Consistent regression when oblivious outliers overwhelm

Tommaso d'Orsi [*1] **Gleb Novikov** [*1] **David Steurer** [*1]

Abstract

We consider a robust linear regression model $y = X\beta^* + \eta$, where an adversary oblivious to the design $X \in \mathbb{R}^{n\times d}$ may choose η to corrupt all but an α fraction of the observations y in an arbitrary way. Prior to our work, even for Gaussian X, no estimator for β^* was known to be consistent in this model except for quadratic sample size $n \gtrsim (d/\alpha)^2$ or for logarithmic inlier fraction $\alpha \geq 1/\log n$. We show that consistent estimation is possible with nearly linear sample size and inverse-polynomial inlier fraction. Concretely, we show that the Huber loss estimator is consistent for every sample size $n = \omega(d/\alpha^2)$ and achieves an error rate of $O(d/\alpha^2 n)^{1/2}$ (both bounds are optimal up to constant factors). Our results extend to designs far beyond the Gaussian case and only require the column span of X to not contain approximately sparse vectors (similar to the kind of assumption commonly made about the kernel space for compressed sensing). We provide two technically similar proofs. One proof is phrased in terms of strong convexity, extending work of (Tsakonas et al., 2014), and particularly short. The other proof highlights a connection between the Huber loss estimator and high-dimensional median computations. In the special case of Gaussian designs, this connection leads us to a strikingly simple algorithm based on computing coordinate-wise medians that achieves nearly optimal guarantees in linear time, and that can exploit sparsity of β^*. The model studied here also captures heavy-tailed noise distributions that may not even have a first moment.

[*]Equal contribution [1]Department of Computer Science, ETH Zürich, Switzerland. Correspondence to: Tommaso d'Orsi <tommaso.dorsi@inf.ethz.ch>, Gleb Novikov <gleb.novikov@inf.ethz.ch>, David Steurer <david.steurer@inf.ethz.ch>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

1. Introduction

Linear regression is a fundamental task in statistics: given observations $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{d+1}$ following a linear model $y_i = \langle x_i, \beta^* \rangle + \eta_i$, where $\beta^* \in \mathbb{R}^d$ is the unknown parameter of interest and $\eta_1, \ldots, \eta_n$ is noise, the goal is to recover β^* as accurately as possible.

In the most basic setting, the noise values are drawn independently from a Gaussian distribution with mean 0 and variance σ^2. Here, the classical least-squares estimator $\hat{\boldsymbol{\beta}}$ achieves an optimal error bound $\frac{1}{n}\|X(\beta^* - \hat{\boldsymbol{\beta}})\|^2 \lesssim \sigma^2 \cdot d/n$ with high probability, where the design X is has rows $x_1, \ldots, x_n$. Unfortuantely, this guarantee is fragile and the estimator may experience arbitrarily large error in the presence of a small number of benign outlier noise values.

In many modern applications, including economics (Rousseeuw & Leroy, 2005), image recognition (Wright et al., 2008), and sensor networks (Haupt et al., 2008), there is a desire to cope with such outliers stemming from extreme events, gross errors, skewed and corrupted measurements. It is therefore paramount to design estimators robust to noise distributions that may have substantial probability mass on outlier values.

In this paper, we aim to identify the weakest possible assumptions on the noise distribution such that for a wide range of measurement matrices X, we can efficiently recover the parameter vector β^* with vanishing error.

The design of learning algorithms capable of succeeding on data sets contaminated by adversarial noise has been a central topic in robust statistics (e.g. see (Diakonikolas et al., 2019a; Charikar et al., 2017) and their follow-ups for some recent developments). In the context of regression with adaptive adversarial outliers (i.e. depending on the instance) several results are known (Candes & Tao, 2005; Candes et al., 2005; Klivans et al., 2018; Karmalkar et al., 2019; Liu et al., 2019; 2018; Karmalkar & Price, 2018; Dalalyan & Thompson, 2019; Raghavendra & Yau, 2020). However, it turns out that for adaptive adversaries, vanishing error bounds are only possible if the fraction of outliers is vanishing.

In order to make vanishing error possible in the presence of large fractions of outliers, we consider weaker adversary models that are oblivious to the design X. Different assump-

tions can be used to model oblivious adversarial corruptions. (Sun et al., 2019) assume the noise distribution satisfies $\mathbb{E}[\boldsymbol{\eta}_i \mid x_i] = 0$ and $\mathbb{E}\left[|\boldsymbol{\eta}_i|^{1+\delta}\right] < \infty$ for some $0 \le \delta \le 1$, and show that if X has constant condition number, then (a modification of) the Huber loss estimator (Huber, 1964) is consistent for[1] $n \gtrsim \tilde{O}((\|X\|_\infty \cdot d)^{(1+\delta)/2\delta})$ (an estimator is consistent if the error tends to zero as the number of observation grows, $\frac{1}{n}\|X(\hat{\boldsymbol{\beta}} - \beta^*)\|^2 \to 0$).

Without constraint on moments, a useful model is that of assuming the noise vector $\eta \in \mathbb{R}^n$ to be an arbitrary fixed vector with $\alpha \cdot n$ coordinates bounded by 1 in absolute value. This modelalso captures random vectors $\boldsymbol{\eta} = \zeta + \boldsymbol{w}$, where $\zeta \in \mathbb{R}^n$ is αn-sparse and $\boldsymbol{w}$ is a random vector with i.i.d. entries with bounded variance independent of the measurement matrix X, and conveniently allows us to think of the α fraction of samples with small noise as the set of uncorrupted samples. In these settings, the problem has been mostly studied in the context of Gaussian design $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim N(0, \Sigma)$. (Bhatia et al., 2017b) provided an estimator achieving error $\tilde{O}(d/(\alpha^2 \cdot n))$ for any α larger than some fixed constant. This result was then extended in (Suggala et al., 2019), where the authors proposed a near-linear time algorithm computing a $\tilde{O}(d/(\alpha^2 \cdot n))$-close estimate for any[2] $\alpha \gtrsim 1/\log\log n$. That is, allowing the number of uncorrupted samples to be $o(n)$. Considering even smaller fractions of inliers, (Tsakonas et al., 2014) showed that with high probability the Huber loss estimator is consistent for $n \gtrsim \tilde{O}(d^2/\alpha^2)$, thus requiring sample size quadratic in the ambient dimension.

Prior to this work, little was known for more general settings when the design matrix X is non-Gaussian. From an asymptotic viewpoint, i.e., when d and α are fixed and $n \to \infty$, a similar model was studied 30 years ago in a seminal work by Pollard (Pollard, 1991), albeit under stronger assumptions on the noise vector. Under mild constraints on X, it was shown that the least absolute deviation (LAD) estimator is consistent.

So, the outlined state-of-the-art provides an incomplete picture of the statistical and computational complexity of the problem. The question of what conditions we need to enforce on the measurement matrix X and the noise vector $\boldsymbol{\eta}$ in order to efficiently and consistently recover β^* remains largely unanswered. In high-dimensional settings, no estimator has been shown to be consistent when the fraction of uncontaminated samples α is smaller than $1/\log n$ and the number of samples n is smaller than d^2/α^2, even in the simple settings of spherical Gaussian design. Furthermore, even less is known on how we can regress consistently when the design matrix is non-Gaussian.

In this work, we provide a more comprehensive picture of the problem. Concretely, we analyze the Huber loss estimator in non-asymptotic, high dimensional setting where the fraction of inliers may depend (even polynomially) on the number of samples and ambient dimension. Under *mild* assumptions on the design matrix and the noise vector, we show that such algorithm achieves *optimal* error guarantees and sample complexity.

Furthermore, a by-product of our analysis is an strikingly simple linear-time estimator based on computing coordinate-wise medians, that achieves nearly optimal guarantees for standard Gaussian design, even in the regime where the parameter vector β^* is k-sparse (i.e. β^* has at most k nonzero entries).

1.1. Results about Huber-loss estimator

We provide here guarantees on the error convergence of the Huber-loss estimator, defined as a minimizer of the *Huber loss* $f\colon \mathbb{R}^d \to \mathbb{R}_{\ge 0}$,

$$f(\beta) = \frac{1}{n}\sum_{i=1}^{n} \Phi[(X\beta - y)_i],$$

where $\Phi\colon \mathbb{R} \to \mathbb{R}_{\ge 0}$ is the *Huber penalty*,[3]

$$\Phi[t] \stackrel{\text{def}}{=} \begin{cases} \frac{1}{2}t^2 & \text{if } |t| \le 2, \\ 2|t| - 2 & \text{otherwise.} \end{cases}$$

Gaussian design The following theorem states our the Huber-loss estimator in the case of Gaussian designs. Previous quantitative guarantees for consistent robust linear regression focus on this setting (Tsakonas et al., 2014; Bhatia et al., 2017b; Suggala et al., 2019).

Theorem 1.1 (Guarantees for Huber-loss estimator with Gaussian design)**.** *Let $\eta \in \mathbb{R}^n$ be a deterministic vector. Let $\boldsymbol{X}$ be a random[4] n-by-d matrix with iid standard Gaussian entries $\boldsymbol{X}_{ij} \sim N(0,1)$.*

Suppose $n \ge C \cdot d/\alpha^2$, where α is the fraction of entries in η of magnitude at most 1, and $C > 0$ is large enough absolute constant.

Then, with probability at least $1 - 2^{-d}$ over $\boldsymbol{X}$, for every $\beta^* \in \mathbb{R}^d$, given $\boldsymbol{X}$ and $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$, the Huber-loss estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2 \le O\left(\frac{d}{\alpha^2 n}\right).$$

[1] We hide absolute constant multiplicative factors using the standard notations $\lesssim, O(\cdot)$. Similarly, we hide multiplicative factors at most logarithmic in n using the notation $\tilde{O}$.

[2] More precisely, their condition is $\alpha \gtrsim \frac{1}{\log n}$ for consistent estimation and $\alpha \gtrsim \frac{1}{\log\log n}$ to get the error bound $\tilde{O}(\frac{d}{\alpha^2 n})$.

[3] Here, we choose 2 as transition point between quadratic and linear penalty. Other transition points can also be used. For example, for a bit more general model where αn entries of η are bounded by some $\sigma > 0$, one can work with transition point 2σ.

[4] As a convention, we use boldface to denote random variables.

The above result improves over previous quantitative analyses of the Huber-loss estimator that require quadratic sample size $n \gtrsim d^2/\alpha^2$ to be consistent (Tsakonas et al., 2014). Other estimators developed for this model (Bhatia et al., 2017a; Suggala et al., 2019) achieve a sample-size bound nearly-linear in d at the cost of an exponential dependence on $1/\alpha$. These results require for consistent estimation a logarithmic bound on the inlier fraction $\alpha \gtrsim 1/\log d$ to achieve sample-size bound nearly-linear in d . In constrast our sample-size bound is nearly-linear in d even for any sub-polynomial inlier fraction $\alpha = 1/d^{o(1)}$. In fact, our sample-size bound and estimation-error bound is statistically optimal up to constant factors.[5]

The proof of the above theorem also applies to approximate minimizers of the Huber loss and it shows that such approximations can be computed in polynomial time.

We remark that related to (one of) our analyses of the Huber-loss estimator, we develop a fast algorithm based on (one-dimensional) median computations that achieves estimation guarantees comparable to the ones above but in linear time $O(nd)$. A drawback of this fast algorithm is that its guarantees depend (mildly) on the norm of β^*.

Several results (Candes & Tao, 2005; Candes et al., 2005; Karmalkar & Price, 2018; Diakonikolas et al., 2019b; Dalalyan & Thompson, 2019) considered settings where the noise vector is adaptively chosen by an adversary. In this setting, it is possible to obtain a unique estimate only if the fraction of outliers is smaller than $1/2$. In contrast, Theorem 1.1 implies consistency even when the fraction of corruptions tends to 1 but applies to settings where the noise vector η is fixed *before* sampling $\boldsymbol{X}$ and thus it is oblivious to the data.

Deterministic design The previous theorem makes the strong assumption that the design is Gaussian. However, it turns out that our proof extends to a much broader class of designs with the property that their columns spans are well-spread (in the sense that they don't contain vectors whose ℓ_2-mass is concentrated on a small number of coordinates, see (Guruswami et al., 2008)). In order to formulate this more general results it is convenient to move the randomness from the design to the noise vector and consider deterministic designs $X \in \mathbb{R}^{n\times d}$ with probabilistic n-dimensional noise vector $\boldsymbol{\eta}$,

$$\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}\,. \tag{1.1}$$

Here, we assume that $\boldsymbol{\eta}$ has independent, symmetrically distributed entries satisfying $\mathbb{P}\{|\boldsymbol{\eta}_i| \le 1\} \ge \alpha$ for all $i \in [n]$.

[5] In the case $\eta \sim N(0, \sigma^2 \cdot \mathrm{Id})$, it's well known that the optimal Bayesian estimator achieves expected error $\sigma^2 \cdot d/n$. For $\sigma \ge 1$, the vector η has a $\Theta(1/\sigma)$ fraction of entries of magnitude at most 1 with high probability.

This model turns out to generalize the one considered in the previous theorem. Indeed, given data following the previous model with Gaussian design and deterministic noise, we can generate data following the above model randomly subsampling the given data and multiplying with random signs.

Theorem 1.2 (Guarantees for Huber-loss estimator with general design)**.** *Let $X \in \mathbb{R}^{n\times d}$ be a deterministic matrix and let $\boldsymbol{\eta}$ be an n-dimensional random vector with independent, symmetrically distributed (about zero) entries and $\alpha = \min_{i\in[n]} \mathbb{P}\{|\boldsymbol{\eta}_i| \le 1\}$.*

Suppose that for every vector v in the column span of X and every subset $S \subseteq [n]$ with $|S| \le C \cdot d/\alpha^2$,

$$\|v_S\| \le 0.9 \cdot \|v\|\,, \tag{1.2}$$

where v_S denotes the restriction of v to the coordinates in S, and $C > 0$ is large enough absolute constant.

Then, with probability at least $1 - 2^{-d}$ over $\boldsymbol{\eta}$, for every $\beta^* \in \mathbb{R}^d$, given X and $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$, the Huber-loss estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\tfrac{1}{n}\left\|X(\beta^* - \hat{\boldsymbol{\beta}})\right\|^2 \le O\left(\frac{d}{\alpha^2 n}\right)\,.$$

In particular, Theorem 1.2 implies that under condition Eq. (1.2) and mild noise assumptions, the Huber loss estimator is consistent for $n \ge \omega\left(d/\alpha^2\right)$.

We say a vector subspace of $\mathbb{R}^n$ is *well-spread*, if all vectors from this subspace satisfy Eq. (1.2). As we only assume the column span of X to be well-spread, the result applies to a substantially broader class of design matrices X than Gaussian, naturally including those studied in (Tsakonas et al., 2014; Bhatia et al., 2017b; Suggala et al., 2019). Well-spread subspaces are closely related to ℓ_1-vs-ℓ_2 distortion[6], and have some resemblance with restricted isometry properties (RIP). Indeed both RIP and distortion assumptions have been successfully used in compressed sensing (Candes & Tao, 2005; Candes et al., 2005; Kashin & Temlyakov, 2007; Donoho, 2006) but, to the best of our knowledge, they were never observed to play a fundamental role in the context of robust linear regression. This is a key difference between our analysis and that of previous works. Understanding how crucial this well-spread property is and how to leverage it allows us to simultaneously obtain nearly optimal error guarantees, while also relaxing the design matrix assumptions. It is important to remark that a weaker version of property Eq. (1.2) is necessary as otherwise it may be *information theoretically impossible* to solve the problem. For example, if all but $o(1/\alpha)$ rows of X are zero, then with high probability all meaningful entries of $\boldsymbol{y}$ are corrupted by arbitrarily large noise.

[6] Our analysis also applies to design matrices whose column span has bounded distortion.

We derive both Theorem 1.1 and Theorem 1.2 using the same proof techniques explained in Section 2.

Remark (Small failure probability). For both Theorem 1.1 and Theorem 1.2 our proof also gives that for any $\delta \in (0,1)$, the Huber loss estimator achieves error $O\left(\frac{d+\log(1/\delta)}{\alpha^2 n}\right)$ with probability at least $1-\delta$ as long as $n \gtrsim \frac{d+\ln(1/\delta)}{\alpha^2}$, and, in Theorem 1.2, the well-spread property is satisfied for all sets $S \subseteq [n]$ of size $|S| \leq O\left(\frac{d+\log(1/\delta)}{\alpha^2}\right)$.

1.2. Results about fast algorithms

The Huber loss estimator has been extensively applied to robust regression problems (Tan et al., 2018; Tsakonas et al., 2014; Elsener et al., 2018). However, one possible drawback of such algorithm (as well as other standard approaches such as L_1-minimization (Pollard, 1991; Karmalkar & Price, 2018; Nguyen & Tran, 2013)) is the non-linear running time. In real-world applications with large, high dimensional datasets, an algorithm running in linear time $O(nd)$ may make the difference between feasible and unfeasible.

In the special case of Gaussian design, previous results (Suggala et al., 2019) already obtained estimators computable in linear time. However these algorithms require a logarithmic bound on the fraction of inliers $\alpha \gtrsim 1/\log n$. We present here a strikingly simple algorithm that achieves similar guarantees as the ones shown in Theorem 1.1 and runs in *linear time*: for each coordinate $j \in [d]$ compute the median $\hat{\boldsymbol{\beta}}_j$ of $\boldsymbol{y}_1/\boldsymbol{X}_{1j}, \ldots, \boldsymbol{y}_n/\boldsymbol{X}_{nj}$ subtract the resulting estimation $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and repeat, logarithmically many times, with fresh samples.

Theorem 1.3 (Guarantees for fast estimator with Gaussian design)**.** *Let $\eta \in \mathbb{R}^n$ and $\beta^* \in \mathbb{R}^d$ be deterministic vectors. Let $\boldsymbol{X}$ be a random n-by-d matrix with iid standard Gaussian entries $\boldsymbol{X}_{ij} \sim N(0,1)$.*

Let α be the fraction of entries in η of magnitude at most 1, and let $\Delta \geq 10 + \|\beta^*\|$. Suppose that

$$n \geq C \cdot \frac{d}{\alpha^2} \cdot \ln \Delta \cdot (\ln d + \ln\ln \Delta),$$

where C is a large enough absolute constant.

Then, there exists an algorithm that given Δ, $\boldsymbol{X}$ and $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$ as input, in time[7] $O(nd)$ finds a vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ such that

$$\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2 \leq O\left(\frac{d}{\alpha^2 n} \cdot \log d\right),$$

with probability at least $1 - d^{-10}$.

The algorithm in Theorem 1.3 requires knowledge of an upper bound Δ on the norm of the parameter vector. The sample complexity of the estimator has logarithmic dependency on this upper bound. This phenomenon is a consequence of the iterative nature of the algorithm and also appears in other results (Suggala et al., 2019).

Theorem 1.3 also works for non-spherical settings Gaussian design matrix and provides nearly optimal error convergence with nearly optimal sample complexity, albeit with running time $\tilde{O}(nd^2)$. The algorithm doesn't require prior knowledge of the covariance matrix Σ. In these settings, even though time complexity is not linear in d, it is linear in n, and if n is considerably larger than d, the algorithm may be very efficient.

Sparse linear regression For spherical Gaussian design, the median-based algorithm introduced above can naturally be extended to the sparse settings, yielding the following theorem.

Theorem 1.4 (Guarantees of fast estimator for sparse regression with Gaussian design)**.** *Let $\eta \in \mathbb{R}^n$ and $\beta^* \in \mathbb{R}^d$ be deterministic vectors, and assume that β^* has at most $k \leq d$ nonzero entries. Let $\boldsymbol{X}$ be a random n-by-d matrix with iid standard Gaussian entries $\boldsymbol{X}_{ij} \sim N(0,1)$.*

Let α be the fraction of entries in η of magnitude at most 1, and let $\Delta \geq 10 + \|\beta^*\|$. Suppose that

$$n \geq C \cdot \frac{k}{\alpha^2} \cdot \ln \Delta \cdot (\ln d + \ln\ln \Delta),$$

where C is a large enough absolute constant.

Then, there exists an algorithm that given k, Δ, $\boldsymbol{X}$ and $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$ as input, in time $O(nd)$ finds a vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ such that

$$\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2 \leq O\left(\frac{k}{\alpha^2 n} \cdot \log d\right),$$

with probability at least $1 - d^{-10}$.

2. Techniques

In this section we discuss the model from Theorem 1.2 (with deterministic X and random $\boldsymbol{\eta}$). The model from Theorem 1.1 (with Gaussian $\boldsymbol{X}$ and deterministic η) can be studied in a very similar way.

Recall our linear regression model,

$$\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}, \tag{2.1}$$

where we observe (a realization of) the random vector $\boldsymbol{y}$, the matrix $X \in \mathbb{R}^{n\times d}$ is a known design, the vector $\beta^* \in \mathbb{R}^n$ is the unknown parameter of interest, and the noise vector $\boldsymbol{\eta}$ has independent, symmetrically distributed[8] coordinates

[7] By time we mean number of arithmetic operations and comparisons of entries of $\boldsymbol{y}$ and $\boldsymbol{X}$. We do not take bit complexity into account.

[8] The distributions of the coordinates are not known to the algorithm designer and can be non-identical.

with[9] $\alpha = \min_{i\in[n]} \mathbb{P}\{|\boldsymbol{\eta}_i| \le 1\}$.

To simplify notation in our proofs, we assume $\frac{1}{n}X^\mathsf{T}X = \mathrm{Id}$. (For general X, we can ensure this property by orthogonalizing and scaling the columns of X.)

We consider the *Huber loss estimator* $\hat{\boldsymbol{\beta}}$, defined as a minimizer of the *Huber loss* $\boldsymbol{f}$,

$$\boldsymbol{f}(\beta) := \tfrac{1}{n}\sum_{i=1}^{n} \Phi[(X\beta - \boldsymbol{y})_i],$$

where $\Phi\colon \mathbb{R} \to \mathbb{R}_{\ge 0}$ is the *Huber penalty*,[10]

$$\Phi[t] = \begin{cases} \frac{1}{2}t^2 & \text{if } |t| \le 2, \\ 2|t| - 2 & \text{otherwise.} \end{cases}$$

2.1. Statistical guarantees from strong convexity

In order to prove statistical guarantees for this estimator, we follow a well-known approach that applies to a wide range of estimators based on convex optimization (see (Negahban et al., 2009) for a more general exposition), which also earlier analyses of the Huber loss estimator (Tsakonas et al., 2014) employ. This approach has two ingredients: (1) an upper bound on the norm of the gradient of the loss function $\boldsymbol{f}$ at the desired parameter β^* and (2) a lower bound on the strong-convexity curvature parameter of $\boldsymbol{f}$ within a ball centered at β^*. Taken together, these ingredients allow us to construct a global lower bound for $\boldsymbol{f}$ that implies that all (approximate) minimizers of $\boldsymbol{f}$ are close to β^*.

An important feature of this approach is that it only requires strong convexity to hold locally around β^*. (Due to its linear parts, the Huber loss function doesn't satisfy strong convexity globally.) It turns out that the radius of strong convexity we can prove is the main factor determining the strength of the statistical guarantee we obtain. Indeed, the reason why previous analyses[11] of the Huber loss estimator (Tsakonas et al., 2014) require quadratic sample size $n \gtrsim (d/\alpha)^2$ to ensure consistency is that they can establish strong convexity only within inverse-polynomial radius $\Omega(1/\sqrt{d})$ even for Gaussian $X \sim N(0,1)^{n\times d}$. In contrast, our analysis gives consistency for any super-linear sample size $n = \omega(d/\alpha^2)$ for Gaussian X because we can establish strong convexity within constant radius.

[9] The value of α need not be known to the algorithm designer and only affects the error guarantees of the algorithms.

[10] Here, in order to streamline the presentation, we choose $\{\pm 2\}$ as the transition points between quadratic and linear penalty. Changing these points to $\{\pm 2\delta\}$ is achieved by scaling $t \mapsto \delta^2\Phi(t/\delta)$.

[11] We remark that the results in (Tsakonas et al., 2014) are phrased asymptotically, i.e., fixed d and $n \to \infty$. Therefore, a radius bound independent of n is enough for them. However, their proof is quantiative and yields a radius bound of $1/\sqrt{d}$ a we will discuss.

Compared to the strong-convexity bound, which we disucss next, the gradient bound is straightforward to prove. The gradient of the Huber loss at β^* for response vector $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$ takes the following form,

$$\nabla\boldsymbol{f}(\beta^*) = \tfrac{1}{n}\sum_{i=1}^{n} \Phi'[\boldsymbol{\eta}_i]\cdot x_i \quad \text{with} \quad \Phi'[t] = \operatorname{sign}(t)\cdot\min\{|t|, 2\},$$

where $x_1, \dots, x_n \in \mathbb{R}^d$ form the rows of X. Since $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n$ are independent and symmetrically distributed, the random variables $\Phi'[\boldsymbol{\eta}_i]$ are zero-mean, independent and bounded by 2 in absolute value. Now, for a unit vector $u \in \mathbb{R}^d$, using Hoeffding's inequality, we get with probability at least $1 - e^{-t}$,

$$\langle \nabla\boldsymbol{f}(\beta^*), u\rangle \le \frac{1}{n}\cdot O\left(\sqrt{t}\cdot\|Xu\|\right) \le O\left(\sqrt{t/n}\right),$$

where we use the assumption $\frac{1}{n}X^\mathsf{T}X = \mathrm{Id}$. Finally, using a union bound over $1/2$-net in the unit sphere in $\mathbb{R}^d$, we get

$$\|\boldsymbol{f}(\beta^*)\| \le O\left(\sqrt{d/n}\right)$$

with high probability.

Proving local strong convexity for Huber loss For response vector $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$ and arbitrary $u \in \mathbb{R}^d$, the Hessian[12] of the Huber loss at $\beta^* + u$ has the following form,

$$H\boldsymbol{f}(\beta^* + u) = \tfrac{1}{n}\sum_{i=1}^{n} \Phi''[(Xu)_i - \boldsymbol{\eta}_i]\cdot x_i x_i^\mathsf{T} \quad \text{with} \quad \Phi''[t] = [\![|t| \le 2]\!].$$

Here, $[\![\cdot]\!]$ is the Iverson bracket (0/1 indicator). To prove local strong convexity within radius R, we are to lower bound $\langle u, H\boldsymbol{f}(\beta^* + u)u\rangle$ uniformly over all vectors $u \in \mathbb{R}^d$ with $\|u\| \le R$.

We do not attempt to exploit any cancellations between Xu and $\boldsymbol{\eta}$ and work with the following lower bound $\boldsymbol{M}(u)$ for the Hessian,

$$H\boldsymbol{f}(\beta^* + u) \ge \boldsymbol{M}(u) := \tfrac{1}{n}\sum_{i=1}^{n} [\![|\langle x_i, u\rangle| \le 1]\!]\cdot[\![|\boldsymbol{\eta}_i| \le 1]\!]\cdot x_i x_i^\mathsf{T}. \tag{2.2}$$

Here, $\ge$ denotes the Löwner order.

[12] The second derivative of the Huber penalty doesn't exit at the transition points $\{\pm 2\}$ between its quadratic and linear parts. Nevertheless, the second derivative exists as an L_1-function in the sense that $\Phi'[b] - \Phi'[a] = \int_a^b [\![|t| \le 2]\!]\,\mathrm{d}t$ for all $a, b \in \mathbb{R}$. This property is enough for our purposes.

It's instructive to first consider $u = 0$. Here, the above lower bound for the Hessian satisfies,

$$\mathbb{E}\,[\boldsymbol{M}(0)] = \tfrac{1}{n}\sum_{i=1}^{n} \mathbb{P}\{|\boldsymbol{\eta}_i| \le 1\} \cdot x_i x_i^\mathsf{T} \succeq \alpha \mathrm{Id}\,.$$

Using standard (matrix) concentration inequalities, we can also argue that this random matrix is close to its expectation with high-probability if $n \ge \tilde{O}(d/\alpha)$ under some mild assumption on X (e.g., that the row norms are balanced so that $\|x_1\|, \ldots, \|x_n\| \le O(\sqrt{d})$).

The main remaining challenge is dealing with the quantification over u. Earlier analyses (Tsakonas et al., 2014) observe that the Hessian lower bound $\boldsymbol{M}(\cdot)$ is constant over balls of small enough radius. Concretely, for all $u \in \mathbb{R}^d$ with $\|u\| \le 1/\max_i \|x_i\|$, we have

$$\boldsymbol{M}(u) = \boldsymbol{M}(0)\,,$$

because $|\langle x_i, u\rangle| \le \|x_i\| \cdot \|u\| \le 1$ by Cauchy-Schwarz. Thus, strong convexity with curvature parameter α within radius $1/\max_i \|x_i\|$ follows from the aforementioned concentration argument for $\boldsymbol{M}(0)$. However, since $\max_i \|x_i\| \ge \sqrt{d}$, this argument cannot give a better radius bound than $1/\sqrt{d}$, which leads to a quadratic sample-size bound $n \gtrsim d^2/\alpha^2$ as mentioned before.

For balls of larger radius, the lower bound $\boldsymbol{M}(\cdot)$ can vary significantly. For illustration, let us consider the case $\boldsymbol{\eta} = 0$ and let us denote the Hessian lower bound by $M(\cdot)$ for this case. (The deterministic choice of $\boldsymbol{\eta} = 0$ would satisfy all of our assumptions about $\boldsymbol{\eta}$.) As we will see, a uniform lower bound on $\langle u, M(u)u\rangle$ over a ball of radius $R > 0$ implies that the column span of X is well-spread in the sense that every vector v in this subspace has a constant fraction of its ℓ_2 mass on entries with squared magnitude at most a $1/R^2$ factor times the average squared entry of v. (Since we aim for $R > 0$ to be a small constant, the number $1/R^2$ is a large constant.) Concretely,

$$\begin{aligned}
&\min_{\|u\|=R} \tfrac{1}{R^2}\langle u, M(u)u\rangle \\
&\quad= \min_{\|u\|=R} \tfrac{1}{R^2} \cdot \tfrac{1}{n}\sum_{i=1}^{n} [\![\langle x_i, u\rangle^2 \le 1]\!] \cdot \langle x_i, u\rangle^2 \\
&\quad= \min_{v \in \mathrm{col.span}(X)} \tfrac{1}{\|v\|^2}\sum_{i=1}^{n} \Big[\!\Big[R^2 \cdot v_i^2 \le \tfrac{1}{n}\|v\|^2\Big]\!\Big] \cdot v_i^2 \\
&\quad=: \kappa_R\,.
\end{aligned} \tag{2.3}$$

(The second step uses our assumption $X^\mathsf{T}X = \mathrm{Id}$.)

It turns out that the above quantity κ_R in Eq. (2.3) indeed captures up to constant factors the radius and curvature parameter of strong convexity of the Huber loss function around β^* for $\boldsymbol{\eta} = 0$. In this sense, the well-spreadness of the column span of X is required for the current approach of analyzing the Huber-loss estimator based on strong convexity. The quantity κ_R in Eq. (2.3) is closely related to previously studied notions of well-spreadness for subspaces (Guruswami et al., 2008; 2010) in the context of compressed sensing and error-correction over the reals.

Finally, we use a covering argument to show that a well-spread subspace remains well-spread even when restricted to a random fraction of the coordinates (namely the coordinates satisfying $|\boldsymbol{\eta}_i| \le 1$). This fact turns out to imply the desired lower bound on the local strong convexity parameter. Concretely, if the column space of X is well-spread in the sense of Eq. (2.3) with parameter κ_R for some $R \ge \tilde{O}_{\kappa_R}(\frac{d}{\alpha n})^{1/2}$, we show that the Huber loss function is locally $\Omega(\alpha \cdot \kappa_R)$-strong convex at β^* within radius $\Omega(R)$. Recall that we are interested in the regime $n \gtrsim d/\alpha^2$ (otherwise, consistent estimation is impossible). In this case, with high probability Gaussian X satisfies $\kappa_R \ge 0.1$ even for constant R.

Final error bound The aforementioned general framework for analyzing estimators via strong convexity allows us to bound the error $\|\hat{\boldsymbol{\beta}} - \beta^*\|$ by the norm of the gradient $\|\nabla \boldsymbol{f}(\beta^*)\|$ divided by the strong-convexity parameter, assuming that this upper bound is smaller than the strong-convexity radius.

Consequently, for the case that our design X satisfies $\kappa_R \ge 0.1$ (corresponding to the setting of Theorem 1.2), the previously discussed gradient bound and strong-convexity bound together imply that, with high probability over $\boldsymbol{\eta}$, the error bound satisfies

$$\|\hat{\boldsymbol{\beta}} - \beta^*\| \le \underbrace{O\left(\sqrt{\frac{d}{n}}\right)}_{\text{gradient bound}} \cdot \underbrace{O\left(\frac{1}{\alpha}\right)}_{\text{strong-convexity bound}} = O\left(\frac{d}{\alpha^2 n}\right)^{1/2},$$

assuming $R \gtrsim \sqrt{d/\alpha^2 n}$.

2.2. Huber-loss estimator and high-dimensional medians

We discuss here some connections between high-dimensional median computations and efficient estimators such as Huber loss or the LAD estimator. This connection leads to a better understanding of *why* these estimators are not susceptible to heavy-tailed noise. Through this analysis we also obtain guarantees similar to the ones shown in Theorem 1.2.

Recall our linear regression model $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$ as in Eq. (2.1). The noise vector $\boldsymbol{\eta}$ has independent, symmetrically distributed coordinates with $\alpha = \min_{i\in[n]} \mathbb{P}\{|\boldsymbol{\eta}_i| \le 1\}$. We further assume the noise entries to satisfy

$$\forall t \in [0,1], \quad \mathbb{P}\,(|\boldsymbol{\eta}_i| \le t) \ge \Omega(\alpha \cdot t)\,.$$

This can be assumed without loss of generality as, for example, we may simply add a Gaussian vector $\boldsymbol{w} \sim N(0, \mathrm{Id}_n)$ (independent of $\boldsymbol{y}$) to $\boldsymbol{y}$ (after this operation parameter α changes only by a constant factor).

The one dimensional case: median algorithm To understand how to design an efficient algorithm robust to $\left(1-\sqrt{d/n}\right)\cdot n$ corruptions, it is instructive to look into the simple settings of one dimensional Gaussian design $\boldsymbol{X} \sim N(0, \mathrm{Id}_n)$. Given samples $(\boldsymbol{y}_1, \boldsymbol{X}_1), \ldots, (\boldsymbol{y}_n, \boldsymbol{X}_n)$ for any $i \in [n]$ such that $|\boldsymbol{X}_i| \geq 1/2$, consider

$$\boldsymbol{y}_i/\boldsymbol{X}_i = \beta^* + \boldsymbol{\eta}_i/\boldsymbol{X}_i \,.$$

By *obliviousness* the random variables $\boldsymbol{\eta}'_i = \boldsymbol{\eta}_i/\boldsymbol{X}_i$ are symmetric about 0 and for any $0 \leq t \leq 1$, still satisfy $\mathbb{P}(-t \leq \boldsymbol{\eta}'_i \leq t) \geq \Omega(\alpha \cdot t)$. Surprisingly, this simple observation is enough to obtain an optimal robust algorithm. Standard tail bounds show that with probability $1-\exp\left\{-\Omega\left(\alpha^2\cdot\varepsilon^2\cdot n\right)\right\}$ the median $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{y}_1/\boldsymbol{X}_1, \ldots, \boldsymbol{y}_n/\boldsymbol{X}_n$ falls in the interval $[-\varepsilon+\beta^*, +\varepsilon+\beta^*]$ for any $\varepsilon \in [0,1]$. Hence, setting $\varepsilon \gtrsim 1/\sqrt{\alpha^2\cdot n}$ we immediately get that with probability at least 0.999, $\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2 \leq \varepsilon^2 \leq O(1/(\alpha^2\cdot n))$.

The high-dimensional case: from the median to the Huber loss In the one dimensional case, studying the median of the samples $\boldsymbol{y}_1/\boldsymbol{X}_1, \ldots \boldsymbol{y}_n/\boldsymbol{X}_n$ turns out to be enough to obtain optimal guarantees. The next logical step is to try to construct a similar argument in high dimensional settings. However, the main problem here is that high dimensional analogs of the median are usually computationally inefficient (e.g. Tukey median (Tukey, 1975)) and so this doesn't seem to be a good strategy to design efficient algorithms. Still in our case one such function provides fundamental insight.

We start by considering the sign pattern of $X\beta^*$, we do not fix any property of X yet. Indeed, note that the median satisfies $\sum_{i\in[n]} \operatorname{sign}\left(\boldsymbol{y}_i/X_i - \hat{\boldsymbol{\beta}}\right) \approx 0$ and so $\sum_{i\in[n]} \operatorname{sign}\left(\boldsymbol{y}_i - \hat{\boldsymbol{\beta}}X_i\right)\operatorname{sign}(X_i) \approx 0$. So a natural generalization to high dimensions is the following candidate estimator

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\beta\in\mathbb{R}^d} \max_{u\in\mathbb{R}^d} \left|\tfrac{1}{n}\langle \operatorname{sign}(\boldsymbol{y} - X\beta), \operatorname{sign}(Xu)\rangle\right| \,. \tag{2.4}$$

Such an estimator may be inefficient to compute, but nonetheless it is instructive to reason about it. We may assume X, β^* are fixed, so that the randomness of the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ only depends on $\boldsymbol{\eta}$. Since for each $i \in [n]$, the distribution of $\boldsymbol{\eta}_i$ has median zero and as there are at most $n^{O(d)}$ sign patterns in $\left\{\operatorname{sign}(Xu) \,\middle|\, u \in \mathbb{R}^d\right\}$, standard ε-net arguments show that with high probability

$$\max_{u\in\mathbb{R}^d} \tfrac{1}{n}\left|\langle \operatorname{sign}\left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right), \operatorname{sign}(Xu)\rangle\right| \leq \tilde{O}\left(\sqrt{d/n}\right) \,, \tag{2.5}$$

and hence

$$\max_{u\in\mathbb{R}^d} \tfrac{1}{n}\left|\langle \operatorname{sign}\left(\boldsymbol{\eta} + X\left(\beta^* - \hat{\boldsymbol{\beta}}\right)\right), \operatorname{sign}(Xu)\rangle\right| \leq \tilde{O}\left(\sqrt{d/n}\right) \,.$$

Consider $\boldsymbol{g}(z) = \frac{1}{n}\langle \operatorname{sign}(\boldsymbol{\eta} + Xz), \operatorname{sign}(Xz)\rangle \leq \tilde{O}(d/n)$ for $z \in \mathbb{R}^d$. Now the central observation is that for any $z \in \mathbb{R}^d$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\eta}} \boldsymbol{g}(z) &= \tfrac{1}{n}\sum_{i\in[n]} \mathop{\mathbb{E}}_{\boldsymbol{\eta}} \operatorname{sign}(\boldsymbol{\eta}_i + \langle X_i, z\rangle)\cdot\operatorname{sign}(\langle X_i, z\rangle) \\
&\geq \tfrac{1}{n}\sum_{i\in[n]} \mathbb{P}\left(0 \geq \operatorname{sign}(\langle X_i, z\rangle)\cdot\boldsymbol{\eta}_i \geq -|\langle X_i, z\rangle|\right) \\
&\geq \tfrac{1}{n}\sum_{i\in[n]} \Omega(\alpha)\cdot\min\{1, |\langle X_i, z\rangle|\} \,.
\end{aligned}$$

By triangle inequality $\mathbb{E}\,\boldsymbol{g}(z) \leq |\boldsymbol{g}(z)| + |\boldsymbol{g}(z) - \mathbb{E}\,\boldsymbol{g}(z)|$ and using a similar argument as in Eq. (2.5), with high probability, for any $z \in \mathbb{R}^d$,

$$|\boldsymbol{g}(z) - \mathbb{E}\,\boldsymbol{g}(z)| \leq \tilde{O}\left(\sqrt{d/n}\right) \,.$$

Denote with $\boldsymbol{z} := \beta^* - \hat{\boldsymbol{\beta}} \in \mathbb{R}^d$. Consider $\boldsymbol{g}(z)$, thinking of $z \in \mathbb{R}^d$ as a *fixed* vector. This allows us to easily study $\mathbb{E}_{\boldsymbol{\eta}}\,\boldsymbol{g}(z)$. On the other hand, since our bounds are based on ε-net argument, we don't have to worry about the dependency of $\boldsymbol{z}$ on $\boldsymbol{\eta}$.

So without any constraint on the measurement X we derived the following inequality:

$$\tfrac{1}{n}\sum_{i\in[n]} \min\{1, |\langle X_i, \boldsymbol{z}\rangle|\} \leq \tilde{O}\left(\sqrt{d/(\alpha^2\cdot n)}\right) \,.$$

Now, our well-spread condition Eq. (1.2) will allow us to relate $\frac{1}{n}\sum_{i\in[n]} \min\{1, |\langle X_i, \boldsymbol{z}\rangle|\}$ with $\frac{1}{n}\sum_{i\in[n]}\langle X_i, \boldsymbol{z}\rangle^2$ and thus obtain a bound of the form

$$\frac{1}{n}\left\|X\left(\beta^* - \hat{\boldsymbol{\beta}}\right)\right\|^2 \leq \tilde{O}\left(d/(\alpha^2 n)\right) \,. \tag{2.6}$$

So far we glossed over the fact that Eq. (2.4) may be hard to compute, however it is easy to see that we can replace such estimator with some well-known efficient estimators and keep a similar proof structure. For instance, one could expect the LAD estimator

$$\hat{\boldsymbol{\beta}} = \min_{\beta\in\mathbb{R}^d} \|\boldsymbol{y} - X\beta\|_1 \tag{2.7}$$

to obtain comparable guarantees. For fixed d and α and n tending to infinity this is indeed the case, as we know by (Pollard, 1991) that such estimator recovers β^*. The

Huber loss function also turns out to be a good proxy for Eq. (2.4). Let $\boldsymbol{g}(u) := \frac{1}{n}\sum_{i\in[n]} \langle \Phi_h'(\boldsymbol{\eta}_i + \langle X_i, u\rangle), Xu\rangle$ where $\Phi_h : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is the Huber penalty function and $\boldsymbol{z} = \beta^* - \hat{\boldsymbol{\beta}}$. Exploiting *only* first order optimality conditions on $\hat{\boldsymbol{\beta}}$ one can show

$$\mathbb{E}\,\boldsymbol{g}(z) \leq |\boldsymbol{g}(z) - \mathbb{E}\,\boldsymbol{g}(z)| \leq \tilde{O}\left(\sqrt{d/n}\right),$$

using a similar argument as the one mentioned for Eq. (2.5). Following a similar proof structure as the one sketched above, we can obtain a bound similar to Eq. (2.6). Note that this approach crucially exploits the fact that the noise $\boldsymbol{\eta}$ has median zero but does not rely on symmetry and so can successfully obtain a good estimate of $X\beta^*$ under *weaker* noise assumptions.

2.3. Fast algorithms for Gaussian design

The one dimensional median approach introduced above can be directly extended to high dimensional settings. This essentially amounts to repeating the procedure for each coordinate, thus resulting in an extremely simple and efficient algorithm. More concretely:

Algorithm 1 Multivariate linear regression iteration via median

Input: (y, X) where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times d}$.
for all $j \in [d]$ **do**
  for all $i \in [n]$ **do**
    Compute $z_{ij} = \frac{y}{X_{ij}}$.
  end for
  Let $\hat{\beta}_j$ be the median of $\{z_{ij}\}_{i\in[n]}$.
end for
Return $\hat{\beta} := \left(\hat{\beta}_1, \ldots, \hat{\beta}_d\right)^\mathsf{T}$.

If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim N(0, \mathrm{Id}_d)$, the analysis of the one dimensional case shows that with high probability, for each $j \in [d]$, the algorithm returns $\hat{\boldsymbol{\beta}}_j$ satisfying $(\beta_j^* - \hat{\boldsymbol{\beta}}_j)^2 \leq O\left(\frac{1+\|\beta^*\|^2}{\alpha^2} \cdot \log d\right)$. Summing up all the coordinate-wise errors, Algorithm 1 returns a $O\left(\frac{d(1+\|\beta^*\|^2)}{\alpha^2} \cdot \log d\right)$-close estimation. This is better than a trivial estimate, but for large $\|\beta^*\|$ it is far from the $O(d \cdot \log d/(\alpha^2 \cdot n))$ error guarantees we aim for. However, using bootstrapping we can indeed improve the accuracy of the estimate. It suffices to iterate $\log \|\beta^*\|$ many times.

Algorithm 2 Multivariate linear regression via median

Input: (y, X, Δ) where $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$ and Δ is an upper bound to $\|\beta^*\|$.
Randomly partition the samples $y_1, \ldots, y_n$ in $t := \Theta(\log \Delta)$ sets $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_t$, such that all $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{t-1}$ have sizes $\Theta\left(\frac{n}{\log \Delta}\right)$ and $\boldsymbol{S}_t$ has size $\lfloor n/2 \rfloor$.
for all $i \in [t]$ **do**
  Run Algorithm 1 on input

$$\left(y_{\boldsymbol{S}_i} - X_{\boldsymbol{S}_i}\left(\sum_{j<i-1} \hat{\boldsymbol{\beta}}^{(j)}\right), X_{\boldsymbol{S}_i}\right),$$

  and let $\hat{\boldsymbol{\beta}}^{(i)}$ be the resulting estimator.
end for
Return $\hat{\beta} := \left(\hat{\beta}_1, \ldots, \hat{\beta}_d\right)^\mathsf{T}$.

As mentioned in Section 1.2, Algorithm 2 requires knowledge of an upper bound Δ on the norm of β^*. The algorithm only obtains meaningful guarantees for

$$n \gtrsim \frac{d}{\alpha^2} \log \Delta \,(\log d + \log\log \Delta)$$

and as such works with nearly optimal (up to polylogarithmic terms) sample complexity whenever $\|\beta^*\|$ is polynomial in d/α^2.

In these settings, since each iteration i requires $O(|\boldsymbol{S}_i| \cdot d)$ steps, Algorithm 2 runs in linear time $O(n \cdot d)$ and outputs a vector $\hat{\boldsymbol{\beta}}$ that with high probability satisfies

$$\left\|\hat{\boldsymbol{\beta}} - \beta^*\right\|^2 \leq O\left(\frac{d}{\alpha^2 \cdot n} \cdot \log d\right).$$

Remark (On learning the norm of β^*). As was noticed in (Suggala et al., 2019), one can obtain a rough estimate of the norm of η by projecting $\boldsymbol{y}$ onto the orthogonal complement of the columns span of $\boldsymbol{X}_{[n/2]}$. Since the ordinary least square estimator obtains an estimate with error $\Delta = O(\sqrt{d}\,\|\eta\|/n)$ with high probability, if $\|\eta\|$ is polynomial in the number of samples, we obtain a vector $\hat{\boldsymbol{\beta}}_{LS}$ such that $\left\|\beta^* - \hat{\boldsymbol{\beta}}_{LS}\right\| \leq \Delta = n^{O(1)}$. The median algorithm can then be applied on $\left(\boldsymbol{y} = \boldsymbol{X}_{[n]\setminus[n/2]}(\beta^* - \hat{\boldsymbol{\beta}}_{LS}) + \eta,\ \boldsymbol{X}_{[n]\setminus[n/2]},\ \Delta\right)$. Note that since $\boldsymbol{X}_{[n/2]}$ and $\boldsymbol{X}_{[n]\setminus[n/2]}$ are independent, $\beta^* - \hat{\boldsymbol{\beta}}_{LS}$ is independent of $\boldsymbol{X}_{[n]\setminus[n/2]}$.

Acknowledgements

The authors thank the anonymous reviewers for useful comments. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 815464).

References

Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. Consistent robust regression. In *NIPS*, pp. 2107–2116, 2017a.

Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. Consistent robust regression. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2110–2119. Curran Associates, Inc., 2017b. URL http://papers.nips.cc/paper/6806-consistent-robust-regression.pdf.

Candes, E. and Tao, T. Decoding by linear programming, 2005.

Candes, E., Romberg, J., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements, 2005.

Charikar, M., Steinhardt, J., and Valiant, G. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 47–60, 2017.

Dalalyan, A. and Thompson, P. Outlier-robust estimation of a sparse linear model using l_1-penalized huber's m-estimator. In *Advances in Neural Information Processing Systems*, pp. 13188–13198, 2019.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019a.

Diakonikolas, I., Kong, W., and Stewart, A. Efficient algorithms and lower bounds for robust linear regression. In Chan, T. M. (ed.), *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pp. 2745–2754. SIAM, 2019b. doi: 10.1137/1.9781611975482.170. URL https://doi.org/10.1137/1.9781611975482.170.

Donoho, D. L. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Elsener, A., van de Geer, S., et al. Robust low-rank matrix estimation. *The Annals of Statistics*, 46(6B):3481–3509, 2018.

Guruswami, V., Lee, J. R., and Wigderson, A. Euclidean sections of with sublinear randomness and error-correction over the reals. In *APPROX-RANDOM*, volume 5171 of *Lecture Notes in Computer Science*, pp. 444–454. Springer, 2008.

Guruswami, V., Lee, J. R., and Razborov, A. A. Almost euclidean subspaces of l_1^n VIA expander codes. *Combinatorica*, 30(1):47–68, 2010.

Haupt, J., Bajwa, W. U., Rabbat, M., and Nowak, R. Compressed sensing for networked data. *IEEE Signal Processing Magazine*, 25(2):92–101, 2008.

Huber, P. J. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

Karmalkar, S. and Price, E. Compressed sensing with adversarial sparse noise via l1 regression. *arXiv preprint arXiv:1809.08055*, 2018.

Karmalkar, S., Klivans, A. R., and Kothari, P. List-decodable linear regression. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 7423–7432, 2019. URL http://papers.nips.cc/paper/8961-list-decodable-linear-regression.

Kashin, B. S. and Temlyakov, V. N. A remark on compressed sensing. *Mathematical notes*, 82(5):748–755, 2007.

Klivans, A. R., Kothari, P. K., and Meka, R. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pp. 1420–1430, 2018. URL http://proceedings.mlr.press/v75/klivans18a.html.

Liu, L., Shen, Y., Li, T., and Caramanis, C. High dimensional robust sparse regression. *arXiv preprint arXiv:1805.11643*, 2018.

Liu, L., Li, T., and Caramanis, C. High dimensional robust m-estimation: Arbitrary corruption and heavy tails, 2019.

Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *NIPS*, pp. 1348–1356. Curran Associates, Inc., 2009.

Nguyen, N. H. and Tran, T. D. Exact recoverability from dense corrupted observations via l_1-minimization. *IEEE transactions on information theory*, 59(4), 2013.

Pollard, D. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

Raghavendra, P. and Yau, M. List decodable learning via sum of squares. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pp. 161–180, 2020. doi: 10.1137/1.9781611975994.10. URL https://doi.org/10.1137/1.9781611975994.10.

Rousseeuw, P. J. and Leroy, A. M. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

Suggala, A. S., Bhatia, K., Ravikumar, P., and Jain, P. Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pp. 2892–2897, 2019. URL http://proceedings.mlr.press/v99/suggala19a.html.

Sun, Q., Zhou, W.-X., and Fan, J. Adaptive huber regression. *Journal of the American Statistical Association*, pp. 1–24, 2019.

Tan, K. M., Sun, Q., and Witten, D. Robust sparse reduced rank regression in high dimensions. *arXiv preprint arXiv:1810.07913*, 2018.

Tsakonas, E., Jaldén, J., Sidiropoulos, N. D., and Ottersten, B. Convergence of the huber regression m-estimate in the presence of dense outliers. *IEEE Signal Processing Letters*, 21(10):1211–1214, 2014.

Tukey, J. W. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pp. 523–531, 1975.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2008.