{f_i \geqslant 0 \mid i \leqslant r\}$ implies $\{g \geqslant 0\}$, we write:

$$\{f_i \geqslant 0 \mid i \leqslant r\} \vdash_\ell \{g \geqslant 0\} \tag{10}$$

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

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 the following inference rules.

The first one derives new inequalities by addition/multiplication:

$$\frac{\mathcal{A} \vdash_{\ell} \{f \geqslant 0, g \geqslant 0\}}{\mathcal{A} \vdash_{\ell} \{f + g \geqslant 0\}}, \frac{\mathcal{A} \vdash_{\ell} \{f \geqslant 0\}, \mathcal{A} \vdash_{\ell'} \{g \geqslant 0\}}{\mathcal{A} \vdash_{\ell+\ell'} \{f \cdot g \geqslant 0\}} \qquad \text{(Addition/Multiplication Rule)}$$

The next one derives new inequalities by transitivity:

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

Finally, the last rule derives new inequalities via substitution:

$$\frac{\{F \geqslant 0\} \vdash_{\ell} \{G \geqslant 0\}}{\{F(H) \geqslant 0\} \vdash_{\ell \cdot \deg(H)} \{G(H) \geqslant 0\}} \qquad \text{(Substitution Rule)}$$

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 3.15 (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 3.16 (Completeness)**.** *Suppose $d \geqslant r' \geqslant r$ and $\mathcal{A}$ is a collection of polynomial constraints with degree at most r, and $\mathcal{A} \vdash \{\sum_{i=1}^{n} x_i^2 \leqslant B\}$ for some finite B.*

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

Basic Sum-of-Squares Proofs. Next, we use the following basic facts regarding sum-of-squares proofs. For further details, we refer the reader to a recent monograph [FKP+19].

Fact 3.17 (Operator norm Bound)**.** *Let A be a symmetric $d \times d$ matrix and v be a vector in $\mathbb{R}^d$. Then,*

$$\vdash_{2} \left\{ v^\top A v \leqslant \|A\|_2 \|v\|_2^2 \right\}.$$

Fact 3.18 (SoS Almost Triangle Inequality). *Let $f_1, f_2, \ldots, f_r$ be indeterminates. Then,*

$$\left| \frac{f_1, f_2, \ldots, f_r}{2t} \right\} \left(\sum_{i \leqslant r} f_i \right)^{2t} \leqslant r^{2t-1} \left(\sum_{i=1}^{r} f_i^{2t} \right) \right\}.$$

Fact 3.19 (SoS Hölder's Inequality). *Let $w_1, \ldots w_n$ be indeterminates and let $f_1, \ldots f_n$ be polynomials of degree m in vector valued variable x. Let k be a power of 2. Then,*

$$\left\{ w_i^2 = w_i, \forall i \in [n] \right\} \left| \frac{x, w}{2km} \right\} \left(\frac{1}{n} \sum_{i=1}^{n} w_i f_i \right)^{k} \leqslant \left(\frac{1}{n} \sum_{i=1}^{n} w_i \right)^{k-1} \left(\frac{1}{n} \sum_{i=1}^{n} f_i^k \right) \right\}.$$

We also use the following fact that allows us to cancel powers of indeterminates within the sum-of-squares proof system.

Fact 3.20 (Cancellation Within SoS, Lemma 9.4 [BK20b]). *Let a, C be indeterminates. Then,*

$$\{a \geqslant 0\} \cup \left\{ a^{2t} \leqslant C a^t \right\} \left| \frac{a, C}{2t} \right\} \left\{ a^{2t} \leqslant C^2 \right\}.$$

4 Robust Certifiability and Information Theoretic Estimators

In this section, we provide an estimator that obtains the information theoretically optimal rate for robust regression. We note that we consider the setting where both the covariates and the noise are hypercontractive and the are independent of each other. This setting displays all the key ideas of our estimator. Further, our estimator extends to the remaining settings, such as bounded dependent noise, by simple modifications to the subsequent analysis.

Theorem 4.1 (Robust Certifiability with Optimal Rate). *Given $\epsilon > 0$, let $\mathcal{D}, \mathcal{D}'$ be distributions over $\mathbb{R}^d \times \mathbb{R}$ such that the respective marginal distributions over $\mathbb{R}^d$, denoted by $\mathcal{D}_{\mathcal{X}}, \mathcal{D}'_{\mathcal{X}}$, are (c_k, k)-hypercontractive and $\|\mathcal{D} - \mathcal{D}'\|_{TV} \leqslant \epsilon$. Let $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma_{\mathcal{D}}, \Theta_{\mathcal{D}})$ and $\mathcal{R}_{\mathcal{D}'}(\epsilon, \Sigma_{\mathcal{D}'}, \Theta_{\mathcal{D}'})$ be the corresponding instances of robust regression such that $\mathcal{D}, \mathcal{D}'$ have negatively correlated moments. Further, for $(x, y) \sim \mathcal{D}, \mathcal{D}'$, let the marginal distribution over $y - \left\langle x, \mathbb{E}\left[xx^\top\right]^{-1} \mathbb{E}\left[xy\right] \right\rangle$ be (η_k, k)-hypercontractive Then,*

$$\left\| \Sigma_{\mathcal{D}}^{1/2} (\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}) \right\|_2 \leqslant \mathcal{O}\Big(\sqrt{c_k\, \eta_k}\, \epsilon^{1-1/k} \Big) \left(err_{\mathcal{D}}(\Theta_{\mathcal{D}})^{1/2} + err_{\mathcal{D}'}(\Theta_{\mathcal{D}'})^{1/2} \right)$$

Further,

$$err_{\mathcal{D}}(\Theta_{\mathcal{D}'}) \leqslant \left(1 + \mathcal{O}\Big(c_k\, \eta_k\, \epsilon^{2-2/k} \Big) \right) err_{\mathcal{D}}(\Theta_{\mathcal{D}}) + \mathcal{O}\Big(c_k\, \eta_k\, \epsilon^{2-2/k} \Big)\, err_{\mathcal{D}'}(\Theta_{\mathcal{D}'})$$

Proof. Consider a maximal coupling of $\mathcal{D}, \mathcal{D}'$ over $(x, y) \times (x', y')$, denoted by $\mathcal{G}$, such that the marginal of $\mathcal{G}$ (x, y) is $\mathcal{D}$, the marginal on (x', y') is $\mathcal{D}'$ and $\mathbb{P}_{\mathcal{G}}[\mathbb{I}\{(x, y) = (x', y')\}] = 1 - \epsilon$. Then, for all v,

$$\begin{aligned} \langle v, \Sigma_{\mathcal{D}}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}) \rangle &= \mathop{\mathbb{E}}_{\mathcal{G}} \left[\left\langle v, xx^\top (\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}) + xy - xy \right\rangle \right] \\ &= \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x \left(\langle x, \Theta_{\mathcal{D}} \rangle - y \right) \rangle \right] + \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x \left(y - \langle x, \Theta_{\mathcal{D}'} \rangle \right) \rangle \right] \end{aligned} \tag{11}$$

Since $\Theta_{\mathcal{D}}$ is the minimizer for the least squares loss, we have the following gradient condition : for all $v \in \mathbb{R}^d$,

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\langle v, (\langle x, \Theta_{\mathcal{D}}\rangle - y)x\rangle] = 0 \tag{12}$$

Since $\mathcal{G}$ is a coupling, using the gradient condition (12) and using that $1 = \mathbb{I}\{(x,y) = (x',y')\}$ $+\mathbb{I}\{(x,y) \neq (x',y')\}$, we can rewrite equation (11) as

$$\begin{aligned}
\langle v, \Sigma_{\mathcal{D}}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle \mathbb{I}\left\{(x,y) = (x',y')\right\}\right] \\
&\quad + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle \mathbb{I}\left\{(x,y) \neq (x',y')\right\}\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle \mathbb{I}\left\{(x,y) = (x',y')\right\}\right] \\
&\quad + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle \mathbb{I}\left\{(x,y) \neq (x',y')\right\}\right]
\end{aligned} \tag{13}$$

Consider the first term in the last equality above. Using the gradient condition for $\Theta_{\mathcal{D}'}$ along with Hölder's Inequality, we have

$$\begin{aligned}
&\left|\mathbb{E}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle \mathbb{I}\left\{(x,y) = (x',y')\right\}\right]\right| \\
&\quad = \left|\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle\right] - \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle \mathbb{I}\left\{(x,y) \neq (x',y')\right\}\right]\right| \\
&\quad = \left|\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle \mathbb{I}\left\{(x,y) \neq (x',y')\right\}\right]\right| \\
&\quad \leqslant \left|\mathop{\mathbb{E}}_{\mathcal{G}}\left[\mathbb{I}\left\{(x,y) \neq (x',y')\right\}^{k/(k-1)}\right]^{(k-1)/k}\right| \cdot \left|\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle^k\right]^{1/k}\right|
\end{aligned} \tag{14}$$

Observe, since $\mathcal{G}$ is a maximal coupling $\mathbb{E}_{\mathcal{G}}\left[\mathbb{I}\left\{(x,y) \neq (x',y')\right\}\right]^{(k-1)/k} \leqslant \epsilon^{1-1/k}$. Further, since $\mathcal{D}'$ has negatively correlated moments,

$$\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left\langle v, x'\right\rangle^k \cdot \left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)^k\right] = \mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left\langle v, x'\right\rangle^k\right] \mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)^k\right]$$

By hypercontractivity of the covariates and the noise, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left\langle v, x'\right\rangle^k\right]^{1/k} \mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)^k\right]^{1/k} \leqslant \mathcal{O}(\sqrt{c_k\,\eta_k})\left(v^\top \Sigma_{\mathcal{D}'} v\right)^{1/2} \mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)^2\right]^{1/2}$$

Therefore, we can restate (14) as follows

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle \mathbb{I}\left\{(x,y) = (x',y')\right\}\right]\right| &\leqslant \mathcal{O}\left(\sqrt{c_k\,\eta_k}\,\epsilon^{\frac{k-1}{k}}\right)\left(v^\top \Sigma_{\mathcal{D}'} v\right)^{\frac{1}{2}} \\
&\quad \mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)^2\right]^{\frac{1}{2}}
\end{aligned} \tag{15}$$

It remains to bound the second term in the last equality of equation (13), and we proceed as follows :

$$\mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y-\langle x,\Theta_{\mathcal{D}'}\rangle\right)\rangle \mathbb{I}\left\{(x,y)\neq(x',y')\right\}\right] = \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle \mathbb{I}\left\{(x,y)\neq(x',y')\right\}\right] + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y-\langle x,\Theta_{\mathcal{D}}\rangle\right)\rangle \mathbb{I}\left\{(x,y)\neq(x',y')\right\}\right] \tag{16}$$

We bound the two terms above separately. Observe, applying Hölder's Inequality to the first term, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle \mathbb{I}\left\{(x,y)\neq(x',y')\right\}\right] &\leqslant \mathop{\mathbb{E}}_{\mathcal{G}}\left[\mathbb{I}\left\{(x,y)\neq(x',y')\right\}\right]^{\frac{k-2}{k}} \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle^{\frac{k}{2}}\right]^{\frac{2}{k}} \\ &\leqslant \epsilon^{\frac{k-2}{k}} \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle^{\frac{k}{2}}\right]^{\frac{2}{k}}\end{aligned} \tag{17}$$

To bound the second term in equation 16, we again use Hölder's Inequality followed $\mathcal{D}$ having negatively correlated moments,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y-\langle x,\Theta_{\mathcal{D}}\rangle\right)\rangle \mathbb{I}\left\{(x,y)\neq(x',y')\right\}\right] &\leqslant \mathop{\mathbb{E}}_{\mathcal{G}}\left[\mathbb{I}\left\{(x,y)\neq(x',y')\right\}\right]^{\frac{k-1}{k}} \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y-\langle x,\Theta_{\mathcal{D}}\rangle\right)\rangle^k\right]^{\frac{1}{k}} \\ &\leqslant \epsilon^{\frac{k-1}{k}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\langle v, x\rangle^k\right]^{1/k} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x,\Theta_{\mathcal{D}}\rangle\right)^k\right]^{1/k} \\ &\leqslant \epsilon^{\frac{k-1}{k}} \sqrt{c_k\,\eta_k}\left(v^\top\Sigma_{\mathcal{D}}v\right)^{1/2} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x,\Theta_{\mathcal{D}}\rangle\right)^2\right]^{1/2}\end{aligned} \tag{18}$$

where the last inequality follows from hypercontractivity of the covariates and noise. Substituting the upper bounds obtained in Equations (17) and (18) back in to (16),

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y-\langle x,\Theta_{\mathcal{D}'}\rangle\right)\rangle \mathbb{I}\left\{(x,y)\neq(x',y')\right\}\right] &\leqslant \epsilon^{\frac{k-2}{k}} \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle^{\frac{k}{2}}\right]^{\frac{2}{k}} \\ &\quad + \epsilon^{\frac{k-1}{k}} \sqrt{c_k\,\eta_k}\left(v^\top\Sigma_{\mathcal{D}}v\right)^{1/2} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x,\Theta_{\mathcal{D}}\rangle\right)^2\right]^{1/2}\end{aligned}$$

Therefore, we can now upper bound both terms in Equation (13) as follows:

$$\begin{aligned}\langle v, \Sigma_{\mathcal{D}}(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'})\rangle &\leqslant \mathcal{O}\left(c_k\,\eta_k\,\epsilon^{\frac{k-1}{k}}\right)\left(v^\top\Sigma_{\mathcal{D}'}v\right)^{1/2} \mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y'-\left\langle x',\Theta_{\mathcal{D}'}\right\rangle\right)^2\right]^{1/2} \\ &\quad + \mathcal{O}\left(\epsilon^{\frac{k-2}{k}}\right)\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle^{k/2}\right]^{2/k} \\ &\quad + \mathcal{O}\left(\epsilon^{\frac{k-1}{k}}\sqrt{c_k\,\eta_k}\right)\left(v^\top\Sigma_{\mathcal{D}}v\right)^{1/2} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x,\Theta_{\mathcal{D}}\rangle\right)^2\right]^{1/2}\end{aligned} \tag{19}$$

Recall, since the marginals of $\mathcal{D}$ and $\mathcal{D}'$ on $\mathbb{R}^d$ are (c_k, k)-hypercontractive and $\|\mathcal{D}-\mathcal{D}'\|_{\mathsf{TV}} \leqslant \epsilon$, it follows from Fact 3.3 that

$$(1-0.1)\,\Sigma_{\mathcal{D}'} \preceq \Sigma_{\mathcal{D}} \preceq (1+0.1)\,\Sigma_{\mathcal{D}'} \tag{20}$$

when $\epsilon \leqslant \mathcal{O}\big((1/c_k k)^{k/k-1}\big)$. Now, consider the substitution $v = \Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}$. Observe,

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}\right)\right\rangle^{k/2}\right]^{2/k} &= \mathop{\mathbb{E}}_{\mathcal{D}}\left[\langle x, (\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle^k\right]^{2/k} \\ &\leqslant c_k \left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2^2 \end{aligned} \tag{21}$$

Then, using the bounds in (20) and (21) along with $v = \Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}$ in Equation 19, we have

$$\begin{aligned} \left(1 - \mathcal{O}\left(\epsilon^{\frac{k-2}{k}} c_k\right)\right)\left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2^2 \leqslant \mathcal{O}\left(\sqrt{c_k\,\eta_k}\,\epsilon^{\frac{k-1}{k}}\right)\left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2 \\ \left(\mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)^2\right]^{\frac{1}{2}} + \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[(y - \langle x, \Theta_{\mathcal{D}}\rangle)^2\right]^{\frac{1}{2}}\right) \end{aligned} \tag{22}$$

Dividing out (22) by $\left(1 - \mathcal{O}\left(\epsilon^{\frac{k-2}{k}} c_k\right)\right)\left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2^2$ and observing that $\mathcal{O}\left(\epsilon^{\frac{k-2}{k}} c_k\right)$ is upper bounded by a fixed constant less than 1 yields the parameter recovery bound.

Given the parameter recovery result above, we bound the least-squares loss between the two hyperplanes on $\mathcal{D}$ as follows:

$$\begin{aligned} \left|\mathrm{err}_{\mathcal{D}}(\Theta_{\mathcal{D}}) - \mathrm{err}_{\mathcal{D}}(\Theta_{\mathcal{D}'})\right| &= \left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left(y - x^\top\Theta_{\mathcal{D}}\right)^2 - \left(y - x^\top\Theta_{\mathcal{D}'} + x^\top\Theta_{\mathcal{D}} - x^\top\Theta_{\mathcal{D}}\right)^2\right]\right| \\ &= \left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\langle x, (\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle^2 + 2(y - x^\top\Theta_{\mathcal{D}})x^\top(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right]\right| \\ &\leqslant \mathcal{O}\left(c_k\,\eta_k\,\epsilon^{2-2/k}\right)\left(\mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)^2\right] + \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[(y - \langle x, \Theta_{\mathcal{D}}\rangle)^2\right]\right) \end{aligned} \tag{23}$$

where the last inequality follows from observing $\mathbb{E}\left[\left\langle\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}, x(y - x^\top\Theta_{\mathcal{D}})\right\rangle\right] = 0$ (gradient condition) and squaring the parameter recovery bound. □

Next, we consider the setting where the noise is allowed to dependent arbitrarily on the covariates, which captures the well-studied agnostic model. With a slightly modification in our certifiability proof above (using Cauchy-Schwarz instead of independence), we obtain the optimal rate in this setting. We defer the details to Appendix A.

Corollary 4.2 (Robust Regression with Dependent Noise). *Let $\mathcal{D}, \mathcal{D}'$ be distributions over $\mathbb{R}^d \times \mathbb{R}$ and let $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma_{\mathcal{D}}, \Theta_{\mathcal{D}})$, $\mathcal{R}_{\mathcal{D}'}(\epsilon, \Sigma_{\mathcal{D}'}, \Theta_{\mathcal{D}'})$ be robust regression instances satisfying the hypothesis in Theorem 4.1 such that the negatively correlated moments condition is not satisfied. Then,*

$$\left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2 \leqslant \mathcal{O}\left(\sqrt{c_k\,\eta_k}\,\epsilon^{1-2/k}\right)\left(err_{\mathcal{D}}(\Theta_{\mathcal{D}})^{1/2} + err_{\mathcal{D}'}(\Theta_{\mathcal{D}'})^{1/2}\right)$$

Further,

$$err_{\mathcal{D}}(\Theta_{\mathcal{D}'}) \leqslant \left(1 + \mathcal{O}\left(c_k\,\eta_k\,\epsilon^{2-4/k}\right)\right) err_{\mathcal{D}}(\Theta_{\mathcal{D}}) + \mathcal{O}\left(c_k\,\eta_k\,\epsilon^{2-4/k}\right) err_{\mathcal{D}'}(\Theta_{\mathcal{D}'})$$

5 Robust Regression in Polynomial Time

In this section, we describe an algorithm to compute our robust estimator for linear regression efficiently. We consider a polynomial system that encodes our robust estimator. We then consider a sum-of-squares relaxation of this program and compute an approximately optimal solution for our relaxation. To analyze our algorithm, we consider the dual of the sum-of-squares relaxation and show that the sum-of-squares proof system caputures a variant of our robust identifiability proof.

We begin by recalling notation: let $\mathcal{D}$ be a distribution over $\mathbb{R}^d \times \mathbb{R}$ such that it is (λ_k, k)-certifiably hypercontractive. Let $\mathcal{X} = \{(x_1^*, y_1^*), (x_2^*, y_2^*) \dots (x_n^*, y_n^*)\}$ denote n uncorrupted i.i.d samples from $\mathcal{D}$ and let $\mathcal{X}_\epsilon = .\{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\}$ be an ϵ-corruption of the samples $\mathcal{X}$, drawn from a Robust Regression model, $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$ (Model 1.1). We consider a polynomial system in the variables $\mathcal{X}' = \{(x_1', y_1'), (x_2', y_2') \dots (x_n', y_n')\}$ and $w_1, w_2, \dots w_n \in \{0,1\}^n$ as follows:

$$\mathcal{A}_{\epsilon,\lambda_k} : \left\{ \begin{array}{ll} & \sum_{i\in[n]} w_i = (1-\epsilon)n \\ \forall i \in [n]. & w_i^2 = w_i \\ \forall i \in [n] & w_i(x_i' - x_i) = 0 \\ \forall i \in [n] & w_i(y_i' - y_i) = 0 \\ & \left\langle v, \frac{1}{n}\sum_{i\in[n]} x_i' \left(\langle x_i', \Theta\rangle - y_i\right) \right\rangle^k = 0 \\ \forall r \leqslant k/2 & \frac{1}{n}\sum_{i\in[n]} \langle x_i', v\rangle^{2r} \leqslant \left(\frac{\lambda_r}{n}\sum_{i\in[n]} \langle x_i', v\rangle^2\right)^r \\ \forall r \leqslant k/2 & \frac{1}{n}\sum_{i\in[n]} \left(y_i' - \langle \Theta, x_i'\rangle\right)^{2r} \leqslant \left(\frac{\lambda_r}{n}\sum_{i\in[n]} \left(y_i' - \langle \Theta, x_i'\rangle\right)^2\right)^r \\ \forall r \leqslant k/2 & \mathbb{E}\left[\left(v^\top x_i' \left(y_i' - (x_i')^\top \Theta\right)\right)^{2r}\right] \leqslant \mathcal{O}(\lambda_r^{2r}) \mathbb{E}\left[\langle v, x_i'\rangle^2\right]^r \mathbb{E}\left[\left(y_i' - \langle x_i', \Theta\rangle\right)^2\right]^r \end{array} \right\}$$

We show that optimizing an appropriate convex function subject to the aforementioned constraint system results in an efficiently computable robust estimator for regression, achieving the information-theoretically optimal rate. Formally,

Theorem 5.1 (Robust Regression with Negatively Correlated Moments, Theorem 1.7 restated)**.** *Given $k \in \mathbb{N}$, $\epsilon > 0$ and $n \geqslant n_0$ samples $\mathcal{X}_\epsilon = \{(x_1, y_1), \dots (x_n, y_n)\}$ from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$, where $\mathcal{D}$ is a (λ_k, k)-certifiably hypercontractive distribution over $\mathbb{R}^d \times \mathbb{R}$. Further, $\mathcal{D}$ has certifiable negatively correlated moments. Then, Algorithm 5.3 runs in $n^{\mathcal{O}(k)}$ time and outputs an estimator $\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta]$ such that when $n_0 = \Omega\left((d\log(d))^{\Omega(k)}/\gamma^2\right)$ with probability $1 - 1/\mathrm{poly}(d)$ (over the draw of the input),*

$$\left\|(\Sigma^*)^{1/2}\left(\Theta^* - \tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta]\right)\right\|_2 \leqslant \mathcal{O}\Big(\lambda_k\, \epsilon^{1-1/k} + \lambda_k \gamma\Big)\, err_{\mathcal{D}}(\Theta^*)^{1/2}$$

Further,

$$err_{\mathcal{D}}\left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta]\right) \leqslant \left(1 + \mathcal{O}\Big(\lambda_k^2\, \epsilon^{2-2/k} + \lambda_k^2\, \gamma^2\Big)\right) err_{\mathcal{D}}(\Theta^*).$$

Efficient Estimator for Arbitrary Noise. We note that an argument similar to the one presented for Theorem 5.1 results in a polynomial time estimator when the regression instance does not have negatively correlated moments (definition 1.5), albeit at a slightly worse rate. Formally,

Corollary 5.2 (Robust Regression with Arbitrary Noise). *Consider the hypothesis of Theorem 5.1, without the negatively correlated moments assumption. Then, there exists an algorithm that runs in time $n^{\mathcal{O}(k)}$ outputs an estimator $\tilde{\Theta}$ such that when $n_0 = (d\log(d))^{\Omega(k)}/\gamma^2$, with probability $1 - 1/\mathrm{poly}(d)$ (over the draw of the input),*

$$\left\|(\Sigma^*)^{1/2}\left(\Theta^* - \tilde{\Theta}\right)\right\|_2 \leqslant \mathcal{O}\Big(\lambda_k\,\epsilon^{1-2/k} + c_2\,\eta_2\,\gamma\Big)\,err_{\mathcal{D}}(\Theta^*)^{1/2}$$

Further,

$$err_{\mathcal{D}}\left(\tilde{\Theta}\right) \leqslant \left(1 + \mathcal{O}\Big(\lambda_k^2\,\epsilon^{2-4/k} + \lambda_2^2\,\gamma\Big)\right) err_{\mathcal{D}}(\Theta^*)$$

Algorithm 5.3 (Optimal Robust Regression in Polynomial Time).

Input: n samples $\mathcal{X}_\epsilon$ from the robust regression model $\mathcal{R}_{\mathcal{D}}(\epsilon, \Theta^*, \Sigma^*)$.

Operation:

1. Find a degree-$\mathcal{O}(k)$ pseudo-distribution $\tilde{\zeta}$ satisfying $\mathcal{A}_{\epsilon,\lambda_k}$ and minimizing

$$\min_{w,x',y',\Theta} \tilde{\mathbb{E}}_{\tilde{\zeta}}\left[\left(\frac{1}{n}\sum_{i\in[n]} w_i\left(y_i' - \langle\Theta, x'\rangle\right)^2\right)^k\right]$$

.

2. Round the pseudo-distribution to obtain an estimator $\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta]$.

Output: A vector $\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta]$ such that the recovery guarantee in Theorem 5.1 is satisfied.

At a high level, we simply do not enforce the negatively correlated moments constraint in our polynomial system $\mathcal{A}_{\epsilon,\lambda_k}$ and instead use the SoS Cauchy-Schwarz inequality in our key technical lemma (Lemma 5.5). For completeness, we provide the proof of the SoS lemma in Appendix B.

5.1 Analysis

We begin by observing that we can efficiently optimize the polynomial program above since it admits a compact representation. In particular, $\mathcal{A}_{\epsilon,\lambda_k}$ can be represented as a system of $\mathrm{poly}(n^k)$ constraints in $n^{O(k)}$ variables. We refer the reader to [FKP+19] for a detailed overview on how to efficiently implement the aforementioned constraints.

Lemma 5.4 (Soundness of the Constraint System). *Given $n \geqslant n_0$ samples from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Theta^*, \Sigma)$, with probability at least $1 - 1/\mathrm{poly}(d)$ over the draw of the samples, there exists an assignment for w, x', y' and Θ such that $\mathcal{A}_{\epsilon,\lambda_k}$ is feasible when $n_0 = \Big((d\log(d))^{\Omega(k)}\Big)$.*

Proof. Consider the following assignment: for all $i \in [n]$ the w_i's indicate the set of uncorrupted points in $\mathcal{X}_\epsilon$, i.e. $w_i = 1$ if $(x_i, y_i) = (x_i^*, y_i^*)$, $x_i' = x_i$ and $y_i' = y_i$. Further, $\Theta = \Theta^*$, the true hyperplane. It is easy to see that the first four constraints (intersection constraints) are satisfied.

We observe that the marginal distribution over the covariates and the noise are both (λ_k, k)-certifiably hypercontractive since they are Affine transformations of $\mathcal{D}$ (Fact 3.7). Next, it follows from Fact 3.6, that for $n_0 = \Omega\left(d \log(d)^{\mathcal{O}(k)}\right)$, the uniform distribution over the samples x_i, is $(2\,\lambda_k, k)$-certifiably hypercontractive with probability at least $1 - 1/\mathrm{poly}(d)$. Similarly, the uniform distribution on $y_i - \langle x_i, \Theta^* \rangle$ is $(2\,\lambda_k, k)$-certifiably hypercontractive.

It remains to show that sampling preserves certifiable negatively correlated moments. Recall, since the joint distribution is hypercontractive, by Fact 3.6 we know that there's a degree $\mathcal{O}(k)$ proof of

$$\begin{aligned} \frac{1}{n} \sum_{i \in [n]} \langle v, x_i \rangle^k \left(y_i - \langle x_i, \Theta^* \rangle\right)^k &\leqslant \mathcal{O}\left(\lambda_k^k\right) \left(\frac{1}{n} \sum_{i \in [n]} \langle v, x_i \rangle^2 \left(y_i - \langle x_i, \Theta^* \rangle\right)^2\right)^{k/2} \\ &= \mathcal{O}\left(\lambda_k^k\right) \left(\frac{1}{n} \sum_{i \in [n]} v^\top x_i (x_i)^\top \left(y_i - \langle x_i, \Theta^* \rangle\right)^2 v\right)^{k/2} \end{aligned} \tag{24}$$

It thus suffices to bound the Operator norm of $\frac{1}{n} \sum_{i \in [n]} x_i x_i^\top \left(y_i - \langle x_i, \Theta^* \rangle\right)^2$. It follows from Lemma 3.4 that with probability at least $1 - 1/\mathrm{poly}(d)$,

$$\frac{1}{n} \sum_{i \in [n]} x_i x_i^\top \left(y_i - \langle x_i, \Theta^* \rangle\right)^2 \preceq \mathcal{O}(1) \mathop{\mathbb{E}}_{x,y \sim \mathcal{D}} \left[x x^\top \left(y - \langle x, \Theta^* \rangle\right)^2\right] \tag{25}$$

when $n \geqslant n_0$. Using that $\mathcal{D}$ has negatively correlated moments,

$$\mathop{\mathbb{E}}_{x,y \sim \mathcal{D}} \left[x x^\top \left(y - \langle x, \Theta^* \rangle\right)^2\right] \preceq \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[x x^\top\right] \mathop{\mathbb{E}}_{x,y \sim \mathcal{D}} \left[\left(y - \langle x, \Theta^* \rangle\right)^2\right] \tag{26}$$

Using Lemma 3.4 on $x x^\top$ and $\left(y - \langle x, \Theta^* \rangle\right)^2$, we can bound (26) as follows:

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[x x^\top\right] \mathop{\mathbb{E}}_{x,y \sim \mathcal{D}} \left[\left(y - \langle x, \Theta^* \rangle\right)^2\right] \preceq \mathcal{O}(1)\, \mathbb{E}\left[x_i x_i^\top\right] \left(y_i - \langle x_i, \Theta^* \rangle\right)^2 \tag{27}$$

Combining Equations (25), (26), and (27), and substituting in (24), we have

$$\frac{1}{n} \sum_{i \in [n]} \langle v, x_i \rangle^k \left(y_i - \langle x_i, \Theta^* \rangle\right)^k \leqslant \mathcal{O}\left(\lambda_k^k\right) \left(\frac{1}{n} \sum_{i \in [n]} \langle x_i, v \rangle^2\right)^{\frac{k}{2}} \left(\frac{1}{n} \sum_{i \in [n]} \left(y_i - \langle x_i, \Theta^* \rangle\right)^2\right)^{\frac{k}{2}}$$

which concludes the proof. ∎

Let $\hat{\Sigma}$ be the empirical covariance of the uncorrupted samples $\mathcal{X}$ and let $\hat{\Theta}$ be an optimizer for the empirical loss. Applying Theorem 4.1 with $\mathcal{D}$ being the uniform distribution on the uncorrupted samples $\mathcal{X}$ and $\mathcal{D}'$ being the uniform distribution on x_i', we get

$$\left\| \hat{\Sigma}^{1/2} \left(\Theta - \hat{\Theta}\right) \right\|_2 \leqslant \mathcal{O}\left(\lambda_k\, \epsilon^{1-1/k}\right) \mathrm{err}_{\mathcal{D}}(\Theta^*)^{1/2}$$

Observe, the aforementioned bound is not a polynomial identity and thus cannot be expressed in the SoS framework. Therefore, we provide a low-degree SoS proof of a slightly modified version of the inequality above, that is inspired by our information theoretic identifiability proof in Theorem 4.1.

Lemma 5.5 (Robust Identifiability in SoS)**.** *Consider the hypothesis of Theorem 5.1. Let w, x', y' and Θ be feasible solutions for the polynomial constraint system $\mathcal{A}$. Let $\hat{\Theta} = \arg\min_\Theta \frac{1}{n}\sum_{i\in[n]}(y_i^* - \langle x_i^*, \Theta\rangle)^2$ be the empirical loss minimizer on the uncorrupted samples and let $\hat{\Sigma} = \mathbb{E}\left[x_i^*(x_i^*)^\top\right]$ be the covariance of the uncorrupted samples. Then,*

$$\mathcal{A} \left|\frac{w,x',y',\Theta}{4k}\right. \left\{ \left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{2k} \leqslant 2^{3k}(2\epsilon)^{k-1}\lambda_k^k\,\sigma^{k/2}\left\|\mathbb{E}\left[x_i'(x_i')^\top\right]^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^k \right.$$
$$+ 2^{3k}(2\epsilon)^{k-2}\lambda_k^{2k}\left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{2k}$$
$$\left. + 2^{3k}(2\epsilon)^{k-1}\lambda_k^k\,\mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^2\right]^{k/2}\left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^k \right\}$$

Proof. Consider the empirical covariance of the uncorrupted set given by $\hat{\Sigma} = \mathbb{E}\left[x_i^*(x_i^*)^\top\right]$. Then, using the Substitution Rule, along with SoS Almost Triangle Inequality (Fact 3.18),

$$\left|\frac{\Theta}{2k}\right. \left\{ \left\langle v, \hat{\Sigma}\left(\hat{\Theta}-\Theta\right)\right\rangle^k = \left\langle v, \mathbb{E}\left[x_i^*(x_i^*)^\top\left(\hat{\Theta}-\Theta\right) + x_i^*y_i^* - x_i^*y_i^*\right]\right\rangle^k \right.$$
$$= \left\langle v, \mathbb{E}\left[x_i^*\left(\langle x_i^*, \hat{\Theta}\rangle - y_i^*\right)\right] + \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \tag{28}$$
$$\left. \leqslant 2^k\left\langle v, \mathbb{E}\left[x_i^*\left(\langle x_i^*, \hat{\Theta}\rangle - y_i^*\right)\right]\right\rangle^k + 2^k\left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \right\}$$

Observe, the first term in (28) only consists of constants of the proof system. Since $\hat{\Theta}$ is the minimizer of $\mathbb{E}\left[\left(\langle x_i^*, \Theta\rangle - y_i^*\right)^2\right]$, the gradient condition on the samples (appearing in Equation (12) of the indentifiability proof) implies this term is 0. Therefore, applying the Substitution Rule it suffices to bound the second term.

To this end, we introduce the following auxiliary variables : for all $i \in [n]$, let $w_i' = w_i$ iff the i-th sample is uncorrupted in $\mathcal{X}_\epsilon$, i.e. $x_i = x_i^*$. Then, it is easy to see that $\sum_i w_i' \geqslant (1-2\epsilon)n$. Further, since $\mathcal{A}\left|\frac{w}{2}\right. \left\{(1-w_i'w_i)^2 = (1-w_i'w_i)\right\}$,

$$\mathcal{A}\left|\frac{w}{2}\right. \left\{ \frac{1}{n}\sum_{i\in[n]}(1-w_i'w_i)^2 = \frac{1}{n}\sum_{i\in[n]}(1-w_i'w_i) \leqslant 2\epsilon \right\} \tag{29}$$

The above equation bounds the uncorrupted points in $\mathcal{X}_\epsilon$ that are not indicated by w. Then, using the Substitution Rule, along with the SoS Almost Triangle Inequality (Fact 3.18),

$$\mathcal{A}\left|\frac{\Theta,w'}{2k}\right. \left\{ \left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k = \left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\left(w_i' + 1 - w_i'\right)\right)\right]\right\rangle^k \right.$$
$$= \left\langle v, \mathbb{E}\left[w_i'x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right] + \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k$$
$$\leqslant 2^k\left\langle v, \mathbb{E}\left[w_i'x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k$$
$$\left. + 2^k\left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \right\} \tag{30}$$

Consider the first term of the last inequality in (30). Observe, since $w_i' x_i^* = w_i w_i' x_i'$ and similarly, $w_i' y_i^* = w_i w_i' y_i'$,

$$\mathcal{A} \left|\frac{\Theta, w'}{4}\right. \left\{ \mathbb{E}\left[w_i' x_i^* \left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right] = \mathbb{E}\left[w_i' w_i x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right] \right\}$$

For the sake of brevity, the subsequent statements hold for relevant SoS variables and have degree $O(k)$ proofs. Using the Substitution Rule,

$$\begin{aligned}
\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[w_i' x_i^* \left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k &= \left\langle v, \mathbb{E}\left[w_i' w_i x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k \\
&= \left\langle v, \mathbb{E}\left[x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right] + \mathbb{E}\left[(1 - w_i' w_i) x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k \\
&\leqslant 2^k \left\langle v, \mathbb{E}\left[x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k \\
&\quad + 2^k \left\langle v, \mathbb{E}\left[(1 - w_i' w_i) x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k \Bigg\}
\end{aligned} \tag{31}$$

Observe, the first term in the last inequality above is identically 0, since we enforce the gradient condition on the SoS variables x', y' and Θ. We can then rewrite the second term using linearity of expectation, followed by applying SoS Hölder's Inequality (Fact 3.19) combined with $\mathcal{A} \left|\frac{w}{2}\right. \{(1 - w_i' w_i)^2 = 1 - w_i' w_i\}$ to get

$$\begin{aligned}
\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1 - w_i' w_i) x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k &= \mathbb{E}\left[\left\langle v, (1 - w_i') w_i x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right\rangle\right]^k \\
&= \mathbb{E}\left[(1 - w_i' w_i) \left\langle v, x_i'\right\rangle \left(y_i' - \langle x_i', \Theta\rangle\right)\right]^k \\
&\leqslant \mathbb{E}\left[(1 - w_i' w_i)\right]^{k-1} \mathbb{E}\left[\left\langle v, x_i'\right\rangle^k \left(y_i' - \langle x_i', \Theta\rangle\right)^k\right] \\
&\leqslant (2\epsilon)^{k-1} \mathbb{E}\left[\left\langle v, x_i'\right\rangle^k \left(y_i' - \langle x_i', \Theta\rangle\right)^k\right] \Bigg\}
\end{aligned} \tag{32}$$

where the last inequality follows from Equation (29). Next, we use the certifiable negatively correlated moments constraint with the Substitution Rule,

$$\mathcal{A} \vdash \left\{ \mathbb{E}\left[\left\langle v, x_i'\right\rangle^k \left(y_i' - \langle x_i', \Theta\rangle\right)^k\right] \leqslant \mathcal{O}\left(\lambda_k^k\right) \mathbb{E}\left[\left\langle v, x_i'\right\rangle^2\right]^{\frac{k}{2}} \mathbb{E}\left[\left(y_i' - \langle x_i', \Theta\rangle\right)^2\right]^{\frac{k}{2}} \right\} \tag{33}$$

For brevity, let $\sigma = \mathbb{E}\left[\left(y_i' - \langle x_i', \Theta\rangle\right)^2\right]$. Using the Substitution Rule, plugging Equation (33) back into (32), we get

$$\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1 - w_i') x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k \leqslant (2\epsilon)^{k-1} \lambda_k^k \sigma^{k/2} \left\langle v, \mathbb{E}\left[x_i' (x_i')^\top\right] v\right\rangle^{k/2} \right\} \tag{34}$$

Recall, we have now bounded the first term of the last inequality in (30). Therefore, it remains to bound the second term of the last inequality in (30). Using the Substitution Rule, we have

$$\begin{aligned}\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k &= \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta - \hat{\Theta} + \hat{\Theta}\rangle\right)\right]\right\rangle^k \\ &\leqslant 2^k \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)\right]\right\rangle^k \\ &\quad + 2^k \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(\langle x_i^*, \Theta - \hat{\Theta}\rangle\right)\right]\right\rangle^k \Bigg\} \end{aligned} \tag{35}$$

We again handle each term separately. Observe, the first term when decoupled is a statement about the uncorrupted samples. Therefore, using the SoS Hölder's Inequality (Fact 3.19),

$$\begin{aligned}\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)\right]\right\rangle^k &= \mathbb{E}\left[(1-w_i')\left\langle v, x_i^*\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)\right\rangle\right]^k \\ &\leqslant \mathbb{E}\left[(1-w_i')\right]^{k-1} \mathbb{E}\left[\left\langle v, x_i^*\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)\right\rangle^k\right] \\ &\leqslant (2\epsilon)^{k-1}\, \mathbb{E}\left[\langle v, x_i^*\rangle^k \left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^k\right] \Bigg\} \end{aligned} \tag{36}$$

Observe, the uncorrupted samples have negatively correlated moments, and thus

$$\mathbb{E}\left[\langle v, x_i^*\rangle^k \left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^k\right] \leqslant \mathcal{O}\left(\lambda_k^k\right) \mathbb{E}\left[\langle v, x_i^*\rangle^2\right]^{k/2} \mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^2\right]^{k/2}$$

Then, by the Substitution Rule, we can bound (36) as follows:

$$\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)\right]\right\rangle^k \leqslant (2\epsilon)^{k-1}\lambda_k^k\, \mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^2\right]^{k/2} \langle v, \hat{\Sigma} v\rangle^{k/2} \right\} \tag{37}$$

In order to bound the second term in (35), we use the SoS Hölder's Inequality,

$$\begin{aligned}\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(\langle x_i^*, \Theta - \hat{\Theta}\rangle\right)\right]\right\rangle^k &= \mathbb{E}\left[(1-w_i')^{k-2}\left\langle v, x_i^*\left(\langle x_i^*, \Theta - \hat{\Theta}\rangle\right)\right\rangle\right] \\ &\leqslant \mathbb{E}\left[1-w_i'\right]^{k-2} \mathbb{E}\left[\left(v^\top x_i^* (x_i^*)^\top (\Theta - \hat{\Theta})\right)^{\frac{k}{2}}\right]^2 \\ &\leqslant (2\epsilon)^{k-2}\, \mathbb{E}\left[\left(v^\top x_i^* (x_i^*)^\top (\Theta - \hat{\Theta})\right)^{\frac{k}{2}}\right]^2 \Bigg\} \end{aligned} \tag{38}$$

Combining the bounds obtained in (37) and (38), we can restate Equation (35) as follows

$$\begin{aligned}\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k &\leqslant 2^k(2\epsilon)^{k-1}\lambda_k^k\, \mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^2\right]^{k/2} \langle v, \hat{\Sigma} v\rangle^{k/2} \\ &\quad + 2^k(2\epsilon)^{k-2}\, \mathbb{E}\left[\left(v^\top x_i^* (x_i^*)^\top (\Theta - \hat{\Theta})\right)^{k/2}\right]^2 \Bigg\} \end{aligned} \tag{39}$$

Combining (39) with (34), we obtain an upper bound for the last inequality in Equation (30). Therefore, using the Substitution Rule, we obtain

$$\begin{aligned}\mathcal{A} \vdash \Bigg\{ \left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \leqslant 2^k(2\epsilon)^{k-1}\lambda_k^k\,\sigma^{k/2}\left\langle v, \mathbb{E}\left[x_i'(x_i')^\top\right]v\right\rangle^{k/2} \\ + 2^{2k}(2\epsilon)^{k-2}\,\mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta-\hat{\Theta})\right)^{\frac{k}{2}}\right]^2 \\ + 2^{2k}(2\epsilon)^{k-1}\lambda_k^k\,\mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^2\right]^{k/2}\langle v, \hat{\Sigma}v\rangle^{k/2}\Bigg\}\end{aligned} \tag{40}$$

Recall, an upper bound on Equation (28) suffices to obtain an upper bound on $\left\langle v, \hat{\Sigma}\left(\hat{\Theta}-\Theta\right)\right\rangle$ as follows:

$$\begin{aligned}\mathcal{A} \vdash \Bigg\{ \left\langle v, \hat{\Sigma}\left(\hat{\Theta}-\Theta\right)\right\rangle^k \leqslant 2^{2k}(2\epsilon)^{k-1}\lambda_k^k\,\sigma^{k/2}\left\langle v, \mathbb{E}\left[x_i'(x_i')^\top\right]v\right\rangle^{k/2} \\ + 2^{3k}(2\epsilon)^{k-2}\,\mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta-\hat{\Theta})\right)^{\frac{k}{2}}\right]^2 \\ + 2^{3k}(2\epsilon)^{k-1}\lambda_k^k\,\mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^2\right]^{k/2}\langle v, \hat{\Sigma}v\rangle^{k/2}\Bigg\}\end{aligned} \tag{41}$$

Consider the substitution $v \mapsto \left(\hat{\Theta}-\Theta\right)$. Then,

$$\begin{aligned}\left\langle v, \hat{\Sigma}\left(\hat{\Theta}-\Theta\right)\right\rangle^k &= \left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{2k} \\ \left\langle v, \mathbb{E}\left[x_i'(x_i')^\top\right]v\right\rangle^{k/2} &= \left\|\mathbb{E}\left[x_i'(x_i')^\top\right]^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{k} \\ \mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta-\hat{\Theta})\right)^{\frac{k}{2}}\right]^2 &= \mathbb{E}\left[\langle x_i^*, \hat{\Theta}-\Theta\rangle^k\right]^2 \leqslant \lambda_k^{2k}\left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{2k} \\ \langle v, \hat{\Sigma}v\rangle^{k/2} &= \left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{k}\end{aligned}$$

Combining the above with (41), we conclude

$$\begin{aligned}\mathcal{A} \vdash \Bigg\{ \left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{2k} \leqslant 2^{3k}(2\epsilon)^{k-1}\lambda_k^k\,\sigma^{k/2}\left\|\mathbb{E}\left[x_i'(x_i')^\top\right]^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{k} \\ + 2^{3k}(2\epsilon)^{k-2}\lambda_k^{2k}\left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{2k} \\ + 2^{3k}(2\epsilon)^{k-1}\lambda_k^k\,\mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^2\right]^{k/2}\left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{k}\Bigg\}\end{aligned} \tag{42}$$

□

Next, we relate the covariance of the samples indicated by w to the covariance on the uncorrupted points. Observe, a real world proof of this follows simply from Fact 3.3.

Lemma 5.6 (Bounding Sample Covariance). *Consider the hypothesis of Theorem 5.1. Let w, x', y' and Θ be feasible solutions for the polynomial constraint system $\mathcal{A}$. Then, for $\delta \leqslant \mathcal{O}(\lambda_k \epsilon^{1-1/k}) < 1$,*

$$\mathcal{A} \left|\frac{w,x'}{2k}\right. \left\{ \left\langle v, \mathbb{E}\left[x_i'(x_i')^\top\right] v \right\rangle^{k/2} \leqslant \left(1 + \mathcal{O}\left(\delta^{k/2}\right)\right) \left\langle v, \hat{\Sigma} v \right\rangle^{k/2} \right\}$$

Proof. Our proof closely follows Lemma 4.5 in [KS17c]. For $i \in [n]$, let z_i be an indicator variable such $z_i(x_i^* - x_i') = 0$. Observe, there exists an assignment to z_i such that $\sum_{i\in[n]} z_i = (1-\epsilon)n$, since at most ϵn points were corrupted. Further, $z_i^2 = z_i$ and $\frac{1}{n} z_i = \epsilon$. Then, using the Substitution Rule,

$$\begin{aligned}
\mathcal{A} \left|\frac{w,x'}{2k}\right. \left\{ \left\langle v, \left(\mathbb{E}\left[x_i'(x_i')^\top\right] - \hat{\Sigma}\right) v \right\rangle^k \right. &= \left\langle v, \mathbb{E}\left[(1+z_i-z_i)\left(x_i'(x_i')^\top - x_i^*(x_i^*)^\top\right)\right] v \right\rangle^k \\
&= \mathbb{E}\left[(1-z_i)\left\langle v, \left(x_i'(x_i')^\top - x_i^*(x_i^*)^\top\right), v\right\rangle\right]^k \\
&\leqslant \epsilon^{k-2} \cdot \mathbb{E}\left[\left(\left\langle v, x_i'\right\rangle^2 - \left\langle v, x_i^*\right\rangle^2\right)^{k/2}\right]^2 \\
&\leqslant \epsilon^{k-2}\, \mathbb{E}\left[2^{k/2}\left\langle v, x_i'\right\rangle^k + 2^{k/2}\left\langle v, x_i^*\right\rangle^k\right]^2 \\
&\leqslant \epsilon^{k-2} 2^k \left(c_k^k\ \mathbb{E}\left[\left\langle v, x_i'\right\rangle^2\right]^{k/2} + \lambda_k^k\ \mathbb{E}\left[\left\langle v, x_i^*\right\rangle^2\right]^{k/2}\right)^2 \left.\right\}
\end{aligned} \tag{43}$$

where the first inequality follows from applying the SoS Hölder's Inequality, the second follows from the SoS Almost Triangle Inequality and the third inequality follows from certifiable hypercontractivity of the SoS variables and the uncorrupted samples. Using the SoS Almost Triangle Inequality again, we have

$$\mathcal{A} \vdash \left\{ \left(c_k^k\ \mathbb{E}\left[\left\langle v, x_i'\right\rangle^2\right]^{k/2} + \lambda_k^k\ \mathbb{E}\left[\left\langle v, x_i^*\right\rangle^2\right]^{k/2}\right)^2 \leqslant \lambda_k^{2k}\, 2^2 \left(\left\langle v, \mathbb{E}\left[x_i'(x_i')^\top v\right]\right\rangle^k + \left\langle v, \hat{\Sigma} v\right\rangle^k\right) \right\} \tag{44}$$

Combining Equations 43, 44, we obtain

$$\mathcal{A} \vdash \left\{ \left\langle v, \left(\mathbb{E}\left[x_i'(x_i')^\top\right] - \hat{\Sigma}\right) v \right\rangle^k \leqslant \epsilon^{k-2}\ \lambda_k^{2k}\ 2^{k+2} \left\langle v, \left(\mathbb{E}\left[x_i'(x_i')^\top\right] + \hat{\Sigma}\right) v \right\rangle^k \right\} \tag{45}$$

Using Lemma A.4 from [KS17c], rearranging and setting $k = k/2$ yields the claim. □

Lemma 5.7 (Rounding). *Consider the hypothesis of Theorem 5.1. Let $\hat{\Theta} = \arg\min_\Theta \frac{1}{n}\sum_{i\in[n]}(y_i^* - \langle x_i^*, \Theta\rangle)^2$ be the empirical loss minimizer on the uncorrupted samples. Then,*

$$\left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta} - \tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta]\right)\right\|_2 \leqslant \mathcal{O}\left(\epsilon^{1-\frac{1}{k}}\ \lambda_k\right) \left(\tilde{\mathbb{E}}_{\tilde{\zeta}}\left[\mathbb{E}\left[\left(y_i' - \left\langle x_i', \Theta\right\rangle\right)^2\right]^k\right]^{\frac{1}{2k}} + \mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^2\right]^{\frac{1}{2}}\right)$$

Proof. Observe, combining Lemma 5.5 and Lemma 5.6, we obtain

$$\mathcal{A} \vdash \left\{ \left\| \hat{\Sigma}^{1/2} \left(\hat{\Theta} - \Theta \right) \right\|_2^{2k} \leqslant \mathcal{O}\left(\frac{2^{3k} \epsilon^{k-1} \lambda_k^k}{1 + 2^{3k} (2\epsilon)^{k-2} \lambda_k^{2k}} \right) \left\| \hat{\Sigma}^{1/2} \left(\hat{\Theta} - \Theta \right) \right\|_2^k \left(\mathbb{E} \left[\left(y_i' - \left\langle x_i', \Theta \right\rangle \right)^2 \right]^{\frac{k}{2}} + \mathbb{E} \left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right)^2 \right]^{\frac{k}{2}} \right) \right\} \tag{46}$$

Using Cancellation within SoS (Fact 3.20) along with the SoS Almost Triangle Inequality, we can conclude

$$\mathcal{A} \vdash \left\{ \left\| \hat{\Sigma}^{1/2} \left(\hat{\Theta} - \Theta \right) \right\|_2^{2k} \leqslant \mathcal{O}\left(2^{3k} \epsilon^{k-1} \lambda_k^k \right)^2 \left(\mathbb{E} \left[\left(y_i' - \left\langle x_i', \Theta \right\rangle \right)^2 \right]^k + \mathbb{E} \left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right)^2 \right]^k \right) \right\} \tag{47}$$

Recall, $\tilde{\zeta}$ is a degree-$\mathcal{O}(k)$ pseudo-expectation satisfying $\mathcal{A}$. Therefore, it follows from Fact 3.15 along with Equation 46,

$$\tilde{\mathbb{E}}_{\tilde{\zeta}} \left[\left\| \hat{\Sigma}^{\frac{1}{2}} \left(\hat{\Theta} - \Theta \right) \right\|_2^{2k} \right] \leqslant \mathcal{O}\left(2^{4k} \epsilon^{k-1} \lambda_k^k \right)^2 \left(\tilde{\mathbb{E}}_{\tilde{\zeta}} \left[\mathbb{E} \left[\left(y_i' - \left\langle x_i', \Theta \right\rangle \right)^2 \right]^k \right] + \mathbb{E} \left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right)^2 \right]^k \right) \tag{48}$$

Further, using Fact 3.14, we have $\left\| \hat{\Sigma}^{\frac{1}{2}} \left(\hat{\Theta} - \tilde{\mathbb{E}}_{\tilde{\zeta}} \Theta \right) \right\|_2^{2k} \leqslant \tilde{\mathbb{E}}_{\tilde{\zeta}} \left[\left\| \hat{\Sigma}^{\frac{1}{2}} \left(\hat{\Theta} - \Theta \right) \right\|_2^{2k} \right]$. Substituting above and taking the $(1/2k)$-th root,

$$\begin{aligned} \left\| \hat{\Sigma}^{\frac{1}{2}} \left(\hat{\Theta} - \tilde{\mathbb{E}}_{\tilde{\zeta}} [\Theta] \right) \right\|_2 &\leqslant \mathcal{O}\left(\epsilon^{1-\frac{1}{k}} \lambda_k \right) \left(\tilde{\mathbb{E}}_{\tilde{\zeta}} \left[\mathbb{E} \left[\left(y_i' - \left\langle x_i', \Theta \right\rangle \right)^2 \right]^k \right] + \mathbb{E} \left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right)^2 \right]^k \right)^{1/2k} \\ &\leqslant \mathcal{O}\left(\epsilon^{1-\frac{1}{k}} \lambda_k \right) \left(\tilde{\mathbb{E}}_{\tilde{\zeta}} \left[\mathbb{E} \left[\left(y_i' - \left\langle x_i', \Theta \right\rangle \right)^2 \right]^k \right]^{\frac{1}{2k}} + \mathbb{E} \left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right)^2 \right]^{\frac{1}{2}} \right) \end{aligned} \tag{49}$$

which concludes the proof. □

Lemma 5.8 (Bounding Optimization and Generalization Error)**.** *Under the hypothesis of Theorem 5.1,*

1. $\tilde{\mathbb{E}}_{\tilde{\zeta}} \left[\mathbb{E} \left[\left(y_i' - \left\langle x_i', \Theta \right\rangle \right)^2 \right]^k \right]^{\frac{1}{2k}} \leqslant \mathbb{E} \left[y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle^2 \right]^{\frac{1}{2}}$, *and*
2. *For any* $\zeta > 0$, *if* $n \geqslant n_0$, *such that* $n_0 = \Omega\left(\max\{c_4 d/\zeta^2, d^{\mathcal{O}(k)}\} \right)$, *with probability at least* $1 - 1/\mathrm{poly}(d)$, $\mathbb{E} \left[y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle^2 \right]^{\frac{1}{2}} \leqslant (1 + \zeta) \, \mathbb{E}_{x,y \sim \mathcal{D}} \left[y - \langle x, \Theta^* \rangle^2 \right]^{\frac{1}{2}}$.

Proof. We exhibit a degree-$\mathcal{O}(k)$ pseudo-distribution $\hat{\zeta}$ such that it is supported on a point mass and attains objective value at most $\mathbb{E} \left[y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle^2 \right]^{\frac{1}{2}}$. Since our objective function minimizes

over all degree-$\mathcal{O}(k)$ pseudo-distributions, the resulting objective value w.r.t. $\tilde{\zeta}$ can only be better. Let $\hat{\zeta}$ be the pseudo-distribution supported on $(w, x^*, y^*, \hat{\Theta})$ such that $w_i = 1$ if $x_i = x_i^*$ (i.e. the i-th sample is not corrupted.) It follows from $n \geqslant n_0$ and Lemma 5.4 that this assignment satisfies the constraint system $\mathcal{A}_{\epsilon,\lambda_k}$. Then, the objective value satisfies

$$\tilde{\mathbb{E}}_{\tilde{\zeta}}\left[\mathbb{E}\left[\left(y_i' - \langle x_i', \Theta\rangle\right)^2\right]^k\right] \leqslant \tilde{\mathbb{E}}_{\hat{\zeta}}\left[\mathbb{E}\left[\left(y_i' - \langle x_i', \Theta\rangle\right)^2\right]^k\right] = \mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^2\right]^k \tag{50}$$

Taking $(1/2k)$-th roots yields the first claim.

To bound the second claim, let $\mathcal{U}$ be the uniform distribution on the uncorrupted samples, x_i^*, y_i^*. Observe, by optimality of $\hat{\Theta}$ on the uncorrupted samples, $\mathrm{err}_{\mathcal{U}}(\hat{\Theta}) \leqslant \mathrm{err}_{\mathcal{U}}(\Theta^*)$. Consider the random variable $z_i = \left(y_i^* - \langle x_i^*, \Theta^*\rangle\right)^2 - \mathbb{E}_{x,y\sim\mathcal{D}}\left[\left(y - \langle x, \Theta^*\rangle\right)^2\right]$. Since $\mathbb{E}[z_i] = 0$, we apply Chebyschev's inequality to obtain

$$\begin{aligned}\mathbf{Pr}\left[\frac{1}{n}\sum_{i\in[n]} z_i \geqslant \zeta\right] = \frac{\mathbb{E}[z_1^2]}{\zeta^2 n} &\leqslant \frac{\mathbb{E}\left[\left(y - \langle x, \Theta\rangle\right)^4\right]}{\zeta^2 n} \\ &\leqslant c_4 \frac{\mathrm{err}_{\mathcal{D}}(\Theta^*)^2}{n\zeta^2}\end{aligned}$$

Therefore, with probability at least $1 - \delta$,

$$\mathrm{err}_{\mathcal{U}}(\hat{\Theta}) \leqslant \left(1 + \sqrt{\frac{c_4}{n\delta}}\right)\mathrm{err}_{\mathcal{D}}(\Theta^*)$$

Therefore, setting $n = \Omega(c_4 d/\zeta^2)$, it follows that with probability $1 - 1/\mathrm{poly}(d)$, for any $\zeta > 0$,

$$\mathrm{err}_{\mathcal{U}}(\hat{\Theta}) \leqslant (1 + \zeta)\,\mathrm{err}_{\mathcal{D}}(\Theta^*)$$

Taking square-roots concludes the proof. □

Proof of Theorem 5.1. Given $n \geqslant n_0$ samples, it follows from Lemma 5.4, that with probability $1 - 1/\mathrm{poly}(d)$, the constraint system $\mathcal{A}_{\epsilon,\lambda_k}$ is feasible. Let ξ_1 be the event that the system is feasible and condition on it. Then, it follows from Lemma 5.7 and Lemma 5.8, with probability $1 - 1/\mathrm{poly}(d)$,

$$\left\|\hat{\Sigma}^{1/2}\left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta] - \hat{\Theta}\right)\right\|_2 \leqslant \mathcal{O}\Big(\lambda_k\,\epsilon^{1-1/k}\Big)\,\mathrm{err}_{\mathcal{D}}(\Theta^*)^{1/2} \tag{51}$$

Let ξ_2 be the event that (51) holds and condition on it. It then follows from Fact 3.2, with probability $1 - 1/\mathrm{poly}(d)$,

$$\left\|(\Sigma^*)^{1/2}\left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta] - \hat{\Theta}\right)\right\|_2 \leqslant \mathcal{O}\Big(\lambda_k\,\epsilon^{1-1/k}\Big)\,\mathrm{err}_{\mathcal{D}}(\Theta^*)^{1/2} \tag{52}$$

Let ξ_2 be the event that (52) holds and condition on it. It remains to relate the hyperplanes $\hat{\Theta}$ and Θ^*. By reverse triangle inequality,

$$\left\|(\Sigma^*)^{1/2}\left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta] - \Theta^*\right)\right\|_2 - \left\|(\Sigma^*)^{1/2}\left(\Theta^* - \hat{\Theta}\right)\right\|_2 \leqslant \left\|(\Sigma^*)^{1/2}\left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta] - \hat{\Theta}\right)\right\|_2$$

Using normal equations, we have $\hat{\Theta} = \hat{\Sigma}^{-1}\,\mathbb{E}[x_i y_i]$ and $\Theta^* = (\Sigma^*)^{-1}\,\mathbb{E}[xy]$. Since $\hat{\Sigma} \preceq (1 + 0.01)\Sigma^*$,

$$\begin{aligned}\left\|(\Sigma^*)^{1/2}\left(\Theta^* - \hat{\Theta}\right)\right\|_2 &= \left\|(\Sigma^*)^{1/2}\left(\hat{\Sigma}^{-1}\hat{\Sigma}\Theta^* - \hat{\Sigma}^{-1}\,\mathbb{E}\left[x_i y_i\right]\right)\right\|_2 \\ &= \left\|(\Sigma^*)^{1/2}\,\hat{\Sigma}^{-1}\left(\mathbb{E}\left[x_i\left(y_i - x_i^\top\Theta^*\right)\right]\right)\right\|_2 \\ &\leqslant 1.01\left\|\mathbb{E}\left[(\Sigma^*)^{-1/2}\,x_i\left(y_i - x_i^\top\Theta^*\right)\right]\right\|_2\end{aligned} \tag{53}$$

By Jensen's inequality

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i\in[n]}(\Sigma^*)^{-1/2}\,x_i\left(y_i - x_i^\top\Theta^*\right)\right\|_2\right] \leqslant \sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i\in[n]}(\Sigma^*)^{-1/2}\,x_i\left(y_i - x_i^\top\Theta^*\right)\right\|_2^2\right]}$$

Let $z_i = \sum_{i\in[n]}(\Sigma^*)^{-1/2}\,x_i\left(y_i - x_i^\top\Theta^*\right)$. Let $\left(\sum_{i\in[n]} z_i\right)_1$ denote the first coordinate of the vector. We bound the expectation of this coordinate as follows:

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i\in[n]} z_i\right)_1^2\right] &= \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i,i'\in[n]}\left((\Sigma^*)^{-1}x_i x_{i'}\right)_1\left(y_i - x_i^\top\Theta^*\right)\left(y_{i'} - x_{i'}^\top\Theta^*\right)\right] \\ &= \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i\in[n]}\left((\Sigma^*)^{-1}x_i^2\right)_1\left(y_i - x_i^\top\Theta^*\right)^2\right] \\ &= \frac{1}{n}\,\mathbb{E}\left[(\Sigma^*)^{-1}(x)_1^2\left(y - x^\top\Theta^*\right)\right]\end{aligned} \tag{54}$$

where the second equality follows from independence of the samples. Using negatively correlated moments, we have

$$\mathbb{E}\left[(\Sigma^*)^{-1}(x)_1^2\left(y - x^\top\Theta^*\right)^2\right] \leqslant \mathbb{E}\left[(\Sigma^*)^{-1}(x)_1^2\right]\mathbb{E}\left[\left(y - x^\top\Theta^*\right)^2\right]$$

Setting $v = (\Sigma^*)^{1/2}e_1$ and using Hypercontractivity of the covariates and the noise in the above equation,

$$\mathbb{E}\left[\Sigma^{-1}(x)_1^2\right]\mathbb{E}\left[\left(y - x^\top\Theta^*\right)^2\right] \leqslant \mathcal{O}(c_2^2\,\eta_2^2)\,\mathrm{err}_{\mathcal{D}}(\Theta^*) \tag{55}$$

Summing over the coordinates, and combining (54), (55), we obtain

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i\in[n]}(\Sigma^*)^{-1/2}\,x_i\left(y_i - x_i^\top\Theta^*\right)\right\|_2\right] \leqslant \mathcal{O}(c_2\eta_2)\sqrt{\frac{d\,\mathrm{err}_{\mathcal{D}}(\Theta^*)}{n}} \tag{56}$$

Applying Chebyschev's Inequality , with probability $1-\delta$

$$\left\|(\Sigma^*)^{1/2}\left(\Theta^* - \mathbf{E}_{\tilde{\zeta}}[\Theta]\right)\right\|_2 \leqslant \mathcal{O}\left(\lambda_k\,\epsilon^{1-1/k} + c_2\,\eta_2\sqrt{\frac{d}{\delta n}}\right)\mathrm{err}_{\mathcal{D}}(\Theta^*)^{1/2}$$

Since $n \geqslant n_0$, we can simplify the above bound and obtain the claim.

The running time of our algorithm is clearly dominated by computing a degree-$\mathcal{O}(k)$ pseudo-distribution satisfying $\mathcal{A}_{\epsilon,\lambda_k}$. Given that our constraint system consists of $O(n)$ variables and $\mathrm{poly}(n)$ constraints, it follows from Fact 3.12 that the pseudo-distribution $\tilde{\zeta}$ can be computed in $n^{\mathcal{O}(k)}$ time.

□

6 Lower bounds

In this section, we present information-theoretic lower bounds on the rate of convergence of parameter estimation and least-squares error for robust regression. Our constructions proceed by demonstrating two distributions over regression instances that are ϵ-close in total variation distance and the marginal distribution over the covariates is hypercontractive, yet the true hyperplanes are $f(\epsilon)$-far in scaled ℓ_2 distance.

6.1 True Linear Model

Consider the setting where there exists an optimal hyperplane Θ^* that is used to generate the data, with the addition of independent noise added to each sample, i.e.

$$y = \langle x, \Theta^* \rangle + \omega,$$

where ω is independent of x. Further, we assume that covariates and noise are hypercontractive. In this setting, Theorem 4.1 implies that we can recover a hyperplane close to Θ^* at a rate proportional to $\epsilon^{1-1/k}$. We show that this dependence is tight for $k = 4$. We note that independent noise is a special case of the distribution having negatively correlated moments.

Theorem 6.1 (True Linear Model Lower Bound, Theorem 1.9 restated)**.** *For any $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathbb{R}^2 \times \mathbb{R}$ such that the marginal distribution over $\mathbb{R}^2$ has covariance Σ and is (c_k, k)-hypercontractive yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega\left(\sqrt{c_k}\,\sigma\,\epsilon^{1-1/k}\right)$, where Θ_1, Θ_2 be the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, $\sigma = \max(err_{\mathcal{D}_1}(\Theta_1), err_{\mathcal{D}_2}(\Theta_2)) < 1/\epsilon^{1/k}$ and the noise ω is uniform over $[-\sigma, \sigma]$.*

Proof. We construct a 2-dimensional instance where the marginal distribution over covariates is identical for $\mathcal{D}_1$ and $\mathcal{D}_2$. The pdf is given as follows: for $q \in \{1, 2\}$ on the first coordinate, x_1,

$$\mathcal{D}_q(x_1) = \begin{cases} 1/2, & \text{if } x_1 \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$$

and on the second coordinate, x_2,

$$\mathcal{D}_q(x_2) = \begin{cases} \epsilon/2, & \text{if } x_2 \in \{-1/\epsilon^{1/k}, 1/\epsilon^{1/k}\} \\ \frac{1-\epsilon}{2\epsilon\sigma} & \text{if } x_2 \in [-\epsilon\sigma, \epsilon\sigma] \\ 0 & \text{otherwise} \end{cases}$$

Next, we set $\Theta_1 = (1, 1)$, $\Theta_2 = (1, -1)$ and ω to be uniform over $[-\sigma, \sigma]$. Therefore,

$$\begin{aligned} \mathcal{D}_1(y \mid (x_1, x_2)) &= x_1 + x_2 + \omega \quad \text{and} \\ \mathcal{D}_2(y \mid (x_1, x_2)) &= x_1 - x_2 + \omega \end{aligned} \tag{57}$$

Observe, $\mathbb{E}\left[x_1^k\right] = \int_{-1}^{1} x^k/2 = 1/(k+1)$ and $\mathbb{E}\left[x_1^2\right] = \int_{-1}^{1} x^2/2 = 1/3$. Further,

$$\mathbb{E}[x_2^k] = \frac{(1-\epsilon)}{\epsilon\sigma}\frac{(\epsilon\sigma)^{k+1}}{k+1} + \epsilon \cdot \left(\frac{1}{\epsilon^{1/k}}\right)^k = 1 + \frac{(1-\epsilon)}{(k+1)}(\epsilon\sigma)^k$$
$$\mathbb{E}[x_2^2] = \frac{(1-\epsilon)}{3\epsilon\sigma}(\epsilon\sigma)^3 + \epsilon \cdot \left(\frac{1}{\epsilon^{1/k}}\right)^2 = \epsilon^{1-2/k} + \frac{1-\epsilon}{3}(\epsilon\sigma)^2$$

Observe, $\mathbb{E}[x_2^k] \leqslant (1/(c\epsilon^{k/2-1}))\ \mathbb{E}[x_2^2]^{k/2}$, for a fixed constant c. Then, for any unit vector v,

$$\begin{aligned}\mathbb{E}\left[\langle x, v\rangle^k\right] \leqslant \mathbb{E}\left[(2x_1v_1)^k + (2x_2v_2)^k\right] &\leqslant c_k^{k/2}\ \left(\mathbb{E}\left[(x_1v)^2\right]^{k/2} + \mathbb{E}\left[(x_2v)^2\right]^{k/2}\right)\\ &\leqslant c_k^{k/2}\ \mathbb{E}\left[\langle x, v\rangle^2\right]^{k/2}\end{aligned}$$

where $c_k^{k/2} = 2^k/c\epsilon^{k/2-1}$. Therefore, $\mathcal{D}_1, \mathcal{D}_2$ are (c_k, k)-hypercontractive over $\mathbb{R}^2$. Next, we compute the TV distance between the two distributions.

$$\begin{aligned}d_{\mathsf{TV}}\left(\mathcal{D}_1, \mathcal{D}_2\right) &= \frac{1}{2}\int_{\mathbb{R}^2\times\mathbb{R}} |\mathcal{D}_1(x_1, x_2, y) - \mathcal{D}_2(x_1, x_2, y)|\\ &= \frac{1}{2}\int_{\mathbb{R}^2} \mathcal{D}_1(x_1, x_2)\int_{\mathbb{R}} |\mathcal{D}_1(y \mid (x_1, x_2)) - \mathcal{D}_2(y \mid (x_1, x_2))|\end{aligned} \tag{58}$$

where the last equality follows from the definition of conditional probability. It follows from Equation (57) that $\mathcal{D}_1(y \mid (x_1, x_2)) = \mathcal{U}(x_1 + x_2 - \sigma, x_1 + x_2 + \sigma)$ and $\mathcal{D}_2(y \mid (x_1, x_2)) = \mathcal{U}(x_1 - x_2 - \sigma, x_1 - x_2 + \sigma)$. If $|x_2| \geqslant \sigma$ the intervals are disjoint and $|\mathcal{D}_1(y \mid (x_1, x_2)) - \mathcal{D}_2(y \mid (x_1, x_2))| = 2$. If $|x_2| < \sigma$, then two symmetric non-intersecting regions have mass $2|x_2|/2\sigma$ and the intersection region contributes 0. Therefore, $|\mathcal{D}_1(y \mid (x_1, x_2)) - \mathcal{D}_2(y \mid (x_1, x_2))| = 2|x_2|/\sigma$ and (58) can be evaluated as

$$\begin{aligned}d_{\mathsf{TV}}(\mathcal{D}_1, \mathcal{D}_2) &= \frac{1}{2}\int_{\mathbb{R}} 2\mathbb{I}\left\{|x_2| \geqslant \sigma\right\} + \frac{2|x_2|}{\sigma}\mathbb{I}\left\{|x_2| < \sigma\right\}\\ &= \mathbf{Pr}\left[|x_2| \geqslant \sigma\right] + \frac{1}{\sigma}\mathop{\mathbb{E}}_{x_2\sim\mathcal{D}_1}\left[|x_2|\mathbb{I}\left\{|x_2| < \sigma\right\}\right]\\ &= 2\epsilon\end{aligned}$$

Finally, we lower bound the parameter distance. Since the coordinates are independent, Σ is a diagonal matrix with $\Sigma_{1,1} = \mathbb{E}\left[x_1^2\right] = 1/3$ and $\Sigma_{2,2} = \mathbb{E}\left[x_2^2\right] = \epsilon^{1-2/k} + (\epsilon\sigma)^2/3$. Further, $\Theta_1 - \Theta_2 = (0, 2)$. Thus, $\left\|\Sigma^{1/2}\left(\Theta_1 - \Theta_2\right)\right\|_2 = 2\Sigma_{2,2}^{1/2} \geqslant 2\epsilon^{1/2-1/k}$. For any $\sigma < 1/\epsilon^{1/k}$,

$$\begin{aligned}\left\|\Sigma^{1/2}\left(\Theta_1 - \Theta_2\right)\right\|_2 \geqslant 2\,\epsilon^{1/2-1/k} &> 2\,\sigma\,\epsilon^{1/2}\\ &\geqslant 2\,\sqrt{c_k}\,\sigma\,\epsilon^{1-1/k}\end{aligned}$$

which concludes the proof.

□

6.2 Agnostic Model

Next, consider the setting where we simply observe samples from $(x, y) \sim \mathcal{D}$, and our goal is to return is to return the minimizer of the squared error, given by $\Theta^* = \mathbb{E}\left[xx^\top\right]^{-1}\mathbb{E}\left[xy\right]$. Here, the

distribution of the noise is allowed to depend on the covariates arbitrarily. We further assume the noise is hypercontractive and obtain a lower bound proportional to $\epsilon^{1-2/k}$ for recovering an estimator close to Θ^*. This matches the upper boundd obtained in Corollary 4.2.

Theorem 6.2 (Agnostic Model Lower Bound, Theorem 1.11 restated)**.** *For any $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathbb{R}^2 \times \mathbb{R}$ such that the marginal distribution over $\mathbb{R}^2$ has covariance Σ and is (c_k, k)-hypercontractive yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega\left(\sqrt{c_k}\,\sigma\,\epsilon^{1-2/k}\right)$, where Θ_1, Θ_2 be the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, $\sigma = \max(err_{\mathcal{D}_1}(\Theta_1), err_{\mathcal{D}_2}(\Theta_2)) < 1/\epsilon^{1/k}$ and the noise is a function of the marginal distribution of $\mathbb{R}^2$.*

Proof. We provide a proof for the special case of $k = 4$. The same proof extends to general k. We again construct a 2-dimensional instance where the marginal distribution over covariates is identical for $\mathcal{D}_1$ and $\mathcal{D}_2$. The pdf is given as follows: for $q \in \{1, 2\}$ on the first coordinate, x_1,

$$\mathcal{D}_q(x_1) = \begin{cases} 1/2, & \text{if } x_1 \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$$

and on the second coordinate, x_2,

$$\mathcal{D}_q(x_2) = \begin{cases} \epsilon/2, & \text{if } x_2 \in \{-1/\epsilon^{1/4}, 1/\epsilon^{1/4}\} \\ \frac{1-\epsilon}{2} & \text{if } x_2 \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$$

Observe, $\mathbb{E}\left[x_1^4\right] = 1/5$ and $\mathbb{E}\left[x_1^2\right] = 1/3$. Similarly, $\mathbb{E}\left[x_2^4\right] = 1 + (1-\epsilon)/5$ and $\mathbb{E}\left[x_2^2\right] = \sqrt{\epsilon} + (1-\epsilon)/3$. Therefore, the marginal distribution over $\mathbb{R}^2$ is $(c, 4)$-hypercontractive for a fixed constant c. Next, let

$$\begin{aligned} \mathcal{D}_1(y \mid (x_1, x_2)) &= x_2 \qquad \text{and} \\ \mathcal{D}_2(y \mid (x_1, x_2)) &= \begin{cases} 0 & \text{if } |x_2| = 1/\epsilon^{1/4} \\ x_2 & \text{otherwise} \end{cases} \end{aligned} \tag{59}$$

Then,

$$\begin{aligned} d_{\mathsf{TV}}(\mathcal{D}_1, \mathcal{D}_2) &= \frac{1}{2} \int_{\mathbb{R}^2} \mathcal{D}_1(x_1, x_2) \int_{\mathbb{R}} |\mathcal{D}_1(y \mid (x_1, x_2)) - \mathcal{D}_2(y \mid (x_1, x_2))| \\ &= \frac{1}{2} \int_{\mathbb{R}} |x_2| \mathbb{I}\left\{|x_2| = 1/\epsilon^{1/4}\right\} \\ &= \epsilon \end{aligned}$$

Since the coordinates over $\mathbb{R}^2$ are independent the covariance matrix Σ is diagonal, such that $\Sigma_{1,1} = \mathbb{E}\left[x_1^2\right] = 1/3$ and $\Sigma_{2,2} = \mathbb{E}\left[x_2^2\right] = \sqrt{\epsilon} + (1-\epsilon)/3$. We can then compute the optimal hyperplanes using normal equations:

$$\Theta_1 = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_1}\left[xx^\top\right]^{-1} \mathop{\mathbb{E}}_{x,y \sim \mathcal{D}_1}[xy] = \Sigma^{-1} \mathop{\mathbb{E}}_{x,y \sim \mathcal{D}_1}[xy]$$

Observe, using (59),

$$\mathbb{E}[x_1 y] = \int_{\mathbb{R}} x_1 y \mathcal{D}_1(x_1 y) = \int_{\mathbb{R}} x_1 y \mathcal{D}_1(x_1) \mathcal{D}_1(y) = 0$$

since x_1 and y are independent. Further,

$$\mathbb{E}[x_2 y] = \int_{\mathbb{R}} x_2 y D(x_2, y) = \int_{\mathbb{R}} x_2^2 D(x_2) = \sqrt{\epsilon} + (1-\epsilon)/3$$

Therefore, $\Theta_1 = (0, 1)$. Similarly,

$$\Theta_2 = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_2}\left[xx^\top\right]^{-1} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_2}[xy] = \Sigma^{-1} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_2}[xy]$$

Further, $\mathbb{E}[x_1 y] = 0$. However,

$$\mathbb{E}[x_2 y] = \int_{\mathbb{R}} x_2 y \mathcal{D}_2(x_2, y) = \int_{\mathbb{R}} x_2^2 \mathbb{I}\{|x_2| \leqslant 1\} \mathcal{D}_2(x_2) = 1 - \epsilon$$

Therefore, $\Theta_2 = \left(0, \frac{1-\epsilon}{1+\sqrt{\epsilon}}\right)$. Then,

$$\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \sqrt{\sqrt{\epsilon} + (1-\epsilon)/3} \cdot \frac{\sqrt{\epsilon}+\epsilon}{1+\sqrt{\epsilon}} = \Omega(\sqrt{\epsilon})$$

which concludes the proof. □

7 Bounded Covariance Distributions

In the heavy-tailed setting, the minimal assumption is to consider a distribution over the covariates with bounded covariance. In this setting, we show that robust estimators for linear regression do not exist, even when the underlying linear model has no noise, i.e. the uncorrupted samples are drawn as follows: $y_i = \langle \Theta^*, x_i \rangle$.

Theorem 7.1 (Lower Bound for Bounded Covariance Distributions)**.** *For all $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathbb{R} \times \mathbb{R}$ corresponding to the linear model $y = \langle \Theta_1, x \rangle$ and $y = \langle \Theta_2, x \rangle$ respectively, such that the marginal distribution over $\mathbb{R}$ has variance σ and $d_{TV}(\mathcal{D}_1, \mathcal{D}_2) \leqslant \epsilon$, yet $|\sigma(\Theta_1 - \Theta_2)| = \Omega(\sigma)$.*

Our hard instance relies on the so called *Student's t-distribution*, which has heavy tails when the degrees of freedom are close to 2.

Definition 7.2 (Student's t-distribution)**.** Given $\nu > 1$, Student's t-distribution has the following probability density function:

$$f_\nu(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where $\Gamma(z) = \int_0^\infty x^{z-1} e^x dx$, for $z \in \mathbb{R}$, is the Gamma function.

We use the following facts about Student's t-distribution:

Fact 7.3 (Mean and Variance)**.** *The mean of Student's t-distribution is $\mathbb{E}_{x\sim f_\nu}[x] = 0$ for $\nu > 1$ and undefined otherwise. The variance of Student's t-distribution is*

$$\mathop{\mathbb{E}}_{x\sim f_\nu}[x^2] = \begin{cases} \infty & \text{if } 1 < \nu \leqslant 2 \\ \frac{\nu}{\nu-2} & \text{if } 2 < \nu \\ \text{undefined} & \text{otherwise} \end{cases}$$

The intuition behind our lower bound is to construct a regression instance where the covariates are non-zero only on an ϵ-measure support and are heavy tailed when non-zero. As a consequence, the adversary can introduce a distinct valid regression instance by changing a different ϵ-measure of the support. It is then information-theoretically impossible to distinguish between the true and the planted models.

Proof of Theorem 7.1. We construct a 1-dimensional instance where the marginal distribution over covariates is identical for $\mathcal{D}_1$ and $\mathcal{D}_2$. The pdf is given as follows: for $q \in \{1,2\}$ the marginal distribution on the covariates is given as follows:

$$\mathcal{D}_q(x) = \begin{cases} 1-\epsilon, & \text{if } x = 0 \\ \epsilon \cdot f_{2+\epsilon}(x) & \text{otherwise} \end{cases}$$

The distribution of the labels is gives as follows:

$$\mathcal{D}_1(y \mid x) = x \text{ and } \mathcal{D}_2(y \mid x) = -x$$

Next, we compute the total variation distance between $\mathcal{D}_1$ and $\mathcal{D}_2$. Recall,

$$\begin{aligned} d_{\mathsf{TV}}(\mathcal{D}_1, \mathcal{D}_2) &= \frac{1}{2}\int_{\mathbb{R}\times\mathbb{R}} |\mathcal{D}_1(x,y) - \mathcal{D}_2(x,y)| \\ &= \frac{1}{2}\int_{\mathbb{R}} \mathcal{D}_1(x) \int_{\mathbb{R}} |\mathcal{D}_1(y \mid x) - \mathcal{D}_2(y \mid x)| \\ &= \frac{1}{2}\int_{\mathbb{R}} |\mathcal{D}_1(y \mid x) - \mathcal{D}_2(y \mid x)| \left(\mathbb{I}\{x = 0\} + \mathbb{I}\{x \neq 0\}\right) \\ &= \frac{1}{2}\int_{\mathbb{R}} |2x| \mathbb{I}\{x \neq 0\} \leqslant \epsilon \end{aligned} \tag{60}$$

Observe, since the regression instances have no noise, we can obtain a perfect fit by setting $\Theta_1 = 1$ and $\Theta_2 = -1$. Further, for $q \in \{1,2\}$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_q}[x] = (1-\epsilon)\cdot 0 + \epsilon \cdot \mathop{\mathbb{E}}_{x\sim f_{2+\epsilon}}[x] = 0 \tag{61}$$

and

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_q}\left[x^2\right] = (1-\epsilon)\cdot 0 + \epsilon \cdot \mathop{\mathbb{E}}_{x\sim f_{2+\epsilon}}\left[x^2\right] = \epsilon \cdot \frac{2+\epsilon}{\epsilon} \tag{62}$$

Thus,

$$\left| \mathop{\mathbb{E}}_{x\sim\mathcal{D}_q}\left[x^2\right]^{1/2} (\Theta_1 - \Theta_2) \right| = (2+\epsilon)^{1/2} \cdot 2 = 2\sigma \tag{63}$$

which completes the proof. We note that the 4-th moment of $f_{2+\epsilon}(t)$ is infinite and thus it is not hypercontractive, even for $k = 4$. □

Acknowledgment

We thank Sivaraman Balakrishnan, Sam Hopkins, Pravesh Kothari, Jerry Li, Pradeep Ravikumar and David Wajc for illuminating discussions related to this project. In particular, we thank Pravesh Kothari for answering several technical questions regarding lower bounds appearing in a previous version of [KKM18b].

References

[Bar] Boaz Barak. Proofs, beliefs, and algorithms through the lens of sum-of-squares. 14

[BGL17] Vijay V. S. P. Bhattiprolu, Venkatesan Guruswami, and Euiwoong Lee. Sum-of-squares certificates for maxima of random tensors on the sphere. In APPROX-RANDOM, volume 81 of LIPIcs, pages 31:1–31:20. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017. 6

[BJK15] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In Advances in Neural Information Processing Systems, pages 721–729, 2015. 5

[BJKK17] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In Advances in Neural Information Processing Systems, pages 2110–2119, 2017. 5

[BK20a] Ainesh Bakshi and Pravesh Kothari. List-decodable subspace recovery via sum-of-squares. arXiv preprint arXiv:2002.05139, 2020. 6, 14

[BK20b] Ainesh Bakshi and Pravesh Kothari. Outlier-robust clustering of non-spherical mixtures. arXiv preprint arXiv:2005.02970, 2020. 6, 14, 17

[BKS15] Boaz Barak, Jonathan A. Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In STOC, pages 143–151. ACM, 2015. 6

[CDG19] Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. 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 2755–2771. SIAM, 2019. 1

[CDGW19] Yu Cheng, Ilias Diakonikolas, Rong Ge, and David P. Woodruff. Faster algorithms for high-dimensional robust covariance estimation. In Alina Beygelzimer and Daniel Hsu, editors, Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA, volume 99 of Proceedings of Machine Learning Research, pages 727–757. PMLR, 2019. 1

[CHK+20] Yeshwanth Cherapanamjeri, Samuel B. Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. In Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, page 601–609, New York, NY, USA, 2020. Association for Computing Machinery. 6

[CMY20] Yeshwanth Cherapanamjeri, Sidhanth Mohanty, and Morris Yau. List decodable mean estimation in nearly linear time. arXiv preprint arXiv:2005.09796, 2020. 6

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

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

[DK19] Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. arXiv preprint arXiv:1911.05911, 2019. 2, 5

[DKK+16] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on, pages 655–664. IEEE, 2016. 1, 5

[DKK+17] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In ICML, volume 70 of Proceedings of Machine Learning Research, pages 999–1008. PMLR, 2017. 1

[DKK+18a] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In Artur Czumaj, editor, Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018, pages 2683–2702. SIAM, 2018. 1

[DKK+18b] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. arXiv preprint arXiv:1803.02815, 2018. 4

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 1596–1606. PMLR, 2019. 1, 5

[DKS17] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In FOCS, pages 73–84. IEEE Computer Society, 2017. 1

[DKS18] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. arXiv preprint arXiv:1806.00040, 2018. 4, 5

[DKS19] 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. 1, 5

[FKP+19] Noah Fleming, Pravesh Kothari, Toniann Pitassi, et al. Semialgebraic Proofs and Efficient Algorithm Design. now the essence of knowledge, 2019. 16, 22

[GLS81] M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. Combinatorica, 1(2):169–197, 1981. 15

[HL18] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, pages 1021–1034, 2018. 6

[Hop18] Samuel B Hopkins. Sub-gaussian mean estimation in polynomial time. arXiv preprint arXiv:1809.07425, 2018. 6

[HSS19] Samuel B Hopkins, Tselil Schramm, and Jonathan Shi. A robust spectral algorithm for overcomplete tensor decomposition. In Conference on Learning Theory, pages 1683–1722, 2019. 6

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

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

[Jae72] Louis A Jaeckel. Estimating regression coefficients by minimizing the dispersion of the residuals. The Annals of Mathematical Statistics, pages 1449–1458, 1972. 1

[KKK19] Sushrut Karmalkar, Adam Klivans, and Pravesh Kothari. List-decodable linear regression. In Advances in Neural Information Processing Systems, pages 7423–7432, 2019. 1, 6, 14

[KKM18a] Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. arXiv preprint arXiv:1803.03241, 2018. 14

[KKM18b] 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, 3, 4, 5, 6, 9, 10, 11, 36

[KS17a] Pravesh K. Kothari and Jacob Steinhardt. Better agnostic clustering via relaxed tensor norms. 2017. 2, 6, 11, 12

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

[KS17c] Pravesh K Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. arXiv preprint arXiv:1711.11581, 2017. 2, 5, 6, 11, 12, 13, 14, 28

[Las01] 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. 15

[Li18] Jerry Zheng Li. Principled approaches to robust machine learning and beyond. PhD thesis, Massachusetts Institute of Technology, 2018. 5

[LRV16] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on, pages 665–674. IEEE, 2016. 1, 5

[MSS16a] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In FOCS, pages 438–446. IEEE Computer Society, 2016. 6

[MSS16b] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), pages 438–446. IEEE, 2016. 14

[Nes00] 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. 15

[Par00] Pablo A Parrilo. Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization. PhD thesis, California Institute of Technology, 2000. 15

[PSBR20] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(3):601–627, 2020. 1, 3, 4, 5, 9, 11

[Rou84] Peter J Rousseeuw. Least median of squares regression. Journal of the American statistical association, 79(388):871–880, 1984. 1

[RRS17] Prasad Raghavendra, Satish Rao, and Tselil Schramm. Strongly refuting random csps below the spectral threshold. In STOC, pages 121–131. ACM, 2017. 6

[RSS18] Prasad Raghavendra, Tselil Schramm, and David Steurer. High-dimensional estimation via sum-of-squares proofs. arXiv preprint arXiv:1807.11419, 6, 2018. 5

[RY84] Peter Rousseeuw and Victor Yohai. Robust regression by means of s-estimators. In Robust and nonlinear time series analysis, pages 256–272. Springer, 1984. 1

[RY20a] Prasad Raghavendra and Morris Yau. List decodable learning via sum of squares. In Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 161–180. SIAM, 2020. 1, 6

[RY20b] Prasad Raghavendra and Morris Yau. List decodable subspace recovery, 2020. 6

[SBRJ19] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. Adaptive hard thresholding for near-optimal consistent robust regression. arXiv preprint arXiv:1903.08192, 2019. 5

[SCV17] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. CoRR, abs/1703.04940, 2017. 1

[Sen68] Pranab Kumar Sen. Estimates of the regression coefficient based on kendall's tau. Journal of the American statistical association, 63(324):1379–1389, 1968. 1

[Sho87] N. Z. Shor. Quadratic optimization problems. Izv. Akad. Nauk SSSR Tekhn. Kibernet., (1):128–139, 222, 1987. 15

[SS17] Tselil Schramm and David Steurer. Fast and robust tensor decomposition with applications to dictionary learning. In COLT, volume 65 of Proceedings of Machine Learning Research, pages 1760–1793. PMLR, 2017. 6

[Ste18] Jacob Steinhardt. ROBUST LEARNING: INFORMATION THEORY AND ALGORITHMS. PhD thesis, STANFORD UNIVERSITY, 2018. 5

[The92] Henri Theil. A rank-invariant method of linear and polynomial regression analysis. In Henri Theil's contributions to economics and econometrics, pages 345–381. Springer, 1992. 1

[Wei05] Sanford Weisberg. Applied linear regression, volume 528. John Wiley & Sons, 2005. 1

[ZJS19] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. arXiv preprint arXiv:1909.08755, 2019. 6, 9

[ZJS20] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Robust estimation via generalized quasi-gradients. arXiv preprint arXiv:2005.14073, 2020. 3, 4, 5, 6

A Robust Identifiability for Arbitrary Noise

Proof of Corollary 4.2. Consider a maximal coupling of $\mathcal{D}, \mathcal{D}'$ over $(x,y) \times (x',y')$, denoted by $\mathcal{G}$, such that the marginal of $\mathcal{G}$ (x,y) is $\mathcal{D}$, the marginal on (x',y') is $\mathcal{D}'$ and $\mathbb{P}_{\mathcal{G}}[\mathbb{I}\{(x,y) = (x',y')\}] = 1-\epsilon$. Then, for all v,

$$\begin{aligned}\langle v, \Sigma_{\mathcal{D}}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}) + xy - xy\right\rangle\right] \\ &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(\langle x, \Theta_{\mathcal{D}}\rangle - y\right)\rangle\right] + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle\right]\end{aligned} \tag{64}$$

Since $\Theta_{\mathcal{D}}$ is the minimizer for the least squares loss, we have the following gradient condition : for all $v \in \mathbb{R}^d$,

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\langle v, (\langle x, \Theta_{\mathcal{D}}\rangle - y)x\rangle\right] = 0 \tag{65}$$

Since $\mathcal{G}$ is a coupling, using the gradient condition (65) and using that $1 = \mathbb{I}\{(x,y) = (x',y')\}$ $+\mathbb{I}\{(x,y) \neq (x',y')\}$, we can rewrite equation (64) as

$$\begin{aligned}\langle v, \Sigma_{\mathcal{D}}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle \mathbb{I}\{(x,y) = (x',y')\}\right] \\ &\quad + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle \mathbb{I}\{(x,y) \neq (x',y')\}\right] \\ &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle \mathbb{I}\{(x,y) = (x',y')\}\right] \\ &\quad + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle \mathbb{I}\{(x,y) \neq (x',y')\}\right]\end{aligned} \tag{66}$$

Consider the first term in the last equality above. Using the gradient condition for $\Theta_{\mathcal{D}'}$ along with Hölder's Inequality, we have

$$\begin{aligned}&\left|\mathbb{E}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle \mathbb{I}\{(x,y) = (x',y')\}\right]\right| \\ &\quad = \left|\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle\right] - \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle \mathbb{I}\{(x,y) \neq (x',y')\}\right]\right| \\ &\quad = \left|\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle \mathbb{I}\{(x,y) \neq (x',y')\}\right]\right| \\ &\quad \leqslant \left|\mathop{\mathbb{E}}_{\mathcal{G}}\left[\mathbb{I}\{(x,y) \neq (x',y')\}^{k/(k-2)}\right]^{(k-2)/k}\right| \cdot \left|\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left\langle v, x'\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)\right\rangle^{k/2}\right]^{2/k}\right|\end{aligned} \tag{67}$$

Observe, since $\mathcal{G}$ is a maximal coupling $\mathbb{E}_{\mathcal{G}}\left[\mathbb{I}\{(x,y) \neq (x',y')\}\right]^{(k-2)/k} \leqslant \epsilon^{1-2/k}$. Here, we no longer have independence of the noise and the covariates, therefore using Cauchy-Schwarz

$$\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left\langle v, x'\right\rangle^{k/2} \cdot \left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)^{k/2}\right] \leqslant \left(\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left\langle v, x'\right\rangle^{k}\right] \mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)^{k}\right]\right)^{1/2}$$

By hypercontractivity of the covariates and the noise, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left\langle v, x'\right\rangle^{k}\right]^{1/k} \mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)^{k}\right]^{1/k} \leqslant \mathcal{O}(c_k\, \eta_k)\left(v^\top \Sigma_{\mathcal{D}'} v\right)^{1/2} \mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y' - \left\langle x', \Theta_{\mathcal{D}'}\right\rangle\right)^2\right]^{1/2}$$

Therefore, we can restate (67) as follows

$$\left| \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x' \left(y' - \langle x', \Theta_{\mathcal{D}'} \rangle \right) \rangle \mathbb{I} \left\{ (x,y) = (x',y') \right\} \right] \right| \leqslant \mathcal{O}\Big(c_k \, \eta_k \, \epsilon^{\frac{k-2}{k}} \Big) \left(v^\top \Sigma_{\mathcal{D}'} v \right)^{\frac{1}{2}} \mathop{\mathbb{E}}_{x',y' \sim \mathcal{D}'} \left[\left(y' - \langle x', \Theta_{\mathcal{D}'} \rangle \right)^2 \right]^{\frac{1}{2}} \tag{68}$$

It remains to bound the second term in the last equality of equation (66), and we proceed as follows :

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x \left(y - \langle x, \Theta_{\mathcal{D}'} \rangle \right) \rangle \mathbb{I} \left\{ (x,y) \neq (x',y') \right\} \right] &= \mathop{\mathbb{E}}_{\mathcal{G}} \left[\left\langle v, xx^\top \left(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'} \right) \right\rangle \mathbb{I} \left\{ (x,y) \neq (x',y') \right\} \right] \\ &\quad + \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x \left(y - \langle x, \Theta_{\mathcal{D}} \rangle \right) \rangle \mathbb{I} \left\{ (x,y) \neq (x',y') \right\} \right] \end{aligned} \tag{69}$$

We bound the two terms above separately. Observe, applying Hölder's Inequality to the first term, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\left\langle v, xx^\top \left(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'} \right) \right\rangle \mathbb{I} \left\{ (x,y) \neq (x',y') \right\} \right] &\leqslant \mathop{\mathbb{E}}_{\mathcal{G}} \left[\mathbb{I} \left\{ (x,y) \neq (x',y') \right\} \right]^{\frac{k-2}{k}} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\left\langle v, xx^\top \left(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'} \right) \right\rangle^{\frac{k}{2}} \right]^{\frac{2}{k}} \\ &\leqslant \epsilon^{\frac{k-2}{k}} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\left\langle v, xx^\top \left(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'} \right) \right\rangle^{\frac{k}{2}} \right]^{\frac{2}{k}} \end{aligned} \tag{70}$$

To bound the second term in equation 69, we again use Hölder's Inequality followed by Cauchy-Schwarz noise and covariates.

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x \left(y - \langle x, \Theta_{\mathcal{D}} \rangle \right) \rangle \mathbb{I} \left\{ (x,y) \neq (x',y') \right\} \right] &\leqslant \mathop{\mathbb{E}}_{\mathcal{G}} \left[\mathbb{I} \left\{ (x,y) \neq (x',y') \right\} \right]^{\frac{k-1}{k}} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x \left(y - \langle x, \Theta_{\mathcal{D}} \rangle \right) \rangle^k \right]^{\frac{1}{k}} \\ &\leqslant \epsilon^{\frac{k-2}{k}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\langle v, x \rangle^{k/2} \right]^{2/k} \mathop{\mathbb{E}}_{x,y \sim \mathcal{D}} \left[\left(y - \langle x, \Theta_{\mathcal{D}} \rangle \right)^{k/2} \right]^{2/k} \\ &\leqslant \epsilon^{\frac{k-2}{k}} \, c_k \, \eta_k \left(v^\top \Sigma_{\mathcal{D}} v \right)^{1/2} \mathop{\mathbb{E}}_{x,y \sim \mathcal{D}} \left[\left(y - \langle x, \Theta_{\mathcal{D}} \rangle \right)^2 \right]^{1/2} \end{aligned} \tag{71}$$

where the last inequality follows from hypercontractivity of the covariates and noise. Substituting the upper bounds obtained in Equations (70) and (71) back in to (69),

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x \left(y - \langle x, \Theta_{\mathcal{D}'} \rangle \right) \rangle \mathbb{I} \left\{ (x,y) \neq (x',y') \right\} \right] &\leqslant \epsilon^{\frac{k-2}{k}} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\left\langle v, xx^\top \left(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'} \right) \right\rangle^{\frac{k}{2}} \right]^{\frac{2}{k}} \\ &\quad + \epsilon^{\frac{k-2}{k}} \, c_k \, \eta_k \left(v^\top \Sigma_{\mathcal{D}} v \right)^{1/2} \mathop{\mathbb{E}}_{x,y \sim \mathcal{D}} \left[\left(y - \langle x, \Theta_{\mathcal{D}} \rangle \right)^2 \right]^{1/2} \end{aligned}$$

Therefore, we can now upper bound both terms in Equation (66) as follows:

$$\begin{aligned}\langle v, \Sigma_{\mathcal{D}}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle &\leqslant \mathcal{O}\Big(c_k\,\eta_k\,\epsilon^{\frac{k-2}{k}}\Big)\left(v^\top \Sigma_{\mathcal{D}'} v\right)^{1/2} \mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)^2\right]^{1/2} \\ &+ \mathcal{O}\Big(\epsilon^{\frac{k-2}{k}}\Big) \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}\right)\right\rangle^{k/2}\right]^{2/k} \\ &+ \mathcal{O}\Big(\epsilon^{\frac{k-2}{k}}\,c_k\,\eta_k\Big)\left(v^\top \Sigma_{\mathcal{D}} v\right)^{1/2} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[(y - \langle x, \Theta_{\mathcal{D}}\rangle)^2\right]^{1/2}\end{aligned} \tag{72}$$

Recall, since the marginals of $\mathcal{D}$ and $\mathcal{D}'$ on $\mathbb{R}^d$ are (c_k, k)-hypercontractive and $\|\mathcal{D} - \mathcal{D}'\|_{\mathsf{TV}} \leqslant \epsilon$, it follows from Fact 3.3 that

$$(1 - 0.1)\,\Sigma_{\mathcal{D}'} \preceq \Sigma_{\mathcal{D}} \preceq (1 + 0.1)\,\Sigma_{\mathcal{D}'} \tag{73}$$

when $\epsilon \leqslant \mathcal{O}\Big((1/c_k k)^{k/(k-2)}\Big)$. Now, consider the substitution $v = \Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}$. Observe,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}\right)\right\rangle^{k/2}\right]^{2/k} &= \mathop{\mathbb{E}}_{\mathcal{D}}\left[\langle x, (\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle^k\right]^{2/k} \\ &\leqslant c_k^2\left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2^2\end{aligned} \tag{74}$$

Then, using the bounds in (73) and (74) along with $v = \Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}$ in Equation 72, we have

$$\begin{aligned}\left(1 - \mathcal{O}\Big(\epsilon^{\frac{k-2}{k}} c_k^2\Big)\right)\left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2^2 &\leqslant \mathcal{O}\Big(c_k\,\eta_k\,\epsilon^{\frac{k-2}{k}}\Big)\left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2 \\ &\left(\mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)^2\right]^{\frac{1}{2}} + \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[(y - \langle x, \Theta_{\mathcal{D}}\rangle)^2\right]^{\frac{1}{2}}\right)\end{aligned} \tag{75}$$

Dividing out (75) by $\left(1 - \mathcal{O}\Big(\epsilon^{\frac{k-2}{k}} c_k^2\Big)\right)\left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2^2$ and observing that $\mathcal{O}\Big(\epsilon^{\frac{k-2}{k}} c_k^2\Big)$ is upper bounded by a fixed constant less than 1 yields the parameter recovery bound.

Given the parameter recovery result above, we bound the least-squares loss between the two hyperplanes on $\mathcal{D}$ as follows:

$$\begin{aligned}\left|\mathrm{err}_{\mathcal{D}}(\Theta_{\mathcal{D}}) - \mathrm{err}_{\mathcal{D}}(\Theta_{\mathcal{D}'})\right| &= \Big|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left(y - x^\top\Theta_{\mathcal{D}}\right)^2 - \left(y - x^\top\Theta_{\mathcal{D}'} + x^\top\Theta_{\mathcal{D}} - x^\top\Theta_{\mathcal{D}}\right)^2\right]\Big| \\ &= \Big|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\langle x, (\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle^2 + 2(y - x^\top\Theta_{\mathcal{D}})x^\top(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right]\Big| \\ &\leqslant \mathcal{O}\Big(c_k^2\,\eta_k^2\,\epsilon^{2-4/k}\Big)\left(\mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)^2\right] + \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[(y - \langle x, \Theta_{\mathcal{D}}\rangle)^2\right]\right)\end{aligned} \tag{76}$$

where the last inequality follows from observing $\mathbb{E}\left[\left\langle\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}, x(y - x^\top\Theta_{\mathcal{D}})\right\rangle\right] = 0$ (gradient condition) and squaring the parameter recovery bound. □

B Efficient Estimator for Arbitrary Noise

In this section, we provide a proof of the key SoS lemma required to obtain a polynomial time estimator. The remainder of the proof, including the feasibility of the constraints and rounding is identical to the one presented in Section 5.

Lemma B.1 (Robust Identifiability in SoS for Arbitrary Noise)**.** *Consider the hypothesis of Theorem 5.1. Let w, x', y' and Θ be feasible solutions for the polynomial constraint system $\mathcal{A}$. Let $\hat{\Theta} = \arg\min_\Theta \frac{1}{n}\sum_{i\in[n]}(y_i^* - \langle x_i^*, \Theta\rangle)^2$ be the empirical loss minimizer on the uncorrupted samples and let $\hat{\Sigma} = \mathbb{E}\left[x_i^*(x_i^*)^\top\right]$ be the covariance of the uncorrupted samples. Then,*

$$\begin{aligned}\mathcal{A}\Big|^{w,x',y',\Theta}_{4k}\Bigg\{\left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{2k} \leqslant{}& 2^{3k}(2\epsilon)^{k-2}c_k^k\,\eta_k^k\,\sigma^{k/2}\left\|\mathbb{E}\left[x_i'(x_i')^\top\right]^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^k \\ &+ 2^{3k}(2\epsilon)^{k-2}c_k^{2k}\left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{2k} \\ &+ 2^{3k}(2\epsilon)^{k-2}c_k^k\,\eta_k^k\,\mathbb{E}\left[\left(y_i^*-\langle x_i^*,\hat{\Theta}\rangle\right)^2\right]^{k/2}\left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^k\Bigg\}\end{aligned}$$

Proof. Consider the empirical covariance of the uncorrupted set given by $\hat{\Sigma} = \mathbb{E}\left[x_i^*(x_i^*)^\top\right]$. Then, using the Substitution Rule, along with Fact 3.18

$$\begin{aligned}\Big|^{\Theta}_{2k}\left\{\left\langle v, \hat{\Sigma}\left(\hat{\Theta}-\Theta\right)\right\rangle^k = \left\langle v, \mathbb{E}\left[x_i^*(x_i^*)^\top\left(\hat{\Theta}-\Theta\right) + x_i^*y_i^* - x_i^*y_i^*\right]\right\rangle^k \right. \\ = \left\langle v, \mathbb{E}\left[x_i^*\left(\langle x_i^*,\hat{\Theta}\rangle - y_i^*\right)\right] + \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*,\Theta\rangle\right)\right]\right\rangle^k \\ \left. \leqslant 2^k\left\langle v, \mathbb{E}\left[x_i^*\left(\langle x_i^*,\hat{\Theta}\rangle - y_i^*\right)\right]\right\rangle^k + 2^k\left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*,\Theta\rangle\right)\right]\right\rangle^k\right\}\end{aligned} \tag{77}$$

Since $\hat{\Theta}$ is the minimizer of $\mathbb{E}\left[\left(\langle x_i^*,\Theta\rangle - y_i^*\right)^2\right]$, the gradient condition (appearing in Equation (65) of the indentifiability proof) implies this term is 0. Therefore, it suffices to bound the second term.

For all $i \in [n]$, let $w_i' = w_i$ iff the i-th sample is uncorrupted in $\mathcal{X}_\epsilon$, i.e. $x_i = x_i^*$. Then, it is easy to see that $\sum_i w_i' \geqslant (1-2\epsilon)n$. Further, since $\mathcal{A}\Big|^w_2\left\{(1-w_i'w_i)^2 = (1-w_i'w_i)\right\}$,

$$\mathcal{A}\Big|^w_2\left\{\frac{1}{n}\sum_{i\in[n]}(1-w_i'w_i)^2 = \frac{1}{n}\sum_{i\in[n]}(1-w_i'w_i) \leqslant 2\epsilon\right\} \tag{78}$$

The above equation bounds the uncorrupted points in $\mathcal{X}_\epsilon$ that are not indicated by w. Then, using the Substitution Rule, along with the SoS Almost Triangle Inequality (Fact 3.18),

$$\mathcal{A} \left|\frac{\Theta, w'}{2k}\right. \left\{ \left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k = \left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\left(w_i' + 1 - w_i'\right)\right)\right]\right\rangle^k \right.$$
$$= \left\langle v, \mathbb{E}\left[w_i' x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right] + \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k$$
$$\leqslant 2^k \left\langle v, \mathbb{E}\left[w_i' x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k$$
$$\left. + 2^k \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \right\} \tag{79}$$

Consider the first term of the last inequality in (79). Observe, since $w_i' x_i^* = w_i w_i' x_i'$ and similarly, $w_i' y_i^* = w_i w_i' y_i'$,

$$\mathcal{A} \left|\frac{\Theta, w'}{4}\right. \left\{ \mathbb{E}\left[w_i' x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right] = \mathbb{E}\left[w_i' w_i x_i'\left(y_i' - \langle x_i', \Theta\rangle\right)\right] \right\}$$

For the sake of brevity, the subsequent statements hold for relevant SoS variables and have degree $O(k)$ proofs. Using the Substitution Rule,

$$\mathcal{A} \left|\vdash\right. \left\{ \left\langle v, \mathbb{E}\left[w_i' x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k = \left\langle v, \mathbb{E}\left[w_i' w_i x_i'\left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k \right.$$
$$= \left\langle v, \mathbb{E}\left[x_i'\left(y_i' - \langle x_i', \Theta\rangle\right)\right] + \mathbb{E}\left[(1 - w_i' w_i) x_i'\left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k$$
$$\leqslant 2^k \left\langle v, \mathbb{E}\left[x_i'\left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k$$
$$\left. + 2^k \left\langle v, \mathbb{E}\left[(1 - w_i' w_i) x_i'\left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k \right\} \tag{80}$$

Observe, the first term in the last inequality above is identically 0, since we enforce the gradient condition on the SoS variables x', y' and Θ. We can then rewrite the second term using linearity of expectation, followed by applying SoS Hölder's Inequality (Fact 3.19) combined with $\mathcal{A} \left|\frac{w}{2}\right. \left\{(1 - w_i' w_i)^2 = 1 - w_i' w_i\right\}$ to get

$$\mathcal{A} \left|\vdash\right. \left\{ \left\langle v, \mathbb{E}\left[(1 - w_i' w_i) x_i'\left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k = \mathbb{E}\left[\left\langle v, (1 - w_i') w_i x_i'\left(y_i' - \langle x_i', \Theta\rangle\right)\right\rangle\right]^k \right.$$
$$= \mathbb{E}\left[(1 - w_i' w_i)\left\langle v, x_i'\right\rangle\left(y_i' - \langle x_i', \Theta\rangle\right)\right]^k$$
$$\leqslant \mathbb{E}\left[(1 - w_i' w_i)\right]^{k-2} \mathbb{E}\left[\left\langle v, x_i'\right\rangle^{k/2}\left(y_i' - \langle x_i', \Theta\rangle\right)^{k/2}\right]$$
$$\left. \leqslant (2\epsilon)^{k-2} \mathbb{E}\left[\left\langle v, x_i'\right\rangle^k\right] \mathbb{E}\left[\left(y_i' - \langle x_i', \Theta\rangle\right)^k\right] \right\} \tag{81}$$

where the last inequality follows from (78) and the SoS Cauchy Schwarz Inequality. Using the certifiable-hypercontractivity of the covariates,

$$\mathcal{A} \left|\frac{w,x'}{2k}\right. \left\{ \mathbb{E}\left[\langle v, x_i' \rangle^k\right] \leqslant c_k^k \, \mathbb{E}\left[\langle v, x_i' \rangle^2\right]^{k/2} = c_k^k \left\langle v, \mathbb{E}\left[x_i'(x_i')^\top\right] v \right\rangle^{k/2} \right\} \tag{82}$$

Further, using certifiable hypercontractivity of the noise,

$$\mathcal{A} \vdash \left\{ \mathbb{E}\left[\left(y_i' - \langle w_i x_i', \Theta \rangle\right)^k\right] \leqslant \eta_k^k \, \mathbb{E}\left[(y_i' - \langle x_i', \Theta \rangle)^2)\right]^{k/2} \right\} \tag{83}$$

Recall, $\sigma = \mathbb{E}\left[(y_i' - \langle x_i', \Theta \rangle)^2)\right]$ Combining the upper bounds obtained in (82) and (83), and plugging this back into (81), we get

$$\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1 - w_i')x_i' \left(y_i' - \langle x_i', \Theta \rangle\right)\right] \right\rangle^k \leqslant (2\epsilon)^{k-2} c_k^k \, \eta_k^k \, \sigma^{k/2} \left\langle v, \mathbb{E}\left[x_i'(x_i')^\top\right] v \right\rangle^{k/2} \right\} \tag{84}$$

Recall, we have now bounded the first term of the last inequality in (79). Therefore, it remains to bound the second term of the last inequality in (79). Using the Substitution Rule, we have

$$\begin{aligned} \mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1 - w_i')x_i^* \left(y_i^* - \langle x_i^*, \Theta \rangle\right)\right] \right\rangle^k &= \left\langle v, \mathbb{E}\left[(1 - w_i')x_i^* \left(y_i^* - \langle x_i^*, \Theta - \hat{\Theta} + \hat{\Theta} \rangle\right)\right] \right\rangle^k \\ &\leqslant 2^k \left\langle v, \mathbb{E}\left[(1 - w_i')x_i^* \left(y_i^* - \langle x_i^*, \hat{\Theta} \rangle\right)\right] \right\rangle^k \\ &\quad + 2^k \left\langle v, \mathbb{E}\left[(1 - w_i')x_i^* \left(\langle x_i^*, \Theta - \hat{\Theta} \rangle\right)\right] \right\rangle^k \Bigg\} \end{aligned} \tag{85}$$

We again handle each term separately. Observe, the first term when decoupled is a statement about the uncorrupted samples. Therefore, using the SoS Hölder's Inequality (Fact 3.19),

$$\begin{aligned} \mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1 - w_i')x_i^* \left(y_i^* - \langle x_i^*, \hat{\Theta} \rangle\right)\right] \right\rangle^k &= \mathbb{E}\left[(1 - w_i') \left\langle v, x_i^* \left(y_i^* - \langle x_i^*, \hat{\Theta} \rangle\right) \right\rangle\right]^k \\ &\leqslant \mathbb{E}\left[(1 - w_i')\right]^{k-2} \mathbb{E}\left[\left\langle v, x_i^* \left(y_i^* - \langle x_i^*, \hat{\Theta} \rangle\right) \right\rangle^{k/2}\right] \\ &\leqslant (2\epsilon)^{k-2} \, \mathbb{E}\left[\langle v, x_i^* \rangle^k\right] \mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta} \rangle\right)^k\right] \Bigg\} \end{aligned} \tag{86}$$

Using certifiable hypercontractivity of the x_i^*s,

$$\mathbb{E}\left[\langle v, x_i^* \rangle^k\right] \leqslant c_k^k \, \mathbb{E}\left[\langle v, x_i^* \rangle^2\right]^{k/2} = c_k^k \left\langle v, \hat{\Sigma} v \right\rangle^{k/2}$$

where $\hat{\Sigma} = \mathbb{E}\left[x_i^*(x_i^*)^\top\right]$ and similarly using hypercontractivity of the noise,

$$\mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta} \rangle\right)^k\right] \leqslant \eta_k^k \, \mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta} \rangle\right)^2\right]^{k/2}$$

Then, by the Substitution Rule, we can bound (86) as follows:

$$\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)\right]\right\rangle^k \leqslant (2\epsilon)^{k-1} c_k^k\, \eta_k^k\, \mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^2\right]^{k/2} \langle v, \hat{\Sigma} v\rangle^{k/2} \right\} \tag{87}$$

In order to bound the second term in (85), we use the SoS Hölder's Inequality,

$$\begin{aligned}
\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(\langle x_i^*, \Theta - \hat{\Theta}\rangle\right)\right]\right\rangle^k &= \mathbb{E}\left[(1-w_i')^{k-2}\left\langle v, x_i^*\left(\langle x_i^*, \Theta - \hat{\Theta}\rangle\right)\right\rangle\right] \\
&\leqslant \mathbb{E}\left[1-w_i'\right]^{k-2} \mathbb{E}\left[\left(v^\top x_i^* (x_i^*)^\top (\Theta - \hat{\Theta})\right)^{\frac{k}{2}}\right]^2 \\
&\leqslant (2\epsilon)^{k-2}\, \mathbb{E}\left[\left(v^\top x_i^* (x_i^*)^\top (\Theta - \hat{\Theta})\right)^{\frac{k}{2}}\right]^2 \Bigg\}
\end{aligned} \tag{88}$$

Combining the bounds obtained in (87) and (88), we can restate Equation (85) as follows

$$\begin{aligned}
\mathcal{A} \vdash \Bigg\{ \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k &\leqslant 2^k (2\epsilon)^{k-1} c_k^k\, \eta_k^k\, \mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^2\right]^{k/2} \langle v, \hat{\Sigma} v\rangle^{k/2} \\
&\quad + 2^k (2\epsilon)^{k-2}\, \mathbb{E}\left[\left(v^\top x_i^* (x_i^*)^\top (\Theta - \hat{\Theta})\right)^{\frac{k}{2}}\right]^2 \Bigg\}
\end{aligned} \tag{89}$$

Combining (89) with (84), we obtain an upper bound for the last inequality in Equation (79). Therefore, using the Substitution Rule, we obtain

$$\begin{aligned}
\mathcal{A} \vdash \Bigg\{ \left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k &\leqslant 2^k (2\epsilon)^{k-1} c_k^k\, \eta_k^k\, \sigma^{k/2} \left\langle v, \mathbb{E}\left[x_i'(x_i')^\top\right] v\right\rangle^{k/2} \\
&\quad + 2^{2k} (2\epsilon)^{k-2}\, \mathbb{E}\left[\left(v^\top x_i^* (x_i^*)^\top (\Theta - \hat{\Theta})\right)^{\frac{k}{2}}\right]^2 \\
&\quad + 2^{2k} (2\epsilon)^{k-1} c_k^k\, \eta_k^k\, \mathbb{E}\left[\left(y_i^* - \langle x_i^*, \hat{\Theta}\rangle\right)^2\right]^{k/2} \langle v, \hat{\Sigma} v\rangle^{k/2} \Bigg\}
\end{aligned} \tag{90}$$

The remaining proof is identical to Lemma 5.5. □

C Proof of Lemma 3.4

Lemma C.1 (Löwner Ordering for Hypercontractive Samples (restated)). *Let $\mathcal{D}$ be a (c_k, k)-hypercontractive distribution with covariance Σ and and let $\mathcal{U}$ be the uniform distribution over n samples. Then, with probability $1-\delta$,*

$$\left\| \Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - I \right\|_F \leqslant \frac{C_4 d^2}{\sqrt{n}\sqrt{\delta}},$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{i\in[n]} x_i x_i^\top$.

Proof. Let $\tilde{x}_i = \Sigma^{-1/2} x_i$ and observe that $\frac{1}{n}\sum_i \tilde{x}_i \tilde{x}_i{}^T = \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}$. Moreover, we know that $\mathbb{E}[\tilde{x}\tilde{x}^T] = I$. Let $z_{j,k}$ be the (j,k) entry of $\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I$ given by,

$$z_{j,k} = \frac{1}{n}\sum_{i\in[n]} \tilde{x}_i(j)\tilde{x}_i(k) - \mathbb{E}[\tilde{x}(i)\tilde{x}(k)]$$

Using Chebyshev's inequality, we get that with probability at least $1-\delta$,

$$|z_{jk}| \leqslant \frac{\mathbb{E}[\tilde{x}(j)^2\tilde{x}(k)^2]}{\sqrt{n}\sqrt{\delta}} \overset{(i)}{\leqslant} \frac{\mathbb{E}[\tilde{x}(j)^4] + \mathbb{E}[\tilde{x}(k)^4]}{2\sqrt{n}\sqrt{\delta}},$$

where (i) follows from AM-GM inequality. To bound $\mathbb{E}[\tilde{x}(j)^4]$, we use hypercontractivity.

$$\mathbb{E}[\tilde{x}(j)^4] = \mathbb{E}[(v^T x)^4] \leqslant C_4 \mathbb{E}[(v^T x)^2]^2,$$

where $v = \Sigma^{-1/2} e_j$. Plugging this above, we get that $\mathbb{E}[\tilde{x}(j)^4] \leqslant C_4$. which in turn implies that with probability at least $1-\delta$,

$$|z_{jk}| \leqslant \frac{C_4}{\sqrt{n\delta}}.$$

Taking a union bound over d^2 entries of $\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I$, we get that with probability at least $1-\delta$,

$$\left\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I\right\|_F \leqslant \frac{C_4 d^2}{\sqrt{n}\sqrt{\delta}}$$

□