Consistent regression when oblivious outliers overwhelm*

Tommaso d'Orsi[†] Gleb Novikov[‡] David Steurer[§]

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 $\alpha \geqslant 1/\log n$. 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 [TJSO14], 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.

*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).

[†]ETH Zürich.

[‡]ETH Zürich.

[§]ETH Zürich.

Contents

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 [RL05], image recognition [WYG+08], and sensor networks [HBRN08], 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 [DKK+19, CSV17] 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 [CT05, CRT05, KKM18, KKK19, LLC19, LSLC18, KP18, DT19, RY20]. 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 assumptions can be used to model oblivious adversarial corruptions. [SZF19] assume the noise distribution satisfies $\mathbb{E}\left[\boldsymbol{\eta}_i \mid x_i\right] = 0$ and $\mathbb{E}\left[\left|\boldsymbol{\eta}_i\right|^{1+\delta}\right] < \infty$ for some $0 \leqslant \delta \leqslant 1$, and show that if X has constant condition number, then (a modification of) the Huber loss estimator [Hub64] is consistent for[1] $n \geqslant \tilde{O}((\|X\|_\infty \cdot d)^{(1+\delta)/2\delta})$ (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

[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 $\tilde{O}$.

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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim N(0, \Sigma)$. [BJKK17a] 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 [SBRJ19], 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, [TJSO14] showed that with high probability the Huber loss estimator is consistent for $n \geqslant \tilde{O}(d^2/\alpha^2)$, 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 [Pol91], 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* $f : \mathbb{R}^d \to \mathbb{R}_{\geqslant 0}$,

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

[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 $\tilde{O}(\frac{d}{\alpha^2 n})$.

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

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

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 [TJSO14, BJKK17a, SBRJ19].

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 $n \geqslant C \cdot d/\alpha^2$, 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

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

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 [TJSO14]. Other estimators developed for this model [BJKK17b, SBRJ19] 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 [CT05, CRT05, KP18, DKS19, DT19] considered settings where the noise vector is adaptively chosen by an adversary. In this setting, it is possible to obtain a unique

[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.

[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 $\sigma \geqslant 1$, the vector η has a $\Theta(1/\sigma)$ fraction of entries of magnitude at most 1 with high probability.

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 [GLW08]). 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 $\mathbb{P}\{|\boldsymbol{\eta}_i| \leqslant 1\} \geqslant \alpha$ for all $i \in [n]$.

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 (see Appendix A for more details).

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

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

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

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

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

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

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 [TJSO14, BJKK17a, SBRJ19]. 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 [CT05, CRT05, KT07, Don06] 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 (see Lemma A.5).

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 $|S| \leqslant O\left(\frac{d+\log(1/\delta)}{\alpha^2}\right)$.

1.2 Results about fast algorithms

The Huber loss estimator has been extensively applied to robust regression problems [TSW18, TJSO14, EvdG+18]. However, one possible drawback of such algorithm (as well as other standard approaches such as L_1-minimization [Pol91, KP18, NT13]) 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 [SBRJ19] 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 $\Delta \geqslant 10 + \|\beta^*\|$. Suppose that

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

where C is a large enough absolute constant.

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

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*

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

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 [SBRJ19].

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* $k \leqslant d$ *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* $\Delta \geqslant 10 + \|\beta^*\|$. *Suppose that*

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

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*

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

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

[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.

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

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

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

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 [NRWY09] for a more general exposition), which also earlier analyses of the Huber loss estimator [TJSO14] 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^*. (See Theorem 3.1 for the formal statement.)

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

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

[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 $t \mapsto \delta^2\Phi(t/\delta)$.

why previous analyses[11] of the Huber loss estimator [TJSO14] 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.

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 $1 - e^{-t}$,

$$\langle \nabla \boldsymbol{f}(\beta^*), u\rangle \leqslant \frac{1}{n}\cdot O\Big(\sqrt{t}\cdot\|Xu\|\Big)\,.$$

Finally, using a union bound over 1/2-net in unit sphere in $\mathbb{R}^n$, we get

$$\big\|\boldsymbol{f}(\beta^*)\big\| \leqslant O\Big(\sqrt{d/n}\Big)$$

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,

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

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 $\|u\| \leqslant R$.

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,

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

[11] We remark that the results in [TJSO14] 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.

[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 $\Phi'[b] - \Phi'[a] = \int_a^b [\![\,|t| \leqslant 2\,]\!]\,\mathrm{d}t$ for all $a, b \in \mathbb{R}$. This property is enough for our purposes.

Here, $\succeq$ denotes the Löwner order.

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

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

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

The main remaining challenge is dealing with the quantification over u. Earlier analyses [TJSO14] 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 $\|u\| \leqslant 1/\max_i \|x_i\|$, we have

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

because $|\langle x_i, u\rangle| \leqslant \|x_i\| \cdot \|u\| \leqslant 1$ 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 $\max_i \|x_i\| \geqslant \sqrt{d}$, 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,

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

(The last 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 [GLW08, GLR10] 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 $|\eta_i| \leqslant 1$). 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 $R \geqslant \tilde{O}_{\kappa_R}(\frac{d}{\alpha n})^{1/2}$, we show that the Huber loss function is locally $\Omega(\alpha \cdot \kappa_R)$-strong convex at β^* within radius $\Omega(R)$. (See Theorem 3.4.) 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 $\kappa_R \geqslant 0.1$ even for constant R.

Final error bound. The aforementioned general framework for analyzing estimators via strong convexity (see Theorem 3.1) allows us to bound the error $\|\hat{\boldsymbol{\beta}} - \beta^*\|$ by the norm of the gradient $\|\nabla f(\beta^*)\|$ 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 $\kappa_R \geqslant 0.1$ (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

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

assuming $R \gtrsim \sqrt{d/\alpha^2 n}$. (This lower bound on R is required by Theorem 3.1 and is stronger than the lower bound on R required by Theorem 3.4.)

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 $\alpha = \min_{i\in[n]} \mathbb{P}\{|\boldsymbol{\eta}_i| \leqslant 1\}$. We further assume the noise entries to satisfy

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

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 $|\boldsymbol{X}_i| \geqslant 1/2$, 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 $0 \leqslant t \leqslant 1$, still satisfy $\mathbb{P}(-t \leqslant \boldsymbol{\eta}'_i \leqslant t) \geqslant \Omega(\alpha \cdot t)$. Surprisingly, this simple observation is enough to obtain an optimal robust algorithm. Standard tail bounds show that with probability $1 - \exp\{-\Omega(\alpha^2 \cdot \varepsilon^2 \cdot n)\}$ the median $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{y}_1/\boldsymbol{X}_1, \ldots, \boldsymbol{y}_n/\boldsymbol{X}_n$ falls in the interval $\left[-\varepsilon + \beta^*, +\varepsilon + \beta^*\right]$ 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, $\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2 \leqslant \varepsilon^2 \leqslant O(1/(\alpha^2 \cdot n))$.

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 [Tuk75]) 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

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

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

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

and hence

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

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

$$\mathop{\mathbb{E}}_{\boldsymbol{\eta}} \boldsymbol{g}(z) = \tfrac{1}{n} \sum_{i\in[n]} \mathop{\mathbb{E}}_{\boldsymbol{\eta}} \operatorname{sign}\left(\boldsymbol{\eta}_i + \langle X_i, z\rangle\right) \cdot \operatorname{sign}(\langle X_i, z\rangle)$$

$$\geqslant \frac{1}{n}\sum_{i\in[n]} \mathbb{P}\big(0 \geqslant \operatorname{sign}(\langle X_i, z\rangle)\cdot \boldsymbol{\eta}_i \geqslant -|\langle X_i, z\rangle|\big)$$
$$\geqslant \frac{1}{n}\sum_{i\in[n]} \Omega(\alpha)\cdot \min\{1, |\langle X_i, z\rangle|\}\,.$$

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

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

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:

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

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

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

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 [Pol91] 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 $\Phi_h : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ 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

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

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 $\hat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_d)^\mathsf{T}$.

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 $(\beta^*_j - \hat{\boldsymbol{\beta}}_j)^2 \leqslant O\left(\frac{1+\|\beta^*\|^2}{\alpha^2} \cdot \log d\right)$. 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 $\hat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_d)^\mathsf{T}$.

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 poly-logarithmic 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 $\hat{\boldsymbol{\beta}}$ satisfying

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

with high probability.

Remark (On learning the norm of β^*). As was noticed in [SBRJ19], 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 $\|\beta^* - \hat{\boldsymbol{\beta}}_{LS}\| \leqslant \Delta = n^{O(1)}$. 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 $\boldsymbol{X}_{[n]\setminus[n/2]}$.

3 Huber-loss estimation guarantees from strong convexity

In this section, we prove statistical guarantees for the Huber loss estimator by establishing strong convexity bounds for the underlying objective function.

The following theorem allows us to show that the global minimizer of the Huber loss is close to the underlying parameter β^*. To be able to apply the theorem, it remains to prove (1) a bound on the gradient of the Huber loss at β^* and (2) a lower bound on the curvature of the Huber loss within a sufficiently large radius around β^*.

Theorem 3.1 (Error bound from strong convexity, adapted from [NRWY09, TJSO14])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex differentiable function and let $\beta^* \in \mathbb{R}^d$. Suppose f is locally κ-strongly convex at β^* within radius $R > 0$:*

$$\forall u \in \mathbb{R}^d, \|u\| \leqslant R. \quad f(\beta^* + u) \geqslant f(\beta^*) + \langle \nabla f(\beta^*), u \rangle + \frac{\kappa}{2}\|u\|^2. \tag{3.1}$$

If $\|\nabla f(\beta^*)\| < \frac{1}{2} \cdot R\kappa$, then every vector $\beta \in \mathbb{R}^d$ such that $f(\beta) \leqslant f(\beta^*)$ satisfies

$$\|\beta - \beta^*\| \leqslant 2 \cdot \|\nabla f(\beta^*)\|/\kappa. \tag{3.2}$$

Furthermore, if If $\|\nabla f(\beta^*)\| < 0.49 \cdot R\kappa$, then every vector $\beta \in \mathbb{R}^d$ such that $f(\beta) \leqslant f(\beta^*) + \varepsilon$, where $\varepsilon = 0.01 \cdot \|\nabla f(\beta^*)\|^2/\kappa$, satisfies

$$\|\beta - \beta^*\| \leqslant 2.01 \cdot \|\nabla f(\beta^*)\|/\kappa. \tag{3.3}$$

Proof. Let $\beta \in \mathbb{R}^d$ be any vector that satisfies $f(\beta) \leqslant f(\beta^*) + \varepsilon$. Write $\beta = \beta^* + t \cdot u$ such that $\|u\| \leqslant R$ and $t = \max\{1, \|\beta - \beta^*\|/R\}$. Since $\beta' = \beta^* + u$ lies on the line segment joining β^* and β, the convexity of f implies that $f(\beta') \leqslant \max\{f(\beta), f(\beta^*)\} \leqslant f(\beta^*) + \varepsilon$. By local strong convexity and Cauchy–Schwarz,

$$\varepsilon \geqslant f(\beta') - f(\beta^*) \geqslant -\|\nabla f(\beta^*)\| \cdot \|u\| + \frac{\kappa}{2} \cdot \|u\|^2 .$$

If $\varepsilon = 0$, we get $\|u\| \leqslant 2 \cdot \|\nabla f(\beta^*)\|/\kappa < R$. By our choice of t, this bound on $\|u\|$ implies $t = 1$ and we get the desired bound.

If $\varepsilon = 0.01 \cdot \|\nabla f(\beta^*)\|^2/\kappa$ and $\|\nabla f(\beta^*)\| < 0.49 \cdot R\kappa$, by the quadratic formula,

$$\|u\| \leqslant \frac{\|\nabla f(\beta^*)\| + \sqrt{\|\nabla f(\beta^*)\|^2 + 2\kappa\varepsilon}}{\kappa} \leqslant \frac{2.01\|\nabla f(\beta^*)\|}{\kappa} < 2.01 \cdot 0.49 \cdot R < R .$$

Again, by our choice of t, this bound on $\|u\|$ implies $t = 1$. We can conclude $\|\beta - \beta^*\| = \|u\| \leqslant 2.01\|\nabla f(\beta^*)\|/\kappa$ as desired. □

We remark that the notion of local strong convexity Eq. (3.1) differs from the usual notion of strong convexity in that one evaluation point for the function is fixed to be β^*. (For the usual notion of strong convexity both evaluation points for the function may vary within some convex region.) However, it is possible to adapt our proof of local strong convexity to establish also (regular) strong convexity inside a ball centered at β^*.

A more general form of the above theorem suitable for the analysis of regularized M-estimators appears in [NRWY09] (see also [Wai19]). Earlier analyses of the Huber-loss estimator also use this theorem implicitly [TJSO14]. (See the the discussion in Section 2.1.)

The following theorem gives an upper bound on the gradient of the Huber loss at β^* for probabilistic error vectors $\boldsymbol{\eta}$.

Theorem 3.2 (Gradient bound for Huber loss)**.** *Let $X \in \mathbb{R}^{n\times d}$ with $X^\mathsf{T}X = \mathrm{Id}$ and $\beta^* \in \mathbb{R}^d$. Let $\boldsymbol{\eta}$ be an n-dimensional random vector with independent symmetrically-distributed entries.*

Then for any $\delta \in (0,1)$, with probability at least $1 - \delta/2$, the Huber loss function $\boldsymbol{f}(\beta) = \frac{1}{n}\sum_{i=1}^n \Phi[(X\beta - \boldsymbol{y})_i]$ for $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$ satisfies

$$\|\nabla \boldsymbol{f}(\beta^*)\| \leqslant 8\sqrt{\frac{d + \ln(2/\delta)}{n}} .$$

Proof. Let $\boldsymbol{z}$ be the n-dimensional random vector with entries $z_i = \Phi'(\boldsymbol{\eta}_i)$, where $\Phi'(t) = \mathrm{sign}(t) \cdot \min\{2, |t|\}$. Then, $\nabla \boldsymbol{f}(\beta^*) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i \cdot x_i$. Since $\boldsymbol{\eta}_i$ is symmetric, $\mathbb{E}\,\boldsymbol{z}_i = 0$. By the Hoeffding bound, every unit vector $u \in \mathbb{R}^d$ satisfies with probability at least $1 - 2\exp(-2(d + \ln(2/\delta)))$.

$$n \cdot \left|\langle \nabla \boldsymbol{f}(\beta^*), u\rangle\right| = |\langle \boldsymbol{z}, Xu\rangle| \leqslant 4\sqrt{d + \ln(2/\delta)} \cdot \|Xu\| = 4\sqrt{(d + \ln(2/\delta))n} .$$

Hence, by union bound over a 1/2-covering of the d-dimensional unit ball of size at most 5^d, we have with probability at least $1 - 2\exp(-2\ln(2/\delta)) \geqslant 1 - \delta/2$,

$$\begin{aligned}\max_{\|u\|\leqslant 1}\left|\langle \nabla \boldsymbol{f}(\beta^*), u\rangle\right| &\leqslant 4\sqrt{(d + \ln(2/\delta))/n} + \max_{\|u\|\leqslant 1/2}\left|\langle \nabla \boldsymbol{f}(\beta^*), u\rangle\right| \\ &= 4\sqrt{(d + \ln(2/\delta))/n} + \tfrac{1}{2}\max_{\|u\|\leqslant 1}\left|\langle \nabla \boldsymbol{f}(\beta^*), u\rangle\right|.\end{aligned}$$

Since $u = \frac{1}{\|\nabla \boldsymbol{f}(\beta^*)\|}\nabla \boldsymbol{f}(\beta^*)$ satisfies $\langle \nabla \boldsymbol{f}(\beta^*), u\rangle = \|\nabla \boldsymbol{f}(\beta^*)\|$, we get the desired bound. □

Proof of local strong convexity. The following lemma represents the second-order behavior of the Huber penalty as an integral. To prove local strong convexity for the Huber-loss function, we will lower bound this integral summed over all sample points.

Lemma 3.3 (Second-order behavior of Huber penalty)**.** *For all $h, \eta \in \mathbb{R}$,*

$$\Phi(\eta + h) - \Phi(\eta) - \Phi'(\eta)\cdot h = h^2 \cdot \int_0^1 (1-t)\cdot \mathbf{1}_{|\eta + t\cdot h|\leqslant 2}\,\mathrm{d}t \geqslant \frac{h^2}{2}\cdot \mathbf{1}_{|\eta|\leqslant 1}\cdot \mathbf{1}_{|h|\leqslant 1}\,. \tag{3.4}$$

Proof. A direct consequence of Taylor's theorem and the integral form of the remainder of the Tayler approximation. Concretely, consider the function $g\colon \mathbb{R}\to\mathbb{R}$ with $g(t) = \Phi(\eta + t\cdot h)$. The first derivative of g at 0 is $g'(0) = \Phi'(\eta)\cdot h$. The function $g''(t) = h^2\cdot \mathbf{1}_{|\eta+th|\leqslant 2}$ is the second derivative of g as an L_1 function (so that $\int_a^b g''(t)\,\mathrm{d}t = g'(b) - g'(a)$ for all $a, b \in \mathbb{R}$). Then, the lemma follows from the following integral form of the remainder of the first-order Taylor expansion of g at 0,

$$g(1) - g(0) - g'(0) = \int_0^1 (1-t)\cdot g''(t)\,\mathrm{d}t\,.$$

Finally, we lower bound the above right-hand side by $\geqslant \frac{1}{2}\,\mathbf{1}_{|\eta|\leqslant 1}\cdot \mathbf{1}_{|h|\leqslant 1}$ using $\int_0^1 (1-t)\mathrm{d}t = \frac{1}{2}$ and the fact $g''(t) \geqslant h^2\cdot \mathbf{1}_{|\eta|\leqslant 1}\cdot \mathbf{1}_{|h|\leqslant 1}$ for all $t\in[0,1]$. □

Theorem 3.4 (Strong convexity of Huber loss)**.** *Let $X \in \mathbb{R}^{n\times d}$ with $X^\mathsf{T}X = \mathrm{Id}$ and $\beta^* \in \mathbb{R}^d$. Let $\boldsymbol{\eta}$ be an n-dimensional random vector with independent entries such that $\alpha = \min_i \mathbb{P}\{|\boldsymbol{\eta}_i| \leqslant 1\}$. Let $\kappa > 0$ and $\delta \in (0,1)$. Suppose that every vector v in the column span of X satisfies*

$$\sum_{i=1}^n \left[\!\!\left[r^2\cdot v_i^2 \leqslant \tfrac{1}{n}\|v\|^2 \right]\!\!\right]\cdot v_i^2 \geqslant \kappa\cdot\|v\|^2\,, \tag{3.5}$$

with

$$\sqrt{\frac{50\cdot\left(d\cdot\ln\left(\frac{100}{\alpha\kappa}\right) + \ln(2/\delta)\right)}{\kappa\alpha n}} \leqslant r \leqslant 1\,.$$

Then, with probability at least $1-\delta/2$, the Huber loss function $\boldsymbol{f}(\beta) = \frac{1}{n}\sum_{i=1}^n \Phi[(X\beta - \boldsymbol{y})_i]$ for $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$ is locally $0.5\kappa\alpha$-strongly convex at β^* within radius $r/2$ (in the sense of Eq. (3.1)).

Proof. By Lemma 3.3, for every $u \in \mathbb{R}^d$,

$$f(\beta^* + u) - f(\beta^*) - \langle \nabla f(\beta^*), u \rangle = \frac{1}{n} \sum_{i=1}^{n} \Phi(\langle x_i, u \rangle - \eta_i) - \Phi(-\eta_i) - \Phi'(-\eta_i) \cdot \langle x_i, u \rangle$$
$$\geqslant \frac{1}{2n} \sum_{i=1}^{n} \langle x_i, u \rangle^2 \cdot \mathbf{1}_{|\langle x_i, u \rangle| \leqslant 1} \cdot \mathbf{1}_{|\eta_i| \leqslant 1} \,. \tag{3.6}$$

It remains show that with high probability over the realization of η, the right-hand side Eq. (3.6) is bounded from below uniformly over all $u \in \mathbb{R}^d$ in a ball.

To this end, we will show, using a covering argument, that with probability at least $1 - \delta/2$ over η, for every unit vector $u \in \mathbb{R}^d$, the vector $v = Xu$ satisfies the following inequality,

$$\frac{1}{n} \sum_{i=1}^{n} [\![v_i^2 \leqslant 4/r^2]\!] \cdot v_i^2 \cdot [\![|\eta_i| \leqslant 1]\!] \geqslant \alpha\kappa/2 \,. \tag{3.7}$$

(Since $\frac{1}{n} X^\top X = \mathrm{Id}$ and $\|u\| = 1$, the vector v has average squared entry 1.)

Let N_ε be an ε-covering of the unit sphere in $\mathbb{R}^d$ of size $|N_\varepsilon| \leqslant (3/\varepsilon)^d$ for a parameter ε to be determined later. Let $u \in \mathbb{R}^d$ be an arbitrary unit vector and let $v = Xu$. Choose $u' \in N_\varepsilon$ such that $\|u - u'\| \leqslant \varepsilon$ and let $v' = Xu'$. We establish the following lower bound on the left-hand side of Eq. (3.7) in terms of a similar expression for v',

$$\frac{1}{n} \sum_{i=1}^{n} [\![(v_i')^2 \leqslant 1/r^2]\!] \cdot (v_i')^2 \cdot [\![|\eta_i| \leqslant 1]\!] \tag{3.8}$$
$$\leqslant \varepsilon^2 + \frac{1}{n} \sum_{i=1}^{n} [\![v_i^2 \leqslant 4/r^2]\!] \cdot [\![(v_i')^2 \leqslant 1/r^2]\!] \cdot (v_i')^2 \cdot [\![|\eta_i| \leqslant 1]\!] \tag{3.9}$$
$$\leqslant 2\varepsilon + \varepsilon^2 + \frac{1}{n} \sum_{i=1}^{n} [\![v_i^2 \leqslant 4/r^2]\!] \cdot [\![(v_i')^2 \leqslant 1/r^2]\!] \cdot v_i^2 \cdot [\![|\eta_i| \leqslant 1]\!] \tag{3.10}$$

The first step Eq. (3.9) uses that each term in the first sum that doesn't appear in the second sum corresponds to a coordinate i with $|v_i'| \leqslant 1/r$ and $|v_i| \geqslant 2/r$, which means that $(v_i - v_i')^2 \geqslant 1/r^2$. Since each term has value at most $1/r^2$, the sum of those terms is bounded by $\|v - v'\|^2 \leqslant \varepsilon^2 n$. For the second step Eq. (3.10), let $(w_i')^2$ be the terms of the second sum and w_i^2 the terms of the third sum. Then, the difference of the two sums is equal to $\langle w - w', w + w' \rangle \leqslant \|w - w'\| \cdot \|w + w'\|$. We have $\|w - w'\| \leqslant \|v - v'\| \leqslant \varepsilon\sqrt{n}$ and $\|w + w'\| \leqslant \|v\| + \|v'\| = 2\sqrt{n}$.

It remains to lower bound the expression Eq. (3.8) over all $u' \in N_\varepsilon$. Let $z_i = \alpha_i - [\![|\eta_i| \leqslant 1]\!]$, where $\alpha_i = \mathbb{P}\{|\eta_i| \leqslant 1\} \geqslant \alpha$. The random variables $z_1, \ldots, z_n$ are independent, centered, and satisfy $|z_i| \leqslant 1$. Let $c_i = [\![(v_i')^2 \leqslant 1/r^2]\!] \cdot (v_i')^2$. By Bernstein inequality, for all $t \geqslant 1$,

$$\mathbb{P}\left\{ \sum_{i=1}^{n} c_i \cdot z_i \geqslant t \cdot \sqrt{\sum_{i \in [n]} \alpha_i c_i^2} + t^2/r^2 \right\} \leqslant e^{-t^2/4} \,.$$

Since $c_i^2 \leqslant c_i/r^2$, we have $\sum_{i\in[n]} \alpha_i c_i^2 \leqslant \frac{1}{r^2}\sum_{i\in[n]} \alpha_i c_i$. Denote $b = \frac{1}{n}\sum_{i\in[n]} \alpha_i c_i$. Note that $b \geqslant \alpha\kappa$.

Choosing $\varepsilon = 0.03\alpha\kappa$, $t = 2\sqrt{d\ln(3/\varepsilon) + \ln(2/\delta)}$, by the union bound over N_ε, it follows that with probability at least $1 - \delta/2$, for every $u' \in N_\varepsilon$, the vector $v' = Xu'$ satisfies,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} [\![(v_i')^2 \leqslant 1/r^2]\!] \cdot (v_i')^2 \cdot [\![|\boldsymbol{\eta}_i| \leqslant 1]\!] &\geqslant b - \sqrt{\frac{t^2 b}{r^2 n}} - \frac{t^2}{r^2 n} \\ &\geqslant b - \sqrt{0.1}\cdot\sqrt{b}\cdot\sqrt{\alpha\kappa} - 0.08\cdot\alpha\kappa \\ &\geqslant 0.6\alpha\kappa\,.\end{aligned}$$

As discussed before, this event implies that for all unit vectors $u \in \mathbb{R}^d$, the vector $v = Xu$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} [\![v_i^2 \leqslant 4/r^2]\!] \cdot v_i^2 \cdot [\![|\boldsymbol{\eta}_i| \leqslant 1]\!] \geqslant 0.6\alpha\kappa - 2\varepsilon - \varepsilon^2 \geqslant 0.5\alpha\kappa\,.$$

□

Putting things together. In this paragraph, we proof Theorem 1.2 by combining previous results in this section.

We start with the definition of well-spread property:

Definition 3.5. Let $V \subseteq \mathbb{R}^n$ be a vector space. V is called (m, ρ)-*spread*, if for every $v \in V$ and every subset $S \subseteq [n]$ with $|S| \geqslant n - m$,

$$\|v_S\| \geqslant \rho\|v\|\,.$$

Theorem 3.6. *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 entries and $\alpha = \min_{i\in[n]} \mathbb{P}\{|\boldsymbol{\eta}_i| \leqslant 1\}$.*

Let $\rho \in (0,1)$ and $\delta \in (0,1)$ and suppose that column span of X is (m,ρ)-spread for

$$m = \frac{10^4 \cdot \big(d\cdot\ln(10/\rho) + \ln(2/\delta)\big)}{\rho^4\cdot\alpha^2}\,.$$

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

$$\frac{1}{n}\big\|X(\beta^* - \hat{\boldsymbol{\beta}})\big\|^2 \leqslant 2000\cdot\frac{d + \ln(2/\delta)}{\rho^4\cdot\alpha^2\cdot n}\,.$$

Proof. Note that if column span of X is (m,ρ)-spread, then Eq. (3.5) holds for all v from column span of X with $r^2 = m/n$ and $\kappa = \rho^2$. Indeed, the set $\big\{i \in [n] \;\big|\; v_i^2 > \frac{1}{r^2 n}\|v\|^2\big\}$ has size at most $r^2 n = m$, so $\sum_{i\in[n]\setminus S} v_i^2 \geqslant \rho^2\|v\|^2 = \kappa\|v\|^2$. Hence for $m = \frac{10^4\left(d\ln(10/\rho)+\ln(2/\delta)\right)}{\rho^4\alpha^2}$,

the conditions of Theorem 3.4 are satisfied and f is locally $0.5\rho^2\alpha$-strongly convex in the ball of radius $0.5\sqrt{m/n}$ with probability at least $1-\delta/2$.

By Theorem 3.2, with probability at least $1-\delta/2$,

$$\|\nabla f(\beta^*)\| \leqslant 8\sqrt{\frac{d+\ln(2/\delta)}{n}}.$$

Hence with probability at least $1-\delta$, $\|\nabla f(\beta^*)\| < 0.49 \cdot \frac{1}{4} \cdot \rho^2\alpha\sqrt{m/n}$. Therefore, by Theorem 3.1, with probability at least $1-\delta$,

$$\tfrac{1}{n}\left\|X(\beta^* - \hat{\beta})\right\|^2 \leqslant 2.1^2 \cdot \frac{4\|\nabla f(\beta^*)\|^2}{\rho^4 \cdot \alpha^2} \leqslant 2000 \cdot \frac{d+\ln(2/\delta)}{\rho^4 \cdot \alpha^2 \cdot n}.$$

□

Proof of Theorem 1.2. Theorem 1.2 follows from Theorem 3.6 with $\rho = \sqrt{1-0.81} = \sqrt{0.19}$ and $\delta = 2^{-d}$. Note that in this case $m \leqslant 10^7 \cdot d/\alpha^2$. □

3.1 Huber-loss estimator for Gaussian design and deterministic noise

In this section we provide a proof of Theorem 1.1. We will use the same strategy as in the previous section: show that the gradient at β^* is bounded by $O\left(\sqrt{d/n}\right)$, then show that Huber loss is locally strongly convex at β^* in a ball of radius $\Omega(1)$, and then use Theorem 3.1 to obtain the desired bound.

Gradient bound.

Theorem 3.7 (Gradient bound, Gaussian design)**.** *Let $X \sim N(0,1)^{n\times d}$ and $\beta^* \in \mathbb{R}^d$. Let $\eta \in \mathbb{R}^n$ be a deterministic vector.*

Then for every $\delta \in (0,1)$, with probability at least $1-\delta/2$, the Huber loss function $f(\beta) = \frac{1}{n}\sum_{i=1}^n \Phi[(X\beta - y)_i]$ for $y = X\beta^* + \eta$ satisfies

$$\|\nabla f(\beta^*)\| \leqslant 3\sqrt{\frac{d+2\ln(2/\delta)}{n}}.$$

Proof. The distribution of $\nabla f(\beta^*)$ is $N\left(0, \frac{1}{n^2}\sum_{i\in[n]}\left(\Phi'(\eta_i)\right)^2 \cdot \mathrm{Id}_d\right)$. Hence by Fact D.10, with probability at least $1-\delta/2$,

$$\|\nabla f(\beta^*)\|^2 \leqslant \frac{4}{n}\left(d + 2\ln(2/\delta) + 2\sqrt{d\ln(2/\delta)}\right) \leqslant \frac{8d + 12\ln(2/\delta)}{n}.$$

□

Strong convexity.

Theorem 3.8 (Strong convexity, Gaussian design). *Let $\boldsymbol{X} \sim N(0,1)^{n\times d}$ and $\beta^* \in \mathbb{R}^d$. Let $\eta \in \mathbb{R}^n$ be a deterministic vector with αn entries of magnitude at most 1. Suppose that for some $\delta \in (0,1)$,*

$$n \geqslant 200 \cdot \frac{d + 2\ln(4/\delta)}{\alpha^2} .$$

Then with probability at least $1 - \delta/2$, the Huber loss function $\boldsymbol{f}(\beta) = \frac{1}{n}\sum_{i=1}^n \Phi[(\boldsymbol{X}\beta - \boldsymbol{y})_i]$ for $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$ is locally 0.5α-strongly convex at β^* within radius $1/6$ (in the sense of Eq. (3.1)).

Proof. By Lemma 3.3, for every $u \in \mathbb{R}^d$,

$$\begin{aligned}\boldsymbol{f}(\beta^* + u) - \boldsymbol{f}(\beta^*) - \langle \nabla \boldsymbol{f}(\beta^*), u\rangle &= \tfrac{1}{n}\sum_{i=1}^n \Phi(\langle \boldsymbol{x}_i, u\rangle - \eta_i) - \Phi(-\eta_i) - \Phi'(-\eta_i)\cdot\langle \boldsymbol{x}_i, u\rangle \\ &\geqslant \tfrac{1}{2n}\sum_{i=1}^n \langle \boldsymbol{x}_i, u\rangle^2 \cdot \mathbf{1}_{|\langle \boldsymbol{x}_i, u\rangle|\leqslant 1}\cdot \mathbf{1}_{|\eta_i|\leqslant 1} .\end{aligned}$$

Consider the set $C = \left\{i \in [n] \,\middle|\, |\eta_i| \leqslant 1\right\}$. Since $|C| = \alpha n$ and η is deterministic, $\boldsymbol{X}_C \sim N(0,1)^{\alpha n\times d}$.

By Fact D.12, for $k = \alpha n/200$, with probability at least $1 - \delta/4$, for any set $\mathcal{K} \subseteq C$ of size k and every $u \in \mathbb{R}^d$,

$$\begin{aligned}\sum_{i\in\mathcal{K}} \langle \boldsymbol{x}_i, u\rangle^2 &\leqslant \|u\|^2 \cdot \left(\sqrt{d} + \sqrt{k} + \sqrt{2k\ln\left(\frac{e\alpha n}{k}\right)} + \sqrt{2\ln(4/\delta)}\right)^2 \\ &\leqslant \|u\|^2 \cdot \left(\sqrt{\alpha n/200} + \sqrt{0.01\alpha n\ln(200e)} + 0.1\cdot\sqrt{\alpha n}\right)^2 \\ &\leqslant 0.18\|u\|^2\cdot\alpha n .\end{aligned}$$

Now if $\mathcal{K}$ is the set of top k entries of $\boldsymbol{X}u$ for $u \in \mathbb{R}^d$ such that $\|u\| \leqslant 1/6$, then we get that the average squared coordinate in $\mathcal{K}$ is at most 1. Hence

$$\sum_{i=1}^n \langle \boldsymbol{x}_i, u\rangle^2 \cdot \mathbf{1}_{|\langle \boldsymbol{x}_i, u\rangle|\leqslant 1}\cdot\mathbf{1}_{|\eta_i|\leqslant 1} \geqslant \sum_{i\in C\setminus\mathcal{K}} \langle \boldsymbol{x}_i, u\rangle^2 \geqslant \|\boldsymbol{X}_C u\|^2 - 0.18\cdot\|u\|^2\alpha n .$$

Since $\boldsymbol{X}_C$ is a Gaussian matrix, for all $u \in \mathbb{R}^d$, with probability at least $1 - \delta/4$,

$$\|\boldsymbol{X}_C u\|^2 \geqslant \|u\|^2\left(\sqrt{\alpha n} - \sqrt{d} - \sqrt{2\log(4/\delta)}\right)^2 \geqslant \|u\|^2\left(\sqrt{\alpha n} - 0.1\sqrt{\alpha n}\right)^2 \geqslant 0.81\|u\|^2\alpha n .$$

Hence with probability at least $1 - \delta/2$, for all $u \in \mathbb{R}^d$ such that $\|u\| \leqslant 1/6$,

$$\boldsymbol{f}(\beta^* + u) - \boldsymbol{f}(\beta^*) - \langle \nabla \boldsymbol{f}(\beta^*), u\rangle \geqslant 0.25\alpha\|u\|^2 ,$$

and $\boldsymbol{f}$ is locally strongly convex with parameter 0.5α at β^* in the ball of radius $1/6$. □

Putting everything together. The following theorem implies Theorem 1.1.

Theorem 3.9. *Let $\eta \in \mathbb{R}^n$ be a deterministic vector. Let $\boldsymbol{X}$ be a random*[13] *n-by-d matrix with iid standard Gaussian entries $\boldsymbol{X}_{ij} \sim N(0,1)$.*

Suppose that for some $\delta \in (0,1)$,

$$n \geqslant 10^4 \cdot \frac{d + 2\ln(2/\delta)}{\alpha^2},$$

where α is the fraction of entries in η of magnitude at most 1.

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

$$\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2 \leqslant 1000 \cdot \frac{d + 2\ln(2/\delta)}{\alpha^2 n}.$$

Proof. Using bounds from Theorem 3.7 and Theorem 3.8, we can apply Theorem 3.1. Indeed, with probability at least $1-\delta$,

$$R = 1/6 > 2 \cdot 3\sqrt{\frac{d + 2\ln(2/\delta)}{0.25\alpha^2 n}} \geqslant 2 \cdot \frac{\|\nabla f(\beta^*)\|}{0.5\alpha},$$

Hence

$$\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2 \leqslant 4 \cdot \frac{\|\nabla f(\beta^*)\|^2}{0.25\alpha^2} \leqslant 1000 \cdot \frac{d + 2\ln(2/\delta)}{\alpha^2 n}.$$

□

4 Robust regression in linear time

In this section we prove Theorem 1.3 and Theorem 1.4. We consider the linear model $\boldsymbol{y} = \boldsymbol{X}\beta^* + \boldsymbol{\eta}$ where $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ has i.i.d entries $\boldsymbol{X}_{ij} \sim N(0,1)$ and the noise vector η satisfies the following assumption.

Assumption 4.1. $\eta \in \mathbb{R}^n$ is a fixed vector such that for some $\alpha = \alpha(n,d) \in (0,1)$, $\left|\left\{i \in [n] \;\middle|\; |\boldsymbol{\eta}_i| \leqslant 1\right\}\right| \geqslant \alpha n$. We denote $\mathcal{T} := \left\{i \in [n] \;\middle|\; |\eta_i| \leqslant 1\right\}$.

We start showing how to obtain a linear time algorithm for Gaussian design in one dimension. Then generalize it to high dimensional settings. We add the following (linear time) preprocessing step

$$\begin{aligned} \forall i \in [n], \qquad & \boldsymbol{y}'_i = \boldsymbol{\sigma}_i \cdot \boldsymbol{y}_i + \boldsymbol{w}_i, \\ & \boldsymbol{X}'_i = \boldsymbol{\sigma}_i \cdot \boldsymbol{X}_i, \qquad \boldsymbol{w}_i \sim N(0,1), \boldsymbol{\sigma}_i \sim U\{-1,1\}, \end{aligned} \tag{PRE}$$

where $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n, \boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_n, \boldsymbol{X}$ are mutually independent. For simplicity, when the context is clear we denote $\boldsymbol{\sigma}_i\eta_i + \boldsymbol{w}_i$ by $\boldsymbol{\eta}_i$ and $\boldsymbol{y}', \boldsymbol{X}'$ with $\boldsymbol{y}, \boldsymbol{X}$. Note that this preprocessing step takes time linear in nd. Assumption 4.1 implies that after this preprocessing step, $\boldsymbol{\eta}$ satisfies the following assumption:

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

Assumption 4.2. For all $i \in \mathcal{T}$ and for any $t \in [0,1]$,

$$\mathbb{P}\big(0 \leqslant \boldsymbol{\eta}_j \leqslant t\big) = \mathbb{P}\big(-t \leqslant \boldsymbol{\eta}_j \leqslant 0\big) \geqslant t/10\,.$$

4.1 Warm up: one-dimensional settings

For the one-dimensional settings, the only property of the design $n \times 1$ matrix $\boldsymbol{X} \sim N(0, \mathrm{Id}_n)$ we are going to use is anti-concentration.

Fact 4.3. *Let $\boldsymbol{X} \sim N(0, \mathrm{Id}_n)$. Then for any $c \in [0,1]$ and $i \in [n]$,*

$$\mathbb{P}(|\boldsymbol{X}_i| \geqslant c) \geqslant \Omega(1)\,.$$

As shown below, our estimator simply computes a median of the samples.

Algorithm 3 Univariate Linear Regression via Median

Input: (y, X), where $y, X \in \mathbb{R}^n$.

0. Preprocess y, X as in Eq. (PRE) and let $(\boldsymbol{y}', \boldsymbol{X}')$ be the resulting pair.

1. Let $\boldsymbol{\mathcal{M}} = \big\{i \in [n] \;\big|\; |\boldsymbol{X}'_i| \geqslant 1/2\big\}$. For $i \in \boldsymbol{\mathcal{M}}$, compute $\boldsymbol{z}_i = \frac{\boldsymbol{y}'_i}{\boldsymbol{X}'_i}$.

2. Return the median $\hat{\boldsymbol{\beta}}$ of $\{\boldsymbol{z}_i\}_{i \in \boldsymbol{\mathcal{M}}}$.

Remark 4.4 (Running time). Preprocessing takes linear time. Finding $\mathcal{M}$ requires linear time, similarly we can compute all $\boldsymbol{z}_i$ in $O(n)$. The median can then be found in linear time using *quickselect* [Hoa61] with pivot chosen running the *median of medians* algorithm [BFP+73]. Thus the overall running time is $O(n)$.

The guarantees of the algorithm are proved in the following theorem.

Theorem 4.5. *Let $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$ for arbitrary $\beta^* \in \mathbb{R}$, $\boldsymbol{X} \sim N(0,1)^{n \times d}$ and $\eta \in \mathbb{R}^n$ satisfying Assumption 4.1 with parameter α. Let $\hat{\boldsymbol{\beta}}$ be the estimator computed by Algorithm 3 given $(\boldsymbol{X}, \boldsymbol{y})$ as input. Then for any positive $\tau \leqslant \alpha^2 \cdot n$,*

$$\big\|\beta^* - \hat{\boldsymbol{\beta}}\big\|^2 \leqslant \frac{\tau}{\alpha^2 \cdot n}$$

with probability at least $1 - 2\exp\{-\Omega(\tau)\}$.

To prove Theorem 4.5 we will use the following bound on the median, which we prove in Appendix D.

Lemma 4.6. *Let $\mathcal{S} \subseteq [n]$ be a set of size γn and let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \mathbb{R}$ be mutually independent random variables satisfying*

1. *For all* $i \in [n]$, $\mathbb{P}(z_i \geqslant 0) = \mathbb{P}(z_i \leqslant 0)$.

2. *For some* $\varepsilon \geqslant 0$, *for all* $i \in \mathcal{S}$, $\mathbb{P}(z_i \in [0, \varepsilon]) = \mathbb{P}(z_i \in [-\varepsilon, 0]) \geqslant q$.

Then with probability at least $1 - 2\exp\{-\Omega(q^2\gamma^2 n)\}$ *the median* $\hat{z}$ *satisfies*

$$|\hat{z}| \leqslant \varepsilon\,.$$

Proof of Theorem 4.5. Due to the preprocessing step the resulting noise $\boldsymbol{\eta}$ satisfies Assumption 4.2. Let $\boldsymbol{\mathcal{M}} \subseteq [n]$ be the set of entries such that $|X_i'| \geqslant \frac{1}{2}$. Since $\mathcal{T}$ and $\boldsymbol{\mathcal{M}}$ are independent, by Chernoff bound, $|\mathcal{T} \cap \boldsymbol{\mathcal{M}}| \geqslant \Omega(\alpha n)$ with probability at least $1 - 2\exp[-\Omega(\alpha n)] \geqslant 1 - 2\exp[-\Omega(\tau)]$. Now observe that for all $\varepsilon \in (0, 1)$ and for all $i \in \mathcal{T} \cap \boldsymbol{\mathcal{M}}$, by Assumption 4.2,

$$\mathbb{P}\left(\left|z_i - \beta^*\right| \leqslant \varepsilon\right) = \mathbb{P}\left(\left|\frac{\boldsymbol{\eta}_i}{X_i'}\right| \leqslant \varepsilon\right) \geqslant \mathbb{P}\left(\left|\boldsymbol{\eta}_i\right| \leqslant \varepsilon/2\right) \geqslant \frac{\varepsilon}{20}\,.$$

By Lemma 4.6, we get the desired bound for $\tau = \varepsilon^2\alpha^2 n$. □

4.2 High-dimensional settings

The median approach can also be applied in higher dimensions. In these settings we need to assume that an upper bound Δ on $\|\beta^*\|$ is known.

Remark 4.7. As shown in [SBRJ19], under the model of Theorem 1.1 the classical least square estimator obtain an estimate $\hat{\boldsymbol{\beta}}$ with error $\frac{d}{n} \cdot \left\|\eta\right\|$. Thus under the additional assumption that the noise magnitude is polynomial in n it is easy to obtain a good enough estimate of the parameter vector.

We first prove in Section 4.2.1 how to obtain an estimate of the form $\left\|\beta^* - \hat{\beta}\right\|^2 \leqslant \frac{\|\beta^*\|^2}{2}$. Then in Section 4.2.2 we obtain Theorem 1.3, using bootstrapping. In Section 4.2.3 we generalize the results to sparse parameter vector β^*, proving Theorem 1.4. Finally we extend the result of Section 4.2.2 to non-spherical Gaussians in Section 4.2.4.

4.2.1 High-dimensional Estimation via median algorithm

To get an estimate of the form $\left\|\beta^* - \hat{\beta}\right\|^2 \leqslant \frac{\|\beta^*\|^2}{2}$, we use the algorithm below:

Algorithm 4 Multivariate Linear Regression Iteration via Median

Input: (y, X) where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$.

0. Preprocess y, X as in Eq. (PRE) and let $(\boldsymbol{y}', \boldsymbol{X}')$ be the resulting pair.

---

1. For all $j \in [d]$ run Algorithm 3 on input $(\boldsymbol{y}, \boldsymbol{X}'_j)$, where $\boldsymbol{X}'_j$ is a j-th column of $\boldsymbol{X}'$ (without additional preprocessing). Let $\hat{\boldsymbol{\beta}}_j$ be the resulting estimate.
2. Return $\hat{\boldsymbol{\beta}} := \left(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_d\right)^\mathsf{T}$.

Remark 4.8 (Running time). Preprocessing takes linear time. Then the algorithm simply executes Algorithm 3 d times, so it runs in $O(nd)$ time.

The performance of the algorithm is captured by the following theorem.

Theorem 4.9. *Let $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$ for arbitrary $\beta^* \in \mathbb{R}^d$, $\boldsymbol{X} \sim N(0,1)^{n \times d}$ and $\eta \in \mathbb{R}^n$ satisfying Assumption 4.1 with parameter α. Let $\hat{\boldsymbol{\beta}}$ be the estimator computed by Algorithm 4 given $(\boldsymbol{y}, \boldsymbol{X})$ as input. Then for any positive $\tau \leqslant \alpha^2 n$,*

$$\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2 \leqslant \frac{d \cdot \tau}{\alpha^2 \cdot n}\left(1 + \left\|\beta^*\right\|^2\right)$$

with probability at least $1 - 2\exp[\ln d - \Omega(\tau)]$.

Proof. We first show that the algorithm obtains a good estimate for each coordinate. Then it suffices to sum the coordinate-wise errors. For $j \in [d]$, let $\boldsymbol{M}_j \subseteq [n]$ be the set of entries such that $|\boldsymbol{X}_{ij}| \geqslant \frac{1}{2}$. Observe that since $\boldsymbol{M}_j$ doesn't depend on $\mathcal{T}$, by Chernoff bound, $\mathcal{T} \cap \boldsymbol{M}_j \geqslant \Omega(\alpha n)$ with probability at least $1 - 2\exp[-\Omega(\alpha n)] \geqslant 1 - 2\exp[-\Omega(\tau)]$. Now for all $i \in [n]$ let

$$\boldsymbol{z}_{ij} := \frac{1}{\boldsymbol{X}'_{ij}}\left(\sigma_i \eta_i + \boldsymbol{w}_i + \sum_{l \neq j} \boldsymbol{X}'_{il} \beta^*_l\right).$$

Note that $\mathbb{P}\left(\boldsymbol{z}_{ij} \geqslant 0\right) = \mathbb{P}\left(\boldsymbol{z}_{ij} \leqslant 0\right)$. Now let $\bar{\beta} \in \mathbb{R}^d$ be the vector such that for $j \in [d] \setminus \{i\}$, $\bar{\beta}_j = \beta^*_j$ and $\bar{\beta}_i = 0$. By properties of Gaussian distribution, for all $i \in [n]$,

$$\boldsymbol{w}_i + \sum_{l \neq j} \boldsymbol{X}'_{il} \beta^*_l \sim N(0, 1 + \|\bar{\beta}\|^2)\,.$$

Hence for each $i \in \mathcal{T} \cap \boldsymbol{M}_j$, for all $0 \leqslant t \leqslant \sqrt{1 + \|\bar{\beta}\|^2}$,

$$\mathbb{P}\left(|\boldsymbol{z}_{ij}| \leqslant t\right) \geqslant \Omega\left(\frac{t}{\sqrt{1 + \|\bar{\beta}\|^2}}\right).$$

By Lemma 4.6, median $\hat{z}_j$ of z_{ij} satisfies

$$\hat{z}_j^2 \leqslant \frac{\tau}{\alpha^2 \cdot n}\left(1 + \left\|\bar{\beta}\right\|^2\right) \leqslant \frac{\tau}{\alpha^2 \cdot n}\left(1 + \left\|\beta^*\right\|^2\right)$$

with probability at least $1 - 2\exp[-\Omega(\tau)]$. Since $\hat{\boldsymbol{\beta}}_j = \hat{z}_j + \beta_j^*$, applying union bound over all coordinates $j \in [d]$, we get the desired bound. □

4.2.2 Nearly optimal estimation via bootstrapping

Here we show how through multiple executions of Algorithm 4 we can indeed obtain error $\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2 \leqslant \tilde{O}\left(\frac{d}{\alpha^2 \cdot n}\right)$. As already discussed, assume that we know some upper bound on $\|\beta\|$, which we denote by Δ. Consider the following procedure:

Algorithm 5 Multivariate Linear Regression via Median

Input: (y, X, Δ) where $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, and $\Delta \geqslant 3$.

1. Randomly partition the samples $y_1, \ldots, y_n$ in $t := \lceil \ln \Delta \rceil$ sets $\boldsymbol{\mathcal{S}}_1, \ldots, \boldsymbol{\mathcal{S}}_t$, such that all $\boldsymbol{\mathcal{S}}_1, \ldots, \boldsymbol{\mathcal{S}}_{t-1}$ have sizes $\Theta\left(\frac{n}{\log \Delta}\right)$ and $\boldsymbol{\mathcal{S}}_t$ has size $\lfloor n/2 \rfloor$.
2. Denote $\hat{\boldsymbol{\beta}}^{(0)} = 0 \in \mathbb{R}^d$. For $i \in [t]$, run Algorithm 4 on input $\left(y_{\boldsymbol{\mathcal{S}}_i} - X_{\boldsymbol{\mathcal{S}}_i}\hat{\boldsymbol{\beta}}^{(i-1)}, X_{\boldsymbol{\mathcal{S}}_i}\right)$, and let $\hat{\boldsymbol{\beta}}^{(i)}$ be the resulting estimator.
3. Return $\hat{\boldsymbol{\beta}} := \sum_{i \in [t]} \hat{\boldsymbol{\beta}}^{(i)}$.

Remark 4.10 (Running time). Splitting the samples into t sets requires time $O(n)$. For each set $\mathcal{S}_i$, the algorithm simply executes Algorithm 4, so all in all the algorithm takes $O\left(\Theta\left(\frac{n}{\log \Delta}\right) d \cdot \log \Delta\right) = O(nd)$ time.

The theorem below proves correctness of the algorithm.

Theorem 4.11. *Let $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$ for $\beta^* \in \mathbb{R}^d$, $\boldsymbol{X} \sim N(0,1)^{n \times d}$ and $\eta \in \mathbb{R}^n$ satisfying Assumption 4.1 with parameter α. Suppose that $\Delta \geqslant 3\left(1 + \|\beta^*\|\right)$, and that for some positive $\varepsilon \leqslant 1/2$, $n \geqslant C \cdot \frac{d \ln \Delta}{\alpha^2} \cdot (\ln(d/\varepsilon) + \ln \ln \Delta)$ for sufficiently large absolute constant $C > 0$. Let $\hat{\boldsymbol{\beta}}$ be the estimator computed by Algorithm 5 given $(\boldsymbol{y}, \boldsymbol{X}, \Delta)$ as input. Then, with probability at least $1 - \varepsilon$,*

$$\frac{1}{n}\left\|\boldsymbol{X}\left(\beta^* - \hat{\boldsymbol{\beta}}\right)\right\|^2 \leqslant O\left(\frac{d \cdot \log(d/\varepsilon)}{\alpha^2 \cdot n}\right).$$

Proof. Since $n \geqslant C \cdot \frac{d \ln \Delta}{\alpha^2} \cdot (\ln d + \ln \ln \Delta)$, by Theorem 4.9, for each $i \in [t-1]$,

$$\left\|\beta^* - \sum_{j=1}^{i} \hat{\boldsymbol{\beta}}^{(j)}\right\|^2 \leqslant \frac{1 + \left\|\beta^* - \sum_{j=1}^{i-1} \hat{\boldsymbol{\beta}}^{(j)}\right\|^2}{10},$$

with probability at least $1 - 2\exp[\ln d - 10\ln(d/\varepsilon) - 10\ln\ln\Delta]$. By union bound over $i \in [t-1]$, with probability at least $1 - 2\varepsilon^{10}$,

$$\left\|\beta^* - \sum_{j=1}^{t-1} \hat{\boldsymbol{\beta}}^{(j)}\right\|^2 \leqslant 100\,.$$

Hence by Theorem 4.9, with probability at least $1 - 4\varepsilon^{10}$,

$$\left\|\beta^* - \sum_{j=1}^{t} \hat{\boldsymbol{\beta}}^{(j)}\right\|^2 \leqslant O\left(\frac{d \cdot \log(d/\varepsilon)}{\alpha^2 \cdot n}\right).$$

By Fact D.8, with probability at least $1 - \varepsilon^{10}$,

$$\frac{1}{n}\left\|\boldsymbol{X}\left(\beta^* - \hat{\boldsymbol{\beta}}\right)\right\|^2 = \left(\beta^* - \hat{\boldsymbol{\beta}}\right)^{\mathsf{T}}\left(\frac{1}{n}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)\left(\beta^* - \hat{\boldsymbol{\beta}}\right) \leqslant 1.1 \cdot \left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2.$$

□

4.2.3 Nearly optimal sparse estimation

A slight modification of Algorithm 4 can be use in the sparse settings.

Algorithm 6 Multivariate Sparse Linear Regression Iteration via Median

Input: (y, X), where $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$.

1. Run Algorithm 4, let $\boldsymbol{\beta}'$ be the resulting estimator.
2. Denote by $\boldsymbol{a}_k$ the value of the k-th largest (by absolute value) coordinate of $\boldsymbol{\beta}'$. For each $j \in [d]$, let $\hat{\boldsymbol{\beta}}_j = \boldsymbol{\beta}'_j$ if $|\boldsymbol{\beta}'_j| \geqslant \boldsymbol{a}_k$, and $\hat{\boldsymbol{\beta}}_j = 0$ otherwise.
3. Return $\hat{\boldsymbol{\beta}} := \left(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_d\right)^{\mathsf{T}}$.

Remark 4.12 (Running time). Running time of Algorithm 4 is $O(nd)$. Similar to median, a_k can be computed in time $O(d)$ (for example, using procedure from [BFP+73]).

The next theorem is the sparse analog of Theorem 4.9.

Theorem 4.13. *Let $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$ for k-sparse $\beta^* \in \mathbb{R}^d$, $\boldsymbol{X} \sim N(0,1)^{n\times d}$ and $\eta \in \mathbb{R}^n$ satisfying Assumption 4.1 with parameter α. Let $\hat{\boldsymbol{\beta}}$ be the estimator computed by Algorithm 6 given $(\boldsymbol{y}, \boldsymbol{X})$ as input. Then for any positive $\tau \leqslant \alpha^2 n$,*

$$\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2 \leqslant O\left(\frac{k \cdot \tau}{\alpha^2 \cdot n}\left(1 + \left\|\beta^*\right\|^2\right)\right)$$

with probability at least $1 - 2\exp[\ln d - \Omega(\tau)]$.

Proof. The reasoning of Theorem 4.9 shows that for each coordinate $[j]$, the median $\hat{z}_j$ satisfies

$$|\hat{z}_j| \leqslant \sqrt{\frac{\tau}{\alpha^2 \cdot n}\left(1 + \left\|\beta^*\right\|^2\right)}$$

with probability at least $1 - 2\exp[-\Omega(\tau)]$. By union bound over $j \in [d]$, with probability at least $1 - 2\exp[\ln d - \Omega(\tau)]$, for any $j \in [d]$,

$$|\boldsymbol{\beta}'_j - \beta^*_j| \leqslant \sqrt{\frac{\tau}{\alpha^2 \cdot n}\left(1 + \left\|\beta^*\right\|^2\right)}.$$

If $\boldsymbol{\beta}'_j < a_k$, then there should be some $i \notin \operatorname{supp}\{\beta^*\}$ such that $|\boldsymbol{\beta}'_i| \geqslant |\boldsymbol{\beta}'_j|$. Hence for such j,

$$|\beta^*_j| \leqslant |\boldsymbol{\beta}'_j - \beta^*_j| + |\boldsymbol{\beta}'_j| \leqslant |\boldsymbol{\beta}'_j - \beta^*_j| + |\boldsymbol{\beta}'_i - \beta^*_i| \leqslant O\left(\sqrt{\frac{\tau}{\alpha^2 \cdot n}\left(1 + \left\|\beta^*\right\|^2\right)}\right).$$

Note that since random variables $\boldsymbol{X}_{ij}$ are independent and absolutely continuous with positive density, $\boldsymbol{\beta}'_j \neq \boldsymbol{\beta}'_m$ for $m \neq j$ with probability 1. Hence $\left|\left\{j \in [d] \;\middle|\; |\boldsymbol{\beta}'_j| \geqslant a_k\right\}\right| = k$. It follows that

$$\begin{aligned}
\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2 &\leqslant \sum_{j \in \operatorname{supp}\{\beta^*\}} \mathbf{1}_{\left[|\boldsymbol{\beta}'_j| < a_k\right]} \cdot \left(\beta^*_j\right)^2 + \sum_{j=1}^{d} \mathbf{1}_{\left[|\boldsymbol{\beta}'_j| \geqslant a_k\right]} \cdot \left(\beta^*_j - \boldsymbol{\beta}'_j\right)^2 \\
&\leqslant \sum_{j \in \operatorname{supp}\{\beta^*\}} O\left(\frac{\tau}{\alpha^2 \cdot n}\left(1 + \left\|\beta^*\right\|^2\right)\right) + \sum_{j=1}^{d} \mathbf{1}_{\left[|\boldsymbol{\beta}'_j| \geqslant a_k\right]} \cdot O\left(\frac{\tau}{\alpha^2 \cdot n}\left(1 + \left\|\beta^*\right\|^2\right)\right) \\
&\leqslant O\left(\frac{k \cdot \tau}{\alpha^2 \cdot n}\left(1 + \left\|\beta^*\right\|^2\right)\right).
\end{aligned}$$

□

Again through bootstrapping we can obtain a nearly optimal estimate. However, for the first few iterations we need a different subroutine. Instead of taking the top-k entries, we will zeros all entries smaller some specific value.

Algorithm 7 Multivariate Sparse Linear Regression Iteration via Median

Input: (y, X, Δ), where $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, $\Delta > 0$.

1. Run Algorithm 4, let $\boldsymbol{\beta}'$ be the resulting estimator.
2. For each $j \in [d]$, let $\hat{\boldsymbol{\beta}}_j = \boldsymbol{\beta}'_j$ if $|\boldsymbol{\beta}'_j| \geqslant \frac{1}{100\sqrt{k}}\Delta$, and $\hat{\boldsymbol{\beta}}_j = 0$ otherwise.
3. Return $\hat{\boldsymbol{\beta}} := \left(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_d\right)^\mathsf{T}$.

Remark 4.14 (Running time). The running time of this algorithm is the same as the running time of Algorithm 4, i.e. $O(nd)$.

The following theorem proves correctness of Algorithm 7.

Theorem 4.15. *Let $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$ for k-sparse $\beta^* \in \mathbb{R}^d$, $\boldsymbol{X} \sim N(0,1)^{n\times d}$ and $\eta \in \mathbb{R}^n$ satisfying Assumption 4.1 with parameter α. Suppose that $\|\beta^*\| \leqslant \Delta$. Let $\hat{\boldsymbol{\beta}}$ be the estimator computed by Algorithm 7 given $(\boldsymbol{y}, \boldsymbol{X}, \Delta)$ as input. Then, with probability at least $1 - 2\exp\left[\ln d - \Omega\left(\frac{\alpha^2\cdot n\cdot\Delta^2}{k(1+\Delta^2)}\right)\right]$, $\operatorname{supp}\{\hat{\boldsymbol{\beta}}\} \subseteq \operatorname{supp}\{\beta^*\}$ and*

$$\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\| \leqslant \frac{\Delta}{10}.$$

Proof. Fix a coordinate $j \in [d]$. The reasoning of Theorem 4.9 shows that the median $\hat{z}_j$ satisfies

$$\hat{z}_j^2 \leqslant \frac{\tau}{\alpha^2\cdot n}\left(1 + \left\|\beta^*\right\|^2\right) \leqslant \frac{\tau}{\alpha^2\cdot n}\left(1+\Delta^2\right)$$

with probability at least $1 - \exp[-\Omega(\tau)] - \exp[-\Omega(\alpha n)]$. If $\beta_i^* = 0$ then with probability at least $1 - 2\exp\left[-\Omega\left(\frac{\alpha^2\cdot n\cdot\Delta^2}{k(1+\Delta^2)}\right)\right]$ we have $\left|\hat{z}_j\right| \leqslant \frac{\Delta}{100\sqrt{k}}$, so $\hat{\boldsymbol{\beta}}_j = 0$. Conversely if $\beta_i^* \neq 0$ then with probability $1 - 2\exp\left[-\Omega\left(\frac{\alpha^2\cdot n\cdot\Delta^2}{k(1+\Delta^2)}\right)\right]$ the error is at most $2\cdot\frac{\Delta}{100\sqrt{k}}$. Combining the two and repeating the argument for all $i \in [d]$, we get that by union bound, with probability at least $1 - 2\exp\left[\ln d - \Omega\left(\frac{\alpha^2\cdot n\cdot\Delta^2}{k(1+\Delta^2)}\right)\right]$, $\left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2 \leqslant k\cdot 4\cdot\frac{\Delta^2}{10000\cdot k} \leqslant \frac{\Delta^2}{100}$. □

Now, combining Algorithm 6 and Algorithm 7 we can introduce the full algorithm.

Algorithm 8 Multivariate Sparse Linear Regression via Median

Input: (y, X, Δ) where $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$, and $\Delta \geqslant 3$.

1. Randomly partition the samples $y_1, \ldots, y_n$ in $t := \lceil \ln\Delta \rceil$ sets $\boldsymbol{\mathcal{S}}_1, \ldots, \boldsymbol{\mathcal{S}}_t$, such that all $\boldsymbol{\mathcal{S}}_1, \ldots, \boldsymbol{\mathcal{S}}_{t-1}$ have sizes $\Theta\left(\frac{n}{\log\Delta}\right)$ and $\boldsymbol{\mathcal{S}}_t$ has size $\lfloor n/2 \rfloor$.

2. Denote $\hat{\boldsymbol{\beta}}^{(0)} = 0 \in \mathbb{R}^d$ and $\Delta_0 = \Delta$. For $i \in [t-1]$, run Algorithm 7 on input

$$\left(y_{\boldsymbol{\mathcal{S}}_i} - X_{\boldsymbol{\mathcal{S}}_i}\hat{\boldsymbol{\beta}}^{(i-1)}, X_{\boldsymbol{\mathcal{S}}_i}, \Delta_{i-1}\right).$$

   Let $\hat{\boldsymbol{\beta}}^{(i)}$ be the resulting estimator and $\Delta_i = \Delta_{i-1}/2$.

3. Run Algorithm 6 on input $\left(y_{\boldsymbol{\mathcal{S}}_t} - X_{\boldsymbol{\mathcal{S}}_t}\hat{\boldsymbol{\beta}}^{(t-1)}, X_{\boldsymbol{\mathcal{S}}_t}\right)$, and let $\hat{\boldsymbol{\beta}}^{(t)}$ be the resulting estimator.

4. Return $\hat{\boldsymbol{\beta}} := \sum_{i\in[t]} \hat{\boldsymbol{\beta}}^{(i)}$.

Remark 4.16 (Running time). Splitting the samples into t sets requires time $O(n)$. For each set $\mathcal{S}_i$, the algorithm simply executes either Algorithm 7 or Algorithm 6, so all in all the algorithm takes $O\left(\Theta\left(\frac{n}{\log \Delta}\right) d \log \Delta + O(nd)\right) = O(nd)$ time.

Finally, Theorem 1.3 follows from the result below.

Theorem 4.17. *Let $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$ for k-sparse $\beta^* \in \mathbb{R}^d$, $\boldsymbol{X} \sim N(0,1)^{n \times d}$ and $\eta \in \mathbb{R}^n$ satisfying Assumption 4.1 with parameter α. Suppose that $\Delta \geqslant 3\left(1 + \|\beta^*\|\right)$, and that for some positive $\varepsilon < 1/2$, $n \geqslant C \cdot \frac{k \ln \Delta}{\alpha^2} \cdot (\ln(d/\varepsilon) + \ln \ln \Delta)$ for sufficiently large absolute constant $C > 0$. Let $\hat{\boldsymbol{\beta}}$ be the estimator computed by Algorithm 8 given $(\boldsymbol{y}, \boldsymbol{X}, \Delta)$ as input. Then, with probability at least $1 - \varepsilon$,*

$$\frac{1}{n}\left\|\boldsymbol{X}\left(\beta^* - \hat{\boldsymbol{\beta}}\right)\right\|^2 \leqslant O\left(\frac{k \cdot \log(d/\varepsilon)}{\alpha^2 \cdot n}\right).$$

Proof. Since $n \geqslant C \cdot \frac{k \ln \Delta}{\alpha^2} \cdot (\ln d + \ln \ln \Delta)$, by Theorem 4.15 and union bound over $i \in [t-1]$, with probability at least $1 - 2\exp[\ln d + \ln t - 10 \ln(d/\varepsilon) - 10 \ln \ln \Delta]$, for each $i \in [t-1]$,

$$\left\|\beta^* - \sum_{j=1}^{i} \hat{\boldsymbol{\beta}}^{(j)}\right\| \leqslant \frac{\Delta}{10^i}.$$

Hence

$$\left\|\beta^* - \sum_{j=1}^{t-1} \hat{\boldsymbol{\beta}}^{(j)}\right\|^2 \leqslant 100$$

with probability $1 - 2\varepsilon^{10}$. Therefore, by Theorem 4.13,

$$\left\|\beta^* - \sum_{j=1}^{t} \hat{\boldsymbol{\beta}}^{(j)}\right\|^2 \leqslant O\left(\frac{k \cdot \log(d/\varepsilon)}{\alpha^2 \cdot n}\right)$$

with probability $1 - 4\varepsilon^{10}$.

Since $\hat{\boldsymbol{\beta}}^{(t)}$ is k-sparse and with probability $1 - 2\varepsilon^{10}$, $\operatorname{supp}\left\{\sum_{j=1}^{t-1} \hat{\boldsymbol{\beta}}^{(j)}\right\} \subseteq \operatorname{supp}\left\{\beta^*\right\}$, vector $\beta^* - \hat{\boldsymbol{\beta}}$ is $2k$-sparse. By Lemma D.9, with probability at least $1 - \varepsilon^{10}$,

$$\frac{1}{n}\left\|\boldsymbol{X}\left(\beta^* - \hat{\boldsymbol{\beta}}\right)\right\|^2 = \left(\beta^* - \hat{\boldsymbol{\beta}}\right)^{\mathsf{T}}\left(\frac{1}{n}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)\left(\beta^* - \hat{\boldsymbol{\beta}}\right) \leqslant 1.1 \cdot \left\|\beta^* - \hat{\boldsymbol{\beta}}\right\|^2.$$

□

4.2.4 Estimation for non-spherical Gaussians

We further extend the results to non-spherical Gaussian design. In this section we assume $n \geqslant d$. We use the algorithm below. We will assume to have in input an estimate of the

covariance matrix of the rows of $\boldsymbol{X}$: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim N(0, \Sigma)$. For example, if number of samples is large enough, sample covariance matrix is a good estimator of Σ. For more details, see Section 4.2.5.

Algorithm 9 Multivariate Linear Regression Iteration via Median for Non-Spherical Design

Input: $(y, X, \hat{\Sigma})$, where $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$, $\hat{\Sigma}$ is a positive definite symmetric matrix.

1. Compute $\tilde{X} = X\hat{\Sigma}^{-1/2}$.
2. Run Algorithm 4 on input $(y, \tilde{X})$ and let $\boldsymbol{\beta}'$ be the resulting estimator.
3. Return $\hat{\boldsymbol{\beta}} := \hat{\Sigma}^{-1/2}\boldsymbol{\beta}'$.

Remark 4.18 (Running time). Since $n \geqslant d$, computing $\tilde{X}$ requires $O(nT(d)/d)$, where $T(d)$ is a time required for multiplication of two $d \times d$ matrices. Algorithm 4 runs in time $O(nd)$, so the running time is $O(nT(d)/d)$.

The performance of the algorithm is captured by the following theorem.

Theorem 4.19. *Let $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$, such that rows of $\boldsymbol{X}$ are iid $\boldsymbol{x}_i \sim N(0, \Sigma)$, η satisfies Assumption 4.1 with parameter α, and $\hat{\Sigma} \in \mathbb{R}^{d\times d}$ is a symmetric matrix independent of $\boldsymbol{X}$ such that $\|\Sigma^{1/2}\hat{\Sigma}^{-1/2} - \mathrm{Id}_d\| \leqslant \delta$ for some $\delta > 0$. Suppose that for some $N \geqslant n + d$, $\delta \leqslant \frac{1}{100\sqrt{\ln N}}$ and $\alpha^2 n \geqslant 100 \ln N$. Let $\hat{\boldsymbol{\beta}}$ be the estimator computed by Algorithm 9 given $(\boldsymbol{y}, \boldsymbol{X}, \hat{\Sigma})$ as input. Then, with probability at least $1 - O(N^{-5})$,*

$$\left\|\hat{\Sigma}^{1/2}\left(\beta^* - \hat{\boldsymbol{\beta}}\right)\right\|^2 \leqslant O\left(\frac{d \cdot \log N}{\alpha^2 \cdot n}\left(1 + \left\|\hat{\Sigma}^{1/2}\beta^*\right\|^2\right) + \delta^2 \cdot \|\hat{\Sigma}^{1/2}\beta^*\|^2\right).$$

Proof. We first show that the algorithm obtains a good estimate for each coordinate. Then it suffices to add together the coordinate-wise errors. By assumptions on $\hat{\Sigma}$,

$$\tilde{\boldsymbol{X}} = \boldsymbol{X}\hat{\Sigma}^{-1/2} = \boldsymbol{G}\Sigma^{1/2}\hat{\Sigma}^{-1/2} = \boldsymbol{G} + \boldsymbol{G}E\,,$$

where $\boldsymbol{G} \sim N(0,1)^{n\times d}$ and E is a matrix such that $\|E\| \leqslant \delta$ for some $\delta \leqslant \frac{1}{100\sqrt{\ln N}}$. Since E is independent of $\boldsymbol{X}$ and each column of E has norm at most δ, for all $i \in [n]$ and $j \in [d]$, $|\tilde{\boldsymbol{X}}_{ij} - \boldsymbol{G}_{ij}| \leqslant 40\delta\sqrt{\ln N}$ with probability $1 - O(N^{-10})$. For simplicity, we still write $\boldsymbol{y}, \tilde{\boldsymbol{X}}$ after the preprocessing step. Fix $j \in [d]$, let $\boldsymbol{M}_j \subseteq [n]$ be the set of entries such that $|\tilde{\boldsymbol{X}}_{ij}| \geqslant 1/2$. With probability $1 - O(N^{-10})$, $\{i \in [n] \mid |\boldsymbol{G}_{ij}| \geqslant 1\}$ is a subset of $\boldsymbol{M}_j$. Hence by Chernoff bound, $\left|\mathcal{T} \cap \boldsymbol{M}_j\right| \geqslant \Omega(\alpha n)$ with probability at least $1 - O(N^{-10})$. Now for all $i \in \boldsymbol{M}_j$ let

$$\boldsymbol{q}_{ij} := \frac{1}{\tilde{\boldsymbol{X}}_{ij}}\left(\boldsymbol{\eta}_i + \sum_{l\neq j}\tilde{\boldsymbol{X}}_{il}\left(\hat{\Sigma}^{1/2}\beta^*\right)_l\right)$$

$$= \frac{1}{\tilde{\boldsymbol{X}}_{ij}}\left(\boldsymbol{\eta}_i + \sum_{l\neq j}\boldsymbol{G}_{il}\left(\hat{\Sigma}^{1/2}\beta^*\right)_l + \sum_{l\neq j}\sum_{m\neq j}\boldsymbol{G}_{im}E_{ml}\left(\hat{\Sigma}^{1/2}\beta^*\right)_l + \sum_{l\neq j}\boldsymbol{G}_{ij}E_{jl}\left(\hat{\Sigma}^{1/2}\beta^*\right)_l\right).$$

Note that for any $i \in \boldsymbol{\mathcal{M}}_j$, with probability $1 - O(N^{-10})$, $\operatorname{sign}(\tilde{\boldsymbol{X}}_{ij}) = \operatorname{sign}(\boldsymbol{G}_{ij})$. Hence

$$\boldsymbol{z}_{ij} := \frac{1}{\tilde{\boldsymbol{X}}_{ij}}\left(\sigma_i\eta_i + \boldsymbol{w}_i + \sum_{l\neq j}\boldsymbol{G}_{il}\left(\hat{\Sigma}^{1/2}\beta^*\right)_l + \sum_{l\neq j}\sum_{m\neq j}\boldsymbol{G}_{im}E_{ml}\left(\hat{\Sigma}^{1/2}\beta^*\right)_l\right)$$

is symmetric about zero.

Now let $\bar{\beta} \in \mathbb{R}^d$ be the vector such that for $l \in [d] \setminus \{j\}$, $\bar{\beta}_l = \left(\hat{\Sigma}^{1/2}\beta^*\right)_l$ and $\bar{\beta}_i = 0$. Note that $\|\bar{\beta}\| \leqslant \|\hat{\Sigma}^{1/2}\beta^*\|$. By properties of Gaussian distribution,

$$\boldsymbol{w}_i + \sum_{l\neq j}\boldsymbol{G}_{il}\left(\hat{\Sigma}^{1/2}\beta^*\right)_l + \sum_{l\neq j}\sum_{m\neq j}\boldsymbol{G}_{im}E_{ml}\left(\hat{\Sigma}^{1/2}\beta^*\right)_l \sim N(0,\sigma^2)\,,$$

where $1+\|\bar{\beta}\|^2 \leqslant \sigma^2 \leqslant 1+(1+\delta^2)\|\bar{\beta}\|^2$. Hence for each $i \in \mathcal{T}\cap\boldsymbol{\mathcal{M}}_j$, for all $0 \leqslant t \leqslant \sqrt{1+\|\bar{\beta}\|^2}$,

$$\mathbb{P}\left(|\boldsymbol{z}_{ij}| \leqslant t\right) \geqslant \Omega\left(\frac{t}{\sqrt{1+\|\bar{\beta}\|^2}}\right).$$

By Lemma 4.6, median $\hat{\boldsymbol{z}}_j$ of $\{\boldsymbol{z}_{ij}\}_{i\in\boldsymbol{\mathcal{M}}_j}$ satisfies

$$\hat{\boldsymbol{z}}_j^2 \leqslant O\left(\frac{\log N}{\alpha^2\cdot n}\left(1+\left\|\bar{\beta}\right\|^2\right)\right) \leqslant O\left(\frac{\log N}{\alpha^2\cdot n}\left(1+\left\|\hat{\Sigma}^{1/2}\beta^*\right\|^2\right)\right)$$

with probability at least $1 - O(N^{-10})$.

For any $i \in \boldsymbol{\mathcal{M}}_j$, the event

$$\left|\frac{\boldsymbol{G}_{ij}}{\tilde{\boldsymbol{X}}_{ij}}\sum_{l\neq j}E_{jl}\left(\hat{\Sigma}^{1/2}\beta^*\right)_l\right| \leqslant O\left(\delta\cdot\|\hat{\Sigma}^{1/2}\beta^*\|\right)$$

occurs with probability $1-O(N^{-10})$. Moreover, since $\|E\| \leqslant \delta$, with probability $1-O(N^{-10})$, for all $i_1,\ldots,i_d \in \boldsymbol{\mathcal{M}}_j$,

$$\sum_{j=1}^{d}\left(\frac{\boldsymbol{G}_{i_jj}}{\tilde{\boldsymbol{X}}_{i_jj}}\sum_{l\neq j}E_{jl}\left(\hat{\Sigma}^{1/2}\beta^*\right)_l\right)^2 \leqslant O\left(\delta^2\cdot\|\hat{\Sigma}^{1/2}\beta^*\|^2\right).$$

Therefore, with probability $1 - O(N^{-9})$, medians $\hat{\boldsymbol{q}}_j$ of $\{\boldsymbol{q}_{ij}\}_{i\in\boldsymbol{\mathcal{M}}_j}$ satisfiy

$$\sum_{j=1}^{d}\hat{\boldsymbol{q}}_j^2 \leqslant O\left(\frac{d\cdot\log n}{\alpha^2\cdot n}\left(1+\left\|\hat{\Sigma}^{1/2}\beta^*\right\|^2\right) + \delta^2\cdot\|\hat{\Sigma}^{1/2}\beta^*\|^2\right).$$

Since $\boldsymbol{y}_i / \tilde{\boldsymbol{X}}_{ij} = \left(\hat{\Sigma}^{1/2}\beta^*\right)_j + \boldsymbol{q}_{ij}$,

$$\sum_{j=1}^{d}\left(\boldsymbol{\beta}'_j - \left(\hat{\Sigma}^{1/2}\beta^*\right)_j\right)^2 \leqslant O\left(\frac{d\cdot\log n}{\alpha^2\cdot n}\left(1+\left\|\hat{\Sigma}^{1/2}\beta^*\right\|^2\right) + \delta^2\cdot\|\hat{\Sigma}^{1/2}\beta^*\|^2\right).$$

□

Next we show how to do bootstrapping for this general case. In this case we will assume to know an upper bound Δ of $\|\boldsymbol{X}\beta^*\|$.

Algorithm 10 Multivariate Linear Regression via Median for Non-Spherical Design

Input: $\left(y, X, \hat{\Sigma}, \Delta\right)$, where $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n, \Delta \geqslant 3$, and $\hat{\Sigma} \in \mathbb{R}^{d\times d}$ is a positive definite symmetric matrix.

1. Randomly partition the samples $y_1, \ldots, y_n$ in $t := t_1 + t_2$, sets $\boldsymbol{\mathcal{S}}_1, \ldots, \boldsymbol{\mathcal{S}}_t$, where $t_1 = \lceil \ln \Delta \rceil$ and $t_2 = \lceil \ln n \rceil$, such that all $\boldsymbol{\mathcal{S}}_1, \ldots, \boldsymbol{\mathcal{S}}_{t_1}$ have sizes $\Theta\left(\frac{n}{\log \Delta}\right)$ and $\boldsymbol{\mathcal{S}}_{t_1+1}, \ldots \boldsymbol{\mathcal{S}}_{t_2}$ have sizes $\Theta\left(\frac{n}{\log n}\right)$.

2. Denote $\hat{\boldsymbol{\beta}}^{(0)} = 0 \in \mathbb{R}^d$ and $\Delta_0 = \Delta$. For $i \in [t]$, run Algorithm 9 on input

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

   and let $\hat{\boldsymbol{\beta}}^{(i)}$ be the resulting estimator.

3. Return $\hat{\boldsymbol{\beta}} := \sum\limits_{i\in[t]} \hat{\boldsymbol{\beta}}^{(t)}$.

Remark 4.20 (Running time). Running time is $O(nT(d)/d)$, where $T(d)$ is a time required for multiplication of two $d \times d$ matrices.

The theorem below extend Theorem 1.3 to non-spherical Gaussians.

Theorem 4.21. *Let $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$ for $\beta^* \in \mathbb{R}^d$, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with iid rows $\boldsymbol{x}_i \sim N(0, \Sigma)$, η satisfying Assumption 4.1 with parameter α. Let $\hat{\Sigma} \in \mathbb{R}^{d\times d}$ be a positive definite symmetric matrix independent of $\boldsymbol{X}$.*

Denote by $\sigma_{\min}$, $\sigma_{\max}$ and κ the smallest singular value, the largest singular value, and the condition number of Σ. Suppose that $\Delta \geqslant 3\left(1 + \|\boldsymbol{X}\beta^*\|\right)$, $n \geqslant C\cdot\left(\frac{d\ln n}{\alpha^2}\cdot\ln(\Delta\cdot n) + (d+\ln n)\kappa^2\ln n\right)$ for some large enough absolute constant C, and $\|\hat{\Sigma} - \Sigma\| \leqslant \frac{\sigma_{\min}}{C\sqrt{\ln n}}$.

Let $\hat{\boldsymbol{\beta}}$ be the estimator computed by Algorithm 10 given $(\boldsymbol{y}, \boldsymbol{X}, \hat{\Sigma}, \Delta)$ as input. Then

$$\frac{1}{n}\left\|\boldsymbol{X}\left(\beta^* - \hat{\boldsymbol{\beta}}\right)\right\|^2 \leqslant O\left(\frac{d\cdot\ln^2 n}{\alpha^2\cdot n}\right)$$

with probability $1 - o(1)$ as $n \to \infty$.

Proof. Let's show that $\|\Sigma^{1/2}\hat{\Sigma}^{-1/2} - \mathrm{Id}_d\| \leqslant \frac{1}{100\sqrt{\ln n}}$. Since $\|\hat{\Sigma} - \Sigma\| \leqslant \frac{\sigma_{\min}}{C\sqrt{\ln n}}$,

$$\left\|\Sigma^{-1}\hat{\Sigma} - \mathrm{Id}_d\right\| \leqslant \frac{1}{C\sqrt{\ln n}}\,.$$

So $\Sigma^{-1}\hat{\Sigma} = \mathrm{Id}_d + E$, where $\|E\| \leqslant \frac{1}{C\sqrt{\ln n}}$. Hence for large enough C,

$$\left\|\Sigma^{1/2}\hat{\Sigma}^{-1/2} - \mathrm{Id}_d\right\| = \left\|(\mathrm{Id}_d + E)^{-1/2} - \mathrm{Id}_d\right\| \leqslant \frac{1}{100\sqrt{\ln n}}\,.$$

Since $n \geqslant C \cdot \frac{d \ln n}{\alpha^2} \cdot (\ln d + \ln \Delta)$ and and $\delta < \frac{1}{C\sqrt{\ln n}}$, applying Theorem 4.19 with $N = 2n$, for each $i \in [t]$,

$$\left\|\hat{\Sigma}^{1/2}\beta^* - \sum_{j=1}^{i} \hat{\Sigma}^{1/2}\hat{\boldsymbol{\beta}}^{(j)}\right\|^2 \leqslant \frac{1}{10}\left(1 + \left\|\hat{\Sigma}^{1/2}\beta^* - \sum_{j=1}^{i-1} \hat{\Sigma}^{1/2}\hat{\boldsymbol{\beta}}^{(j)}\right\|^2\right),$$

with probability at least $1 - O\left(n^{-5}\right)$. By union bound over $i \in [t_1]$, with probability at least $1 - O\left(n^{-4}\right)$,

$$\left\|\hat{\Sigma}^{1/2}\beta^* - \sum_{j=1}^{t_1} \hat{\Sigma}^{1/2}\hat{\boldsymbol{\beta}}^{(j)}\right\|^2 \leqslant 100\,.$$

Hence by Theorem 4.19, by union bound over $i \in [t] \setminus [t_1]$, with probability at least $1 - O\left(n^{-4}\right)$,

$$\left\|\hat{\Sigma}^{1/2}\beta^* - \sum_{j=1}^{t} \hat{\Sigma}^{1/2}\hat{\boldsymbol{\beta}}^{(j)}\right\|^2 \leqslant O\left(\frac{d \cdot \log n}{\alpha^2 \cdot (n/\log n)} + \frac{1}{n}\right) \leqslant O\left(\frac{d \cdot \log^2 n}{\alpha^2 \cdot n}\right).$$

By Fact D.8, with probability at least $1 - O(n^{-4})$, for large enough C,

$$\left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} - \hat{\Sigma}\right\| \leqslant \left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} - \Sigma\right\| + \left\|\Sigma - \hat{\Sigma}\right\| \leqslant \frac{\sigma_{\min}}{100\sqrt{\ln n}} + \frac{\sigma_{\min}}{C\sqrt{\ln n}} \leqslant \frac{\sigma_{\min}}{10} \leqslant \frac{\sigma_{\min}\left(\hat{\Sigma}\right)}{5}\,,$$

where $\sigma_{\min}\left(\hat{\Sigma}\right)$ is the smallest singular value of $\hat{\Sigma}$. Hence

$$\frac{1}{n}\left\|\boldsymbol{X}\left(\beta^* - \hat{\boldsymbol{\beta}}\right)\right\|^2 = \left(\beta^* - \hat{\boldsymbol{\beta}}\right)^\top\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)\left(\beta^* - \hat{\boldsymbol{\beta}}\right) \leqslant 1.2\left\|\hat{\Sigma}^{1/2}\left(\beta^* - \hat{\boldsymbol{\beta}}\right)\right\|^2.$$

□

4.2.5 Estimating covariance matrix

Algorithm 10 requires $\hat{\Sigma} \in \mathbb{R}^{d\times d}$ which is a symmetric matrix independent of $\boldsymbol{X}$ such that $\|\hat{\Sigma} - \Sigma\| \lesssim \frac{\sigma_{\min}}{\sqrt{\ln n}}$. The same argument as in the proof of Theorem 4.21 shows that if $n \geqslant C \cdot (d + \ln n)\kappa^2 \ln n$ for some large enough absolute constant $C > 0$, then with probability at least $1 - O\left(n^{-4}\right)$ the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{\lfloor n/2 \rfloor}$ satisfies the desired property. So we can use Algorithm 5 with design matrix $\boldsymbol{x}_{\lfloor n/2 \rfloor + 1}, \ldots, \boldsymbol{x}_n$ and covariance estimator $\hat{\boldsymbol{\Sigma}}$. Computation of $\hat{\boldsymbol{\Sigma}}$ takes time $O(nd^2)$.

5 Bounding the Huber-loss estimator via first-order conditions

In this section we study the guarantees of the Huber loss estimator via first order optimality condition. Our analysis exploits the connection with high-dimensional median estimation as described in Section 2.2 and yields guarantees comparable to Theorem 1.2 (up to logarithmic factors) for slightly more general noise assumptions.

We consider the linear regression model

$$y = X\beta^* + \eta \tag{5.1}$$

where $X \in \mathbb{R}^{n\times d}$ is a deterministic matrix, $\beta^* \in \mathbb{R}^d$ is a deterministic vector and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a random vector satisfying the assumption below.

Assumption 5.1 (Noise assumption)**.** Let $\mathcal{R} \subseteq [n]$ be a set chosen uniformly at random among all sets of size[14] αn. Then $\eta \in \mathbb{R}^n$ is a random vector such that for all $i \in [n]$, η_i satisfies:

1. $\eta_1, \dots, \eta_n$ are mutually conditionally independent given $\mathcal{R}$.
2. For all $i \in [n]$, $\mathbb{P}\left(\eta_i \leqslant 0 \,\middle|\, \mathcal{R}\right) = \mathbb{P}\left(\eta_i \geqslant 0 \,\middle|\, \mathcal{R}\right)$,
3. For all $i \in \mathcal{R}$, there exists a conditional density p_i of η_i given $\mathcal{R}$ such that $p_i(t) \geqslant 0.1$ for all $t \in [-1, 1]$.

We remark that Assumption 5.1 is more general than the assumptions of Theorem 1.2 (see Appendix A).

We will also require the design matrix X to be well-spread. We restate here the definition.

Definition 5.2. Let $V \subseteq \mathbb{R}^n$ be a vector space. V is called (m, ρ)-*spread*, if for every $v \in V$ and every subset $S \subseteq [n]$ with $|S| \geqslant n - m$,

$$\|v_S\| \geqslant \rho\|v\|\,.$$

Also recall the definition of Huber loss function.

Definition 5.3 (Huber Loss Function)**.** Let $y = X\beta^* + \eta$ for $\beta^* \in \mathbb{R}^d, \eta \in \mathbb{R}^n, X \in \mathbb{R}^{n\times d}$. For $h > 0$ and $\beta \in \mathbb{R}^d$, define

$$f_h(\beta) = \sum_{i=1}^{n} \Phi_h\left(\langle x_i, \beta\rangle - y_i\right) = \sum_{i=1}^{n} \Phi_h\left(\langle x_i, \beta - \beta^*\rangle - \eta_i\right),$$

where

$$\Phi_h(t) = \begin{cases} \frac{1}{2h}t^2 & \text{if } |t| \leqslant h \\ |t| - \frac{h}{2} & \text{otherwise.} \end{cases}$$

[14] For simplicity we assume that αn is integer.

We are now ready to state the main result of the section. We remark that for simplicity we do not optimize constants in the statement below.

Theorem 5.4. *Let $\alpha = \alpha(n,d) \in (0,1)$ and let $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$ for $\beta^* \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times d}$ and $\boldsymbol{\eta} \in \mathbb{R}^n$ satisfying Assumption 5.1. Let*

$$\delta = \frac{10^7 \cdot d \ln n}{\alpha^2 \cdot n},$$

and suppose the column span of X is $(\delta \cdot n, 1/2)$-spread. Then, for $h = 1/n$, the Huber loss estimator $\hat{\boldsymbol{\beta}} := \operatorname*{argmin}_{\beta\in\mathbb{R}^d} f_h(\beta)$ satisfies

$$\frac{1}{n}\|X\left(\hat{\boldsymbol{\beta}} - \beta^*\right)\|^2 \leqslant \delta$$

with probability at least $1 - 10n^{-d/2}$.

Remark 5.5. If for all $i \in [n]$ the conditional distribution of $\boldsymbol{\eta}_i$ given $\boldsymbol{\mathcal{R}}$ is symmetric about 0 (this assumption is satisfied for the noise from considered in Theorem 1.2), the theorem is also true for Huber parameter $h = 1$.

To prove Theorem 5.4 we need the Lemmata below. We start showing a consequence of the $(\delta \cdot n, 1/2)$-spread property of X.

Lemma 5.6. *Let $X \in \mathbb{R}^{n\times d}$, α and δ be as in Theorem 5.4, and let $\boldsymbol{\mathcal{R}}$ be as in Assumption 5.1. With probability $1 - 2n^{-d/2}$, for any $u \in \mathbb{R}^d$,*

$$\sum_{i\in\boldsymbol{\mathcal{R}}} |\langle x_i, u\rangle| \geqslant \frac{1}{2}\alpha \cdot \sqrt{\delta n} \cdot \|Xu\|.$$

Proof. Let $\zeta_1, \ldots, \zeta_n$ be i.i.d. Bernoulli random variables such that $\mathbb{P}(\zeta_i = 1) = 1 - \mathbb{P}(\zeta_i = 0) = \alpha$. By Lemma D.7, it is enough to show that with probability $1 - n^{-d}$, for any $u \in \mathbb{R}^d$,

$$\sum_{i=1}^n \zeta_i |\langle x_i, u\rangle| \geqslant \frac{1}{2}\alpha \cdot \sqrt{\delta n} \cdot \|Xu\|.$$

Note that the inequality is scale invariant, hence it is enough to prove it for all $u \in \mathbb{R}^d$ such that $\|Xu\| = 1$. Consider arbitrary $u \in \mathbb{R}^d$ such that $\|Xu\| = 1$. Applying Lemma E.5 with $\mathcal{A} = \emptyset$, $m = \lfloor \delta n \rfloor$, $\gamma_1 = 0$, $\gamma_2 = 1/2$ and $v = Xu$, we get

$$\sum_{i=1}^n |\langle x_i, u\rangle| \geqslant \frac{3}{4}\sqrt{\delta n}.$$

Hence

$$\mathop{\mathbb{E}}_{\zeta} \sum_{i=1}^n \zeta_i |\langle x_i, u\rangle| \geqslant \frac{3}{4}\alpha \cdot \sqrt{\delta n}.$$

Let $[\![\cdot]\!]$ is the Iverson bracket (0/1 indicator). Applying Lemma D.6 with $g(x,y) = [\![y = 1]\!] \cdot |x|$, $v = Xu$ and $w = \zeta = (\zeta_1, \dots, \zeta_n)^\mathsf{T}$, we get that with probability $1 - n^{-d}$ for all u such that $\|Xu\| = 1$,

$$\left| \sum_{i=1}^{n} \left(\zeta_i |\langle x_i, u \rangle| - \mathop{\mathbb{E}}_{\zeta} \zeta_i |\langle x_i, u \rangle| \right) \right| \leqslant 20\sqrt{d \ln n} \leqslant \frac{1}{5} \alpha \sqrt{\delta n} \,,$$

which yields the desired bound. □

Next we show that with high probability $\|X\left(\hat{\boldsymbol{\beta}} - \beta^*\right)\| < n$.

Lemma 5.7. *Let $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$ as in Theorem 5.4, and let $h \leqslant 1$. With probability $1 - 4n^{-d/2}$, for any β such that $\|X(\beta - \beta^*)\| \geqslant n$,*

$$f_h(\beta) \geqslant f_h(\beta^*) + 1 \,.$$

Proof. Note that

$$f_h(\beta^*) = \sum_{i=1}^{n} \Phi_h(\eta_i) \leqslant \sum_{i=1}^{n} |\eta_i| \,.$$

Consider some β such that $\|X(\beta - \beta^*)\| = n$. Denote $u = \beta - \beta^*$. Since there exists a conditional density $p_i(t) \geqslant 0.1$ (for $t \in [-1, 1]$), there exist $a, b \in [0, 1]$ such that for all $i \in \mathcal{R}$,

$$\mathbb{P}\left(-a \leqslant \eta_i \leqslant 0 \,\middle|\, \mathcal{R}\right) = \mathbb{P}\left(0 \leqslant \eta_i \leqslant b \,\middle|\, \mathcal{R}\right) \geqslant 0.1 \,.$$

Let $\mathcal{S} = \left\{ i \in [n] \,\middle|\, -a \leqslant \eta_i \leqslant b \right\}$. We get

$$\begin{aligned} f_h(\beta) = \sum_{i=1}^{n} \Phi_h\left(\langle x_i, u \rangle - \eta_i\right) &\geqslant \sum_{i=1}^{n} \left|\langle x_i, u \rangle - \eta_i\right| - hn \\ &\geqslant \sum_{i \in \mathcal{S} \cap \mathcal{R}} |\langle x_i, u \rangle| + \sum_{i \in [n] \setminus \mathcal{S}} \left|\langle x_i, u \rangle - \eta_i\right| - 2n. \end{aligned}$$

Denote $\zeta_i = [\![-a \leqslant \eta_i \leqslant b]\!]$. By Lemma 5.6,

$$\mathbb{E}\left[\sum_{i \in \mathcal{R}} \zeta_i \cdot |\langle x_i, u \rangle| \,\middle|\, \mathcal{R} \right] \geqslant \frac{1}{10} \alpha \cdot \sqrt{\delta n} \cdot \|Xu\| \,.$$

with probability $1 - 2n^{-d/2}$.

By Lemma D.6 with $g(x,y) = [\![y = 1]\!] \cdot |x|$, $v = X_{\mathcal{R}} u$, $R = n$ and $w = \zeta_{\mathcal{R}}$, we get that with probability $1 - n^{-d}$ for all u such that $\|X_{\mathcal{R}} u\| \leqslant n$,

$$\sum_{i \in \mathcal{S} \cap \mathcal{R}} |\langle x_i, u \rangle| \geqslant \frac{1}{10} \alpha \cdot \sqrt{\delta n} \cdot \|Xu\| - 20 \cdot \|X_{\mathcal{R}} u\| \cdot \sqrt{d \ln n} - 1 \geqslant \frac{1}{20} \cdot \alpha \cdot \sqrt{\delta n} \cdot \|Xu\| - 1 \,.$$

Note that

$$\sum_{i\in[n]\setminus\mathcal{S}} \big|\langle x_i, u\rangle - \boldsymbol{\eta}_i\big| = \sum_{i\in[n]\setminus\mathcal{S}} \big||\boldsymbol{\eta}_i| - \operatorname{sign}(\boldsymbol{\eta}_i)\langle x_i, u\rangle\big| \geqslant \sum_{i\in[n]\setminus\mathcal{S}} |\boldsymbol{\eta}_i| - \sum_{i\in[n]\setminus\mathcal{S}} \operatorname{sign}(\boldsymbol{\eta}_i)\langle x_i, u\rangle\,.$$

Applying Lemma D.6 with $g(x,y) = [\![-a \leqslant y \leqslant b]\!]\operatorname{sign}(y)\cdot|x|$, $v = Xu$, $\boldsymbol{w} = \boldsymbol{\eta}$, $R = n$, we get that with probability $1 - n^{-d}$, for all $u \in \mathbb{R}^d$ such that $\|Xu\| = n$,

$$\begin{aligned}\left|\sum_{i\in[n]\setminus\mathcal{S}} \operatorname{sign}(\boldsymbol{\eta}_i)\langle x_i, u\rangle\right| &= \left|\sum_{i=1}^n \big([\![-a \leqslant \boldsymbol{\eta}_i \leqslant b]\!]\operatorname{sign}(\boldsymbol{\eta}_i)\langle x_i, u\rangle - \mathbb{E}\big[[\![-a \leqslant \boldsymbol{\eta}_i \leqslant b]\!]\operatorname{sign}(\boldsymbol{\eta}_i)\langle x_i, u\rangle \,\big|\, \boldsymbol{\mathcal{R}}\big]\big)\right| \\ &\leqslant 20n\sqrt{d\ln n}\|Xu\| + 1\,.\end{aligned}$$

Therefore, with probability $1 - 4n^{-d/2}$, for any $\beta \in \mathbb{R}^d$ such that $\|X(\beta - \beta^*)\| = n$,

$$\boldsymbol{f}_h(\beta) \geqslant \frac{1}{20}\alpha\cdot\sqrt{\delta n}\cdot\|Xu\| + \sum_{i=1}^n|\boldsymbol{\eta}_i| - \sum_{i\in\boldsymbol{\mathcal{S}}}|\boldsymbol{\eta}_i| - 2n - 20n\sqrt{d\ln n}\|Xu\| - 2 \geqslant \sum_{i=1}^n|\boldsymbol{\eta}_i| + 1 \geqslant \boldsymbol{f}_h(\beta^*) + 1\,.$$

Note that since $\boldsymbol{f}$ is convex, with probability $1 - 4n^{-d/2}$, for any β such that $\|X(\beta - \beta^*)\| > n$, $\boldsymbol{f}_h(\beta) \geqslant \boldsymbol{f}_h(\beta^*) + 1$. □

Now observe that $\boldsymbol{f}_h(\beta)$ is differentiable and

$$\nabla\boldsymbol{f}_h(\beta) = \sum_{i=1}^n \phi_h\big(\langle x_i, \beta\rangle - \boldsymbol{y}_i\big)\cdot x_i = \sum_{i=1}^n \phi_h\big(\langle x_i, \beta - \beta^*\rangle - \boldsymbol{\eta}_i\big)\cdot x_i\,,$$

where $\phi_h(t) = \Phi_h'(t) = \operatorname{sign}(t)\cdot\min\{|t|/h, 1\}$, $t \in \mathbb{R}$. We will need the following lemma.

Lemma 5.8. *Let $\boldsymbol{z}$ be a random variable such that $\mathbb{P}(\boldsymbol{z} \leqslant 0) = \mathbb{P}(\boldsymbol{z} \geqslant 0)$. Then for any τ such that $|\tau| \geqslant 2h$,*

$$\tau\cdot\mathop{\mathbb{E}}_{\boldsymbol{z}}\phi_h(\tau - \boldsymbol{z}) \geqslant |\tau|\cdot\mathbb{P}\big(0 \leqslant \operatorname{sign}(\tau)\cdot\boldsymbol{z} \leqslant |\tau|/2\big)\,.$$

Proof. Note that

$$\tau\cdot\mathop{\mathbb{E}}_{\boldsymbol{z}}\phi_h(\tau - \boldsymbol{z}) = |\tau|\cdot\mathop{\mathbb{E}}_{\boldsymbol{z}}\phi_h\big(|\tau| - \operatorname{sign}(\tau)\cdot\boldsymbol{z}\big)\,.$$

We get

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{z}}\phi_h\big(|\tau| - \operatorname{sign}(\tau)\boldsymbol{z}\big) &= \mathbb{P}\big(\operatorname{sign}(\tau)\boldsymbol{z} \leqslant |\tau| - h\big) + \mathop{\mathbb{E}}_{\boldsymbol{z}}[\![\operatorname{sign}(\tau)\boldsymbol{z} > |\tau| - h]\!]\cdot\phi_h\big(|\tau| - \operatorname{sign}(\tau)\boldsymbol{z}\big) \\ &\geqslant \mathbb{P}\big(0 \leqslant \operatorname{sign}(\tau)\boldsymbol{z} \leqslant |\tau| - h\big) + \mathbb{P}\big(\operatorname{sign}(\tau)\boldsymbol{z} < 0\big) - \mathbb{P}\big(\operatorname{sign}(\tau)\boldsymbol{z} > 0\big) \\ &\geqslant \mathbb{P}\big(0 \leqslant \operatorname{sign}(\tau)\boldsymbol{z} \leqslant |\tau|/2\big)\,.\end{aligned}$$

□

Using point 3 of Assumption 5.1, we get for all $i \in [n]$ and for all $\tau \geqslant 2h$,

$$\tau \cdot \mathbb{E}\left[\phi_h(\tau - \boldsymbol{\eta}_i) \,\middle|\, \boldsymbol{\mathcal{R}}\right] \geqslant \frac{1}{20}|\tau| \cdot \min\{|\tau|, 1\}\,.$$

Note that for $h = 1$, if z is symmetric, we can also show it for $\tau \leqslant 2h \leqslant 2$. Indeed,

$$\begin{aligned}
|\tau| \mathop{\mathbb{E}}_{z} \phi_h(|\tau| - \operatorname{sign}(\tau) z) &= |\tau| \mathop{\mathbb{E}}_{z}[\![|z| \leqslant h]\!]\phi_h(|\tau| - \operatorname{sign}(\tau) z) + |\tau| \mathop{\mathbb{E}}_{z}[\![|z| > h]\!]\phi_h(|\tau| - \operatorname{sign}(\tau) z) \\
&\geqslant |\tau| \mathop{\mathbb{E}}_{z}[\![|z| \leqslant h]\!]\phi_h(|\tau| - \operatorname{sign}(\tau) z) \\
&= |\tau| \mathop{\mathbb{E}}_{z}[\![|z| \leqslant h]\!]\big(\phi_h(|\tau| - z) - \phi_h(-z)\big)\,,
\end{aligned}$$

since for symmetric z, $\mathbb{E}_z[\![|z| > h]\!]\phi_h(|\tau| - \operatorname{sign}(\tau) z) \geqslant 0$. Assuming the existence of density p of z such that $p(t) \geqslant 0.1$ for all $t \in [-1, 1]$, we get

$$\begin{aligned}
\mathop{\mathbb{E}}_{z}[\![|z| \leqslant h]\!]\big(\phi_h(|\tau| - z) - \phi_h(-z)\big) &\geqslant 0.1 \int_{-1}^{1} \big(\phi_h(|\tau| - z) - \phi_h(z)\big) dz \\
&\geqslant \frac{1}{20} \min\{|\tau|, 1\}\,.
\end{aligned}$$

We are now ready to prove Theorem 5.4.

Proof of Theorem 5.4. Consider $\boldsymbol{u} = \hat{\boldsymbol{\beta}} - \beta^*$. By Lemma 5.7, with probability $1 - 3n^{-d/2}$, $\|X\boldsymbol{u}\| \leqslant n$. If $\|X\boldsymbol{u}\| < 100$, we get the desired bound. So further we assume that $100 \leqslant \|X\boldsymbol{u}\| \leqslant n$.

Since $\nabla \boldsymbol{f}_h(\hat{\beta}) = 0$,

$$\sum_{i=1}^{n} \phi_h\big(\langle x_i, \boldsymbol{u}\rangle - \boldsymbol{\eta}_i\big) \cdot \langle x_i, \boldsymbol{u}\rangle = 0\,.$$

For each $i \in [n]$, consider the function F_i defined as follows:

$$F_i(a) = \langle x_i, a\rangle\, \mathbb{E}\left[\phi_h\big(\langle x_i, a\rangle - \boldsymbol{\eta}_i\big) \,\middle|\, \boldsymbol{\mathcal{R}}\right]$$

for any $a \in \mathbb{R}^d$. Applying Lemma 5.8 with $z = \boldsymbol{\eta}_i$, $\tau = \langle x_i, a\rangle$ and $h = 1/n$, and using point 3 of Assumption 5.1, we get for any $a \in \mathbb{R}^d$,

$$\begin{aligned}
\sum_{i=1}^{n} F_i(a) &\geqslant -\sum_{i=1}^{n}[\![|\langle x_i, a\rangle| < 2h]\!]|\langle x_i, a\rangle| + \sum_{i=1}^{n}[\![|\langle x_i, a\rangle| \geqslant 2h]\!]F_i(a) && (5.2) \\
&\geqslant -2 + \frac{1}{20}\sum_{i \in \boldsymbol{\mathcal{R}}}|\langle x_i, a\rangle| \cdot \min\{|\langle x_i, a\rangle|, 1\}\,. && (5.3)
\end{aligned}$$

Note that this is the only place in the proof where we use $h \leqslant 1/n$. By the observation described after Lemma 5.8, if for all $i \in [n]$, conditional distribution of $\boldsymbol{\eta}_i$ given $\boldsymbol{\mathcal{R}}$ is symmetric about 0, the proof also works for $h = 1$.

For $x, y \in \mathbb{R}$, consider $g(x,y) = x \cdot \phi_h(x-y)$. For any $\Delta x \in \mathbb{R}$,

$$|g(x+\Delta x, y) - g(x,y)| = \left|(x+\Delta x)\cdot\phi_h(x+\Delta x - y) - x\cdot\phi_h(x-y)\right| \leqslant \Delta x + \frac{x}{h}\Delta x\,.$$

By Lemma D.6 with $v = Xa$, $\boldsymbol{w} = \boldsymbol{\eta}$, $R = n$, and $K = \left(1+n^2\right)$, with probability $1 - 4n^{-d/2}$, for all $a \in \mathbb{R}^d$ such that $\|Xa\| \leqslant n$,

$$\left|\sum_{i=1}^{n}\left(\langle x_i, a\rangle \cdot \phi_h\left(\langle x_i, a\rangle - \boldsymbol{\eta}_i\right) - F_i(a)\right)\right| \leqslant 25\sqrt{d\ln n}\cdot\|Xa\| + 1/n\,. \tag{5.4}$$

Let $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n$ be i.i.d. Bernoulli random variables such that $\mathbb{P}(\boldsymbol{\zeta}_i = 1) = 1 - \mathbb{P}(\boldsymbol{\zeta}_i = 0) = \alpha$. By Lemma D.7 and Lemma D.6 with $g(x,y) = [\![y=1]\!]|x|\cdot\min\{|x|,1\}$, $v = Xa$, $\boldsymbol{w}_i = \boldsymbol{\zeta}_i$, $R = n$ and $K = 2$, with probability $1 - 3n^{-d/2}$, for all $a \in \mathbb{R}^d$ such that $\|Xa\| \leqslant n$,

$$\sum_{i\in\boldsymbol{\mathcal{R}}}|\langle x_i, a\rangle|\cdot\min\{|\langle x_i, a\rangle|, 1\} \geqslant \alpha\sum_{i=1}^{n}|\langle x_i, a\rangle|\cdot\min\{|\langle x_i, a\rangle|, 1\} - 20\sqrt{d\ln n}\,\|Xa\| - 1/n\,. \tag{5.5}$$

Plugging $a = \boldsymbol{u}$ into inequalities 5.3, 5.4 and 5.5, we get

$$\sum_{|\langle x_i,\boldsymbol{u}\rangle|\leqslant 1}\langle x_i, \boldsymbol{u}\rangle^2 + \sum_{|\langle x_i,\boldsymbol{u}\rangle|>1}|\langle x_i, \boldsymbol{u}\rangle| \leqslant \frac{1000}{\alpha}\sqrt{d\ln n}\cdot\|X\boldsymbol{u}\|$$

with probability $1 - 7n^{-d/2}$.

If

$$\sum_{|\langle x_i,\boldsymbol{u}\rangle|\leqslant 1}\langle x_i, \boldsymbol{u}\rangle^2 < \frac{1}{3}\|X\boldsymbol{u}\|^2\,,$$

we get

$$\sum_{|\langle x_i,\boldsymbol{u}\rangle|>1}|\langle x_i, \boldsymbol{u}\rangle| \leqslant \frac{1000}{\alpha}\sqrt{d\ln n}\cdot\|X\boldsymbol{u}\|\,.$$

Applying Lemma E.5 with $m = \lfloor\frac{10^7 d\ln n}{\alpha^2}\rfloor$, $\gamma_1 = 1/\sqrt{3}$, $\mathcal{A} = \{i\in[n] : |\langle x_i, \boldsymbol{u}\rangle| \leqslant 1\}$, $\gamma_2 = 1/2$ and $v = X\boldsymbol{u}$, we get a contradiction.

Hence

$$\sum_{|\langle x_i,\boldsymbol{u}\rangle|\leqslant 1}\langle x_i, \boldsymbol{u}\rangle^2 \geqslant \frac{1}{3}\|X\boldsymbol{u}\|^2$$

and we get

$$\|X\left(\hat{\boldsymbol{\beta}} - \beta^*\right)\| \leqslant \frac{3000}{\alpha}\sqrt{d\ln n}\,,$$

with probability at least $1 - 10n^{-d/2}$, which yields the desired bound. □

References

[BFP+73] Manuel Blum, Robert W. Floyd, Vaughan R. Pratt, Ronald L. Rivest, and Robert Endre Tarjan, *Time bounds for selection*, J. Comput. Syst. Sci. **7** (1973), no. 4, 448–461. 24, 28

[BJKK17a] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar, *Consistent robust regression*, Advances in Neural Information Processing Systems 30 (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), Curran Associates, Inc., 2017, pp. 2110–2119. 4, 5, 6

[BJKK17b] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar, *Consistent robust regression*, NIPS, 2017, pp. 2107–2116. 5

[CRT05] Emmanuel Candes, Justin Romberg, and Terence Tao, *Stable signal recovery from incomplete and inaccurate measurements*, 2005. 3, 5, 7

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant, *Learning from untrusted data*, Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, 2017, pp. 47–60. 3

[CT05] Emmanuel Candes and Terence Tao, *Decoding by linear programming*, 2005. 3, 5, 7

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart, *Robust estimators in high-dimensions without the computational intractability*, SIAM Journal on Computing **48** (2019), no. 2, 742–864. 3

[DKS19] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart, *Efficient algorithms and lower bounds for robust linear regression*, Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019 (Timothy M. Chan, ed.), SIAM, 2019, pp. 2745–2754. 5

[Don06] David L Donoho, *Compressed sensing*, IEEE Transactions on information theory **52** (2006), no. 4, 1289–1306. 7

[DT19] Arnak Dalalyan and Philip Thompson, *Outlier-robust estimation of a sparse linear model using l_1-penalized huber's m-estimator*, Advances in Neural Information Processing Systems, 2019, pp. 13188–13198. 3, 5

[EvdG+18] Andreas Elsener, Sara van de Geer, et al., *Robust low-rank matrix estimation*, The Annals of Statistics **46** (2018), no. 6B, 3481–3509. 7

[GLR10] Venkatesan Guruswami, James R. Lee, and Alexander A. Razborov, *Almost euclidean subspaces of l_1^n VIA expander codes*, Combinatorica **30** (2010), no. 1, 47–68. 12, 61

[GLW08] Venkatesan Guruswami, James R. Lee, and Avi Wigderson, *Euclidean sections of with sublinear randomness and error-correction over the reals*, APPROX-RANDOM, Lecture Notes in Computer Science, vol. 5171, Springer, 2008, pp. 444–454. 6, 12

[HBRN08] Jarvis Haupt, Waheed U Bajwa, Michael Rabbat, and Robert Nowak, *Compressed sensing for networked data*, IEEE Signal Processing Magazine **25** (2008), no. 2, 92–101. 3

[Hoa61] C. A. R. Hoare, *Algorithm 65: Find*, Commun. ACM **4** (1961), no. 7, 321âĂŞ322. 24

[Hub64] Peter J. Huber, *Robust estimation of a location parameter*, Ann. Math. Statist. **35** (1964), no. 1, 73–101. 3

[Jen69] Robert I. Jennrich, *Asymptotic properties of non-linear least squares estimators*, Ann. Math. Statist. **40** (1969), no. 2, 633–643. 55

[KKK19] Sushrut Karmalkar, Adam R. Klivans, and Pravesh Kothari, *List-decodable linear regression*, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada, 2019, pp. 7423–7432. 3

[KKM18] Adam R. Klivans, Pravesh K. Kothari, and Raghu Meka, *Efficient algorithms for outlier-robust regression*, Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018, 2018, pp. 1420–1430. 3

[KP18] Sushrut Karmalkar and Eric Price, *Compressed sensing with adversarial sparse noise via l1 regression*, arXiv preprint arXiv:1809.08055 (2018). 3, 5, 7

[KT07] Boris S Kashin and Vladimir N Temlyakov, *A remark on compressed sensing*, Mathematical notes **82** (2007), no. 5, 748–755. 7

[LLC19] Liu Liu, Tianyang Li, and Constantine Caramanis, *High dimensional robust m-estimation: Arbitrary corruption and heavy tails*, 2019. 3

[LM00] B. Laurent and P. Massart, *Adaptive estimation of a quadratic functional by model selection*, Ann. Statist. **28** (2000), no. 5, 1302–1338. 59

[LSLC18] Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis, *High dimensional robust sparse regression*, arXiv preprint arXiv:1805.11643 (2018). 3

[NRWY09] Sahand Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu, *A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers*, NIPS, Curran Associates, Inc., 2009, pp. 1348–1356. 9, 16, 17

[NT13] Nam H Nguyen and Trac D Tran, *Exact recoverability from dense corrupted observations via l_1-minimization*, IEEE transactions on information theory **59** (2013), no. 4. 7

[Pol91] David Pollard, *Asymptotics for least absolute deviation regression estimators*, Econometric Theory **7** (1991), no. 2, 186–199. 4, 7, 14

[RH15] Phillippe Rigollet and Jan-Christian Hütter, *High dimensional statistics*, Lecture notes for course 18S997 **813** (2015), 814. 46

[RL05] Peter J Rousseeuw and Annick M Leroy, *Robust regression and outlier detection*, vol. 589, John wiley & sons, 2005. 3

[RY20] Prasad Raghavendra and Morris Yau, *List decodable learning via sum of squares*, Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020, 2020, pp. 161–180. 3

[SBRJ19] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain, *Adaptive hard thresholding for near-optimal consistent robust regression*, Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA, 2019, pp. 2892–2897. 4, 5, 6, 7, 8, 16, 25, 50

[SZF19] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan, *Adaptive huber regression*, Journal of the American Statistical Association (2019), 1–24. 3

[TJSO14] Efthymios Tsakonas, Joakim Jaldén, Nicholas D Sidiropoulos, and Björn Ottersten, *Convergence of the huber regression m-estimate in the presence of dense outliers*, IEEE Signal Processing Letters **21** (2014), no. 10, 1211–1214. 1, 4, 5, 6, 7, 9, 10, 11, 16, 17, 50

[TSW18] Kean Ming Tan, Qiang Sun, and Daniela Witten, *Robust sparse reduced rank regression in high dimensions*, arXiv preprint arXiv:1810.07913 (2018). 7

[Tuk75] John W Tukey, *Mathematics and the picturing of data*, Proceedings of the International Congress of Mathematicians, Vancouver, 1975, vol. 2, 1975, pp. 523–531. 13

[Ver18] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2018. 56

[Vis18] Nisheeth K Vishnoi, *Algorithms for convex optimization*, Cambridge University Press, 2018. 52

[Wai19] Martin J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019. 17, 56, 59

[WYG+08] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma, *Robust face recognition via sparse representation*, IEEE transactions on pattern analysis and machine intelligence **31** (2008), no. 2, 210–227. 3

A Error convergence and model assumptions

In this section we discuss the error convergence of our main theorems Theorem 1.1, Theorem 1.2 as well as motivate our model assumptions.

A.1 Lower bounds for consistent oblivious linear regression

We show here that no estimator can obtain expected squared error $o\big(d/(\alpha^2 \cdot n)\big)$ for any $\alpha \in (0,1)$ and that no estimator can have expected error converging to zero for $\alpha \lesssim \sqrt{d/n}$. The first claim is captured by the following statement.

Fact A.1. *Let $X \in \mathbb{R}^{n\times d}$ be a matrix with linearly independent columns. Let $\boldsymbol{\eta} \sim N(0, \sigma^2 \cdot \mathrm{Id}_n)$ with $\sigma > 0$ so that $\alpha = \min_i \mathbb{P}\{|\boldsymbol{\eta}_i| \leqslant 1\} = \Theta(1/\sigma)$.*

Then there exists a distribution over $\boldsymbol{\beta}^*$ independent of $\boldsymbol{\eta}$ such that for every estimator $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$, with probability at least $\Omega(1)$,

$$\frac{1}{n}\left\|X\hat{\beta}(X\boldsymbol{\beta}^* + \boldsymbol{\eta}) - X\beta^*\right\|^2 \geqslant \Omega\left(\frac{d}{\alpha^2 \cdot n}\right).$$

In particular, for every estimator $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$ there exists a vector β^* such that for $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$, with probability at least $\Omega(1)$,

$$\frac{1}{n}\left\|X\hat{\beta}(\boldsymbol{y}) - X\beta^*\right\|^2 \geqslant \Omega\left(\frac{d}{\alpha^2 \cdot n}\right).$$

Fact A.1 is well-known and we omit the proof here (see for example [RH15]). The catch is that the vector $\boldsymbol{\eta} \sim N(0, \sigma^2 \cdot \mathrm{Id})$ satisfies the noise constraints of Theorem 1.2 for $\alpha = \Theta(1/\sigma)$. Hence, for $\alpha \lesssim \sqrt{d/n}$ we obtain the second claim as an immediate corollary.

Corollary A.2. *Let $n, d \in \mathbb{R}$ and $\alpha \lesssim \sqrt{\frac{d}{n}}$. Let $X \in \mathbb{R}^{n\times d}$ be a matrix with linearly independent columns and let $\boldsymbol{\eta} \sim N(0, 1/\alpha^2 \cdot \mathrm{Id}_n)$. Then, for every estimator $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$ there exists a vector β^* such that for $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$, with probability at least $\Omega(1)$,*

$$\frac{1}{n}\left\|X\hat{\beta}(\boldsymbol{y}) - X\beta^*\right\|^2 \geqslant \Omega(1).$$

In other words, in this regime no estimator obtains error converging to zero.

A.2 On the design assumptions

Recall our linear model with Gaussian design:

$$\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta \tag{A.1}$$

for $\beta^* \in \mathbb{R}^d$, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with i.i.d entries $\boldsymbol{X}_{ij} \sim N(0,1)$ and $\eta \in \mathbb{R}^n$ a deterministic vector with $\alpha \cdot n$ entries bounded by 1 in absolute value. We have mentioned in Section 1.1 how to extend these Gaussian design settings to deterministic design settings. We formally show here how to apply Theorem 1.2 to reason about the Gaussian design model Eq. (A.1). For this it suffices to to turn an instance of model Eq. (A.1) into an instance of the model:

$$\boldsymbol{y}' = X'\beta^* + \boldsymbol{\eta}' \tag{A.2}$$

where $\beta^* \in \mathbb{R}^d$, X' is a n-by-d matrix $(\Omega(n), \Omega(1))$-spread and the noise vector $\boldsymbol{\eta}$ has independent, symmetrically distributed entries with $\alpha = \min_{i\in[n]} \mathbb{P}\{|\boldsymbol{\eta}_i| \leqslant 1\}$. This can be done resampling through the following procedure.

Algorithm 11 Resampling

Input: (y, X) where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times d}$.
Sample n indices $\gamma_1, \ldots, \gamma_n$ independently from the uniform distribution over $[n]$.
Sample n i.i.d Rademacher random variables $\sigma_1, \ldots, \sigma_n$.
Return pair $(\boldsymbol{y}', \boldsymbol{X}')$ where $\boldsymbol{y}'$ is an n-dimensional random vector and $\boldsymbol{X}'$ is an n-by-d random matrix:

$$\begin{aligned} \boldsymbol{y}'_i &= \sigma_i \cdot y_{\gamma_i} \\ \boldsymbol{X}'_{i,-} &= \sigma_i \cdot X_{\gamma_i,-}\,. \end{aligned}$$

Note that $\boldsymbol{X}'$ and $\boldsymbol{\eta}$ are independent. The next result shows that for $n \gtrsim d \log d$ with high probability over $\boldsymbol{X}'$ Algorithm 11 outputs an instance of Eq. (A.2) as desired.

Theorem A.3. *Let $n \gtrsim d \ln d$. Let $\boldsymbol{X}' \in \mathbb{R}^{n\times d}$ be a matrix obtained from a Gaussian matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ by choosing (independently of $\boldsymbol{X}$) n rows of $\boldsymbol{X}$ with replacement. Then column span of $\boldsymbol{X}'$ is $(\Omega(n), \Omega(1))$-spread with probability $1 - o(1)$ as $n \to \infty$.*

Proof. Let $\boldsymbol{c}(i)$ be the row chosen at i-th step. Let's show that with high probability for all subsets $\mathcal{M} \subseteq [n]$ of size m, $\left|\boldsymbol{c}^{-1}(\mathcal{M})\right| \leqslant O(m \ln(n/m))$. By Chernoff bound (Fact D.2), for any $j \in [n]$ and any $\Delta \geqslant 1$,

$$\mathbb{P}\left(\sum_{i=1}^{n} [\![\boldsymbol{c}(i) = j]\!] \geqslant 2\Delta\right) \leqslant \Delta^{-\Delta}\,. \tag{A.3}$$

Denote $a_m = en/m$. Let $A_j(2\Delta_j)$ be the event $\left\{\sum_{i=1}^{n} \mathbf{1}_{[\boldsymbol{c}(i)=j]} \geqslant 2\Delta_j\right\}$. If $\sum_{j=1}^{m} \Delta_j = 2m \ln a_m$, then

$$\mathbb{P}\left[\bigcap_{j=1}^{m} A_j(2\Delta_j)\right] = \mathbb{P}[A_m(2\Delta_m)] \cdot \mathbb{P}\left[\bigcap_{j=1}^{m-1} A_j(2\Delta_j) \,\middle|\, A_m(2\Delta_m)\right]$$

$$\leqslant \mathbb{P}[A_m(2\Delta_m)] \cdot \mathbb{P}\left[\bigcap_{j=1}^{m-1} A_j(2\Delta_j)\right]$$
$$\leqslant \exp\left(-\sum_{j=1}^{m} \Delta_j\right)$$
$$= a_m^{-2m}\,.$$

By union bound, with probability at least $1 - a_m^{-m}$, for all subsets $\mathcal{M} \subseteq [n]$ of size m, $\left|c^{-1}(\mathcal{M})\right| \leqslant 4m\ln(en/m)$. Let z be the vector with entries $z_j = \sum_{i=1}^{n} [\![c(i) = j]\!]$. With probability at least $1 - 1/n \leqslant 1 - \sum_{m=1}^{n} a_m^{-m}$, for any $\mathcal{M} \subseteq [n]$,

$$\sum_{j\in\mathcal{M}} |z_j| \leqslant 4|\mathcal{M}| \ln\left(\frac{en}{|\mathcal{M}|}\right).$$

Note that by Fact D.12, with probability at least $1 - 1/n$, for any vector unit v from column span of $\boldsymbol{X}$ and any set $\mathcal{M} \subseteq [n]$,

$$\|v_{\mathcal{M}}\|^2 \leqslant 2d + 20|\mathcal{M}| \ln\left(\frac{en}{|\mathcal{M}|}\right).$$

Now let v' be arbitrary unit vector from column span of $\boldsymbol{X}'$. Let $\mathcal{S}$ be the set of its top $m = c^6 \cdot n$ entries for some small enough constant c. Then for some vector v from column span of $\boldsymbol{X}$, with probability $1 - o(1)$,

$$\sum_{i\in\mathcal{S}} (v_i')^2 \leqslant \sum_{j\in c(\mathcal{S})} |z_j| \cdot v_j^2 = \langle z_{c(\mathcal{S})}, v_{c(\mathcal{S})}^2 \rangle \leqslant c^4 \cdot (d\ln(en) + 6n) \cdot \|v\|^2\,,$$

where the last inequality follows from Lemma E.7.

Now let's bound $\|v\|^2$. Let u be a fixed unit vector, then $\boldsymbol{X}u \sim N(0, \mathrm{Id}_n)$. By Chernoff bound, with probability at least $1 - \exp(-cn)$, number of entries of $\boldsymbol{X}u$ bounded by c is at most $2cn$. Note that with high probability $0.9\sqrt{n}\|u\| \leqslant \|\boldsymbol{X}u\| \leqslant 1.1\sqrt{n}\|u\|$

Let $u' \in \mathbb{R}^d$ be a unit vector such that number of entries of $\boldsymbol{X}u'$ bounded by c is at most $2cn$ and let $u \in \mathbb{R}^d$ be a unit vector such that $\|\boldsymbol{X}u' - \boldsymbol{X}u\|^2 \leqslant 1.1^2 \cdot n\|u' - u\|^2 \leqslant c^2 n/5$. Then Xu cannot have more than $3cn$ entries bounded by $c/2$. Hence by union bound over $c/3$-net in the unit ball in $\mathbb{R}^d$, with probability at least $1 - \exp(-cn/2)$, v has at most $3cn$ entries of magnitude smaller than $c/2$. Hence v' has at most $12cn\ln\left(\frac{e}{3c}\right) \leqslant 0.9n$ entries of magnitude smaller than $\frac{c}{3\sqrt{n}}\|v\|$, and $\|v'\|^2 \geqslant \frac{c^2}{100}\|v\|^2$. Choosing $c = 0.01$, we get that column span of $\boldsymbol{X}'$ is $(10^{-12} \cdot n, 1/2)$-spread with high probability. □

Remark A.4. It is perhaps more evident how the model considered in Theorem 5.4 subsumes the one of Theorem 1.1. Given $(\boldsymbol{y}, \boldsymbol{X})$ as in Eq. (A.1) it suffices to multiply the instance by

an independent random matrix $\boldsymbol{U}$ corresponding to flip of signs and permutation of the entries, then add an independent Gaussian vector $\boldsymbol{w} \sim N(0, \mathrm{Id}_n)$. The model then becomes

$$\boldsymbol{U}\boldsymbol{y} = \boldsymbol{U}\boldsymbol{X}\beta^* + \boldsymbol{U}\eta + \boldsymbol{w}\,,$$

which can be rewritten as

$$\boldsymbol{y}' = \boldsymbol{X}'\beta^* + \boldsymbol{U}\eta + \boldsymbol{w}\,.$$

Here $\boldsymbol{X}' \in \mathbb{R}^{n\times d}$ has i.i.d entries $\boldsymbol{X}_{ij} \sim N(0,1)$, $\boldsymbol{U} = \boldsymbol{S}\boldsymbol{P}$ where $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ is a permutation matrix chosen u.a.r. among all permutation matrices, $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ is a diagonal random matrix with i.i.d. Rademacher variables on the diagonal and $\boldsymbol{U}\eta \in \mathbb{R}^n$ is a symmetrically distributed vector independent of $\boldsymbol{X}'$ such that $\mathbb{P}\{|\boldsymbol{\eta}_i| \leqslant 1\} \geqslant \alpha/2$. Moreover the entries of $\boldsymbol{U}\eta$ are conditionally independent given $\boldsymbol{P}$. At this point, we can relax our Gaussian design assumption and consider

$$\boldsymbol{y} = X\beta^* + \boldsymbol{w} + \boldsymbol{U}\eta\,, \tag{A.4}$$

which corresponds to the model considered in Theorem 5.4.

A.2.1 Relaxing well-spread assumptions

It is natural to ask if under weaker assumptions on X we may design an efficient algorithm that correctly recovers β^*. While it is likely that the assumptions in Theorem 1.2 are not tight, *some* requirements are needed if one hopes to design an estimator with bounded error. Indeed, suppose *there exists a vector* $\beta^* \in \mathbb{R}^d$ *and a set* $\mathcal{S} \subseteq [n]$ *of cardinality* $o(1/\alpha)$ *such that* $\|X_{\mathcal{S}}\beta^*\| = \|X\beta^*\| > 0$. Consider an instance of linear regression $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$ with $\boldsymbol{\eta}$ as in Theorem 1.2. Then with probability $1 - o(1)$ any non-zero row containing information about β^* will be corrupted by (possibly unbounded) noise. More concretely:

Lemma A.5. *Let $\sigma > 0$ be arbitrarily chosen. For large enough absolute constant $C > 0$, let $X \in \mathbb{R}^n$ be an $\frac{n}{C\alpha}$-sparse deterministic vector and let $\boldsymbol{\eta}$ be an n-dimensional random vector with i.i.d coordinates sampled as*

$$\begin{aligned} \boldsymbol{\eta}_i &= 0 \quad \textit{with probability } \alpha \\ \boldsymbol{\eta}_i &\sim N(0,\sigma^2) \quad \textit{otherwise.} \end{aligned}$$

Then for every estimator $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}$ there exists $\beta^* \in \mathbb{R}$ such that for $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$, with probability at least $\Omega(1)$,

$$\tfrac{1}{n}\left\|X\left(\hat{\beta}(\boldsymbol{y}) - \beta^*\right)\right\|^2 \geqslant \Omega\left(\frac{\sigma^2}{n}\right).$$

Proof. Let $C \subseteq [n]$ be the set of zero entries of X and $\overline{C}$ its complement. Notice that with probability $1 - \Omega(1)$ over $\boldsymbol{\eta}$ the set $\boldsymbol{S} = \left\{ i \in [n] \;\middle|\; i \in \overline{C} \text{ and } \boldsymbol{\eta}_i = 0 \right\}$ is empty. Conditioning

on this event $\mathcal{E}$, for any estimator $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{\eta}_C$ define the function $g_{\boldsymbol{\eta}_C} : \mathbb{R}^{n-|C|} \to \mathbb{R}$ such that $g_{\boldsymbol{\eta}_C}(\boldsymbol{y}_{\overline{C}}) = \hat{\beta}(\boldsymbol{y})$. Taking distribution over $\boldsymbol{\beta}^*$ from Fact A.1 (independent of $\boldsymbol{\eta}$), we get with probability $\Omega(1)$

$$\tfrac{1}{n}\left\|X\left(\hat{\beta}(\boldsymbol{y}) - \beta^*\right)\right\|^2 = \tfrac{1}{n}\left\|X\left(g_{\boldsymbol{\eta}_C}(\boldsymbol{y}_{\overline{C}}) - \beta^*\right)\right\|^2 \geqslant \Omega\left(\frac{\sigma^2}{n}\right).$$

Hence for any $\hat{\beta}$ there exists β^* with desired property. □

Notice that the noise vector $\boldsymbol{\eta}$ satisfies the premises of Theorem 1.2. Furthermore, since $\sigma > 0$ can be arbitrarily large, no estimator can obtain bounded error.

A.3 On the noise assumptions

On the scaling of noise. Recall our main regression model,

$$\boldsymbol{y} = X\beta^* + \boldsymbol{\eta} \tag{A.5}$$

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

A slight generalization of Eq. (A.5) can be obtained if we allow $\boldsymbol{\eta}$ to have independent, symmetrically distributed coordinates with $\alpha = \min_{i\in[n]} \mathbb{P}\{|\boldsymbol{\eta}_i| \leqslant \sigma\}$. This parameter σ is closely related to the subgaussian parameter σ from [SBRJ19]. If we assume as in [SBRJ19] that some (good enough) estimator of this parameter is given, we could then simply divide each $\boldsymbol{y}_i$ by this estimator and obtain bounds comparable to those of Theorem 1.2. For unknown $\sigma \ll 1$ better error bounds can be obtained if we decrease Huber loss parameter h (for example, it was shown in the Huber loss minimization analysis of [TJSO14]). It is not difficult to see that our analysis (applied to small enough h) also shows similar effect. However, as was also mentioned in [SBRJ19], it is not known whether in general σ can be estimated using only $\boldsymbol{y}$ and X. So for simplicity we assume that $\sigma = 1$.

A.3.1 Tightness of noise assumptions

We provide here a brief discussion concerning our assumptions on the noise vector $\boldsymbol{\eta}$. We argue that, without further assumptions on X, the assumptions on the noise in Theorem 5.4 are tight (notice that such model is more general than the one considered in Theorem 1.2). Fact A.6 and Fact A.7 provide simple arguments that we cannot relax median zero and independence noise assumptions.

Fact A.6 (Tightness of Zero Median Assumption)**.** *Let $\alpha \in (0,1)$ be such that αn is integer, and let $0 < \varepsilon < 1/100$. There exist $\beta, \beta' \in \mathbb{R}^d$ and $X \in \mathbb{R}^{n\times d}$ satisfying*

- *the column span of X is $(n/2, 1/2)$-spread,*

- $\|X(\beta - \beta')\| \geqslant \frac{\varepsilon}{10}\|X\| \geqslant \frac{\varepsilon}{10} \cdot \sqrt{n}$,

and there exist distributions D and D' over vectors in $\mathbb{R}^n$ such that if $\boldsymbol{z} \sim D$ or $\boldsymbol{z} \sim D'$, then:

1. *$\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are mutually inependent,*
2. *For all $i \in [n]$, $\mathbb{P}(\boldsymbol{z}_i \geqslant 0) \geqslant \mathbb{P}(\boldsymbol{z}_i \leqslant 0) \geqslant (1 - \varepsilon) \cdot \mathbb{P}(\boldsymbol{z}_i \geqslant 0)$,*
3. *For all $i \in [n]$, there exists a density p_i of $\boldsymbol{z_i}$ such that $p_i(t) \geqslant 0.1$ for all $t \in [-1, 1]$,*

and for $\boldsymbol{\eta} \sim D$ and $\boldsymbol{\eta}' \sim D'$, random variables $X\beta + \boldsymbol{\eta}$ and $X\beta' + \boldsymbol{\eta}'$ have the same distribution.

Proof. It suffices to consider the one dimensional case. Let $X \in \mathbb{R}^n$ be a vector with all entries equal to 1, let $\beta = 1$ and $\beta' = 1 - \frac{\varepsilon}{10}$. Then $\boldsymbol{\eta} \sim N(0, \mathrm{Id}_n)$ and $\boldsymbol{\eta}' = N(\mu, \mathrm{Id}_n)$ with $\mu = \left(\frac{\varepsilon}{10}, \ldots, \frac{\varepsilon}{10}\right)^\intercal \in \mathbb{R}^n$ satisfy the assumptions, and random variables $X\beta + \boldsymbol{\eta}$ and $X\beta' + \boldsymbol{\eta}'$ have the same distribution. □

Fact A.7 (Tightness of Independence Assumption)**.** *Let $\alpha \in (0, 1)$ be such that αn is integer. There exist $\beta, \beta' \in \mathbb{R}^d$ and $X \in \mathbb{R}^{n \times d}$ satisfying*

- *the column span of X is $(n/2, 1/2)$-spread,*
- *$\|X(\beta - \beta')\| \geqslant \|X\| \geqslant \sqrt{n}$,*

and there exists a distribution D over vectors in $\mathbb{R}^n$ such that $\boldsymbol{\eta} \sim D$ satisfies:

1. *For all $i \in [n]$, $\mathbb{P}(\boldsymbol{\eta}_i \leqslant 0) \geqslant \mathbb{P}(\boldsymbol{\eta}_i \geqslant 0)$,*
2. *For all $i \in [n]$, there exists a density p_i of $\boldsymbol{z_i}$ such that $p_i(t) \geqslant 0.1$ for all $t \in [-1, 1]$,*

and for some $\boldsymbol{\eta}' \sim D$, with probability $1/2$, $X\beta + \boldsymbol{\eta} = X\beta' + \boldsymbol{\eta}'$.

Proof. Again it suffices to consider the one dimensional case. Let $X \in \mathbb{R}^n$ be a vector with all entries equal to 1, $\beta = 0$, $\beta' = 1$. Let $\boldsymbol{\sigma} \sim U\{-1, 1\}$ and let $\boldsymbol{v} \in \mathbb{R}^n$ be a vector independent of $\boldsymbol{\sigma}$ such that for all $i \in [n]$, the entries of $\boldsymbol{v}$ are iid $\boldsymbol{v}_i = U[0, 1]$. Then $\boldsymbol{\eta} = \boldsymbol{\sigma} \cdot \boldsymbol{v}$ and $\boldsymbol{\eta}' = -\boldsymbol{\sigma}(1 - \boldsymbol{v})$ satisfy the assumptions, and if $\boldsymbol{\sigma} = 1$, $X\beta + \boldsymbol{\eta} = X\beta' + \boldsymbol{\eta}'$. □

Note that the assumptions in both facts do not contain $\boldsymbol{\mathcal{R}}$ as opposed to the assumptions of Theorem 5.4. One can take $\boldsymbol{\mathcal{R}}$ to be a random subset of $[n]$ of size αn independent of $\boldsymbol{\eta}$ and $\boldsymbol{\eta}'$.

B Computing the Huber-loss estimator in polynomial time

In this section we show that in the settings of Theorem 1.2, we can compute the Huber-loss estimator efficiently. For a vector $v \in \mathbb{Q}^N$ we denote by $\mathtt{b}[v]$ its bit complexity. For $r > 0$ we denote by $\mathcal{B}(0, r)$ the Euclidean ball of radius r centered at 0. We consider the Huber loss function as in Definition 5.3 and for simplicity we will assume $X^\intercal X = n\mathrm{Id}_d$.

Theorem B.1. *Let $X \in \mathbb{Q}^{n\times d}$ be a matrix such that $X^\intercal X = n\mathrm{Id}_d$ and $y \in \mathbb{Q}^n$ be a vector. Let*

$$\mathtt{B} := \mathtt{b}[y] + \mathtt{b}[X]$$

Let f be a Huber loss function $f(\beta) = \sum_{i=1}^n \Phi\big((X\beta - y)_i\big)$. Then there exists an algorithm that given X, y and positive $\varepsilon \in \mathbb{Q}$, computes a vector $\overline{\beta} \in \mathbb{R}^d$ such that

$$f(\overline{\beta}) \leqslant \inf_{\beta\in\mathbb{R}^d} f(\beta) + \varepsilon\,,$$

in time

$$\mathtt{B}^{O(1)} \cdot \ln(1/\varepsilon)\,.$$

As an immediate corollary, the theorem implies that for design matrix X with orthogonal columns we can compute an ε-close approximation of the Huber loss estimator in time polynomial in the input size.

To prove Theorem B.1 we will rely on the following standard result concerning the Ellipsoid algorithm.

Theorem B.2 (See [Vis18])**.** *There is an algorithm that, given*

1. *a first-order oracle for a convex function $g : \mathbb{R}^d \to R$,*
2. *a separation oracle for a convex set $K \subseteq R^d$,*
3. *numbers $r > 0$ and $R > 0$ such that $\mathcal{B}(0, r) \subset K \subset \mathcal{B}(0, R)$,*
4. *bounds ℓ, u such that $\forall v \in K$, $\ell \leqslant g(v) \leqslant u$*
5. *$\varepsilon > 0$,*

outputs a point $\overline{x} \in K$ such that

$$g(\overline{x}) \leqslant g(\hat{x}) + \varepsilon\,,$$

where $\hat{x}$ is any minimizer of g over K. The running time of the algorithm is

$$O\left(\left(d^2 + T_K + T_g\right) \cdot d^2 \cdot \log\left(\frac{R}{r} \cdot \frac{u-\ell}{\varepsilon}\right)\right),$$

where T_k, T_g are the running times for the separation oracle for K and the first-order oracle for g respectively.

We only need to apply Theorem B.2 to the settings of Theorem B.1. Let $\hat{\beta}$ be a (global) minimizer of f. Our first step is to show that $\|\hat{\beta}\|$ is bounded by $\exp(O(\mathtt{B}))$.

Lemma B.3. *Consider the settings of Theorem B.1. Then* $\left\|\hat{\beta}\right\| \leqslant 2^{5\mathtt{B}}$. *Moreover, for any* $\beta \in \mathcal{B}(0, 2^{10\mathtt{B}})$

$$0 \leqslant f(\beta) \leqslant 2^{12\mathtt{B}}.$$

Proof. Let $M = 2^{\mathtt{B}}$. By definition $f(0) \leqslant 2\|y\|_1 + \frac{1}{2}\|y\|^2 \leqslant M^4$. On the other hand for any $v \in \mathbb{R}^d$ with $\|v\| \geqslant M^5$ we have

$$\begin{aligned} f(v) &= \sum_{i\in[n]} \Phi\big(y_i - \langle X_i, v\rangle\big) \\ &\geqslant \sum_{i\in[n]} \Phi(\langle X_i, v\rangle) - \left\|y\right\|^2 - \left\|y\right\|_1 \\ &\geqslant M^5 - M^4 \\ &\geqslant M^4. \end{aligned}$$

It follows that $\|\hat{\beta}\| \leqslant M^5$. For the second inequality note that for any $v \in \mathbb{R}^d$ with $\|v\| \leqslant M^{10}$

$$\begin{aligned} f(v) &= \sum_{i\in[n]} \Phi\big(y_i - \langle X_i, v\rangle\big) \\ &\leqslant 2\sum_{i\in[n]} \Phi(\langle X_i, v\rangle) + \left\|y\right\|^2 + 2\left\|y\right\|_1 \\ &\leqslant M^{12}. \end{aligned}$$

□

Next we state a simple fact about the Huber loss function and separation oracles, which proof we omit. Recall the formula for the gradient of the Huber loss function. $\nabla f(\beta^*) = \frac{1}{n}\sum_{i=1}^{n} \phi'[\eta_i] \cdot x_i$ with $\Phi'[t] = \operatorname{sign}(t) \cdot \min\{|t|, h\}$.

Fact B.4. *Consider the settings of Theorem B.1. Let* $R = 2^{10\mathtt{B}}$. *Then*

1. *there exists an algorithm that given* $v \in \mathbb{Q}^d$, *computes* $\nabla f(v)$ *and* $f(v)$ *in time* $\mathtt{B}^{O(1)} \cdot \mathtt{b}^{O(1)}[v]$.
2. *there exists an algorithm that given* $v \in \mathbb{Q}^d$ *outputs*
   - *YES if* $v \in \mathcal{B}(0, R)$
   - *otherwise outputs a hyperplane* $\left\{x \in \mathbb{R}^d \;\middle|\; \langle a, x\rangle = b\right\}$ *with* $a, b \in \mathbb{Q}^d$ *separating* v *from* $\mathcal{B}(0, R)$,

   in time $\mathtt{B}^{O(1)} \cdot \mathtt{b}^{O(1)}[v]$.

We are now ready to prove Theorem B.1.

Proof of Theorem B.1. Let $M = 2^{\mathtt{B}}$. By Lemma B.3 it suffices to set $K = \mathcal{B}(0, M^{10})$, $R = M^{10}$ and $r = R/2$. Then for any $v \in K$, $0 \leqslant f(v) \leqslant M^{12}$. By Fact B.4 $T_f + T_K \leqslant B^{O(1)}$. Thus, puttings things together and plugging Theorem B.2 it follows that there exists an algorithm computing $\overline{\beta}$ with $f(\overline{\beta}) \leqslant f(\hat{\beta}) + \varepsilon$ in time

$$\mathtt{B}^{O(1)} \cdot \ln(1/\varepsilon)$$

for positive $\varepsilon \in \mathbb{Q}$. □

C Consistent estimators in high-dimensional settings

In this section we discuss the generalization of the notion of consistency to the case when the dimension d (and the fraction of inliers α) can depend on n.

Definition C.1 (Estimator)**.** Let $\{d(n)\}_{n=1}^{\infty}$ be a sequence of positive integers. We call a function $\hat{\beta} : \bigcup_{n=1}^{\infty} \mathbb{R}^n \times \mathbb{R}^{n \times d(n)} \to \bigcup_{n=1}^{\infty} \mathbb{R}^{d(n)}$ an *estimator*, if for all $n \in \mathbb{N}$, $\hat{\beta}\big(\mathbb{R}^n \times \mathbb{R}^{n \times d(n)}\big) \subseteq \mathbb{R}^{d(n)}$ and the restriction of $\hat{\beta}$ to $\mathbb{R}^n \times \mathbb{R}^{n \times d(n)}$ is a Borel function.

For example, Huber loss defined at $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times d(n)}$ as

$$\hat{\beta}(y, X) = \operatorname*{argmin}_{\beta \in \mathbb{R}^{d(n)}} \sum_{i=1}^{n} \Phi\big((X\beta - y)_i\big)$$

is an estimator (see Lemma C.4 for formal statement and the proof).

Definition C.2 (Consistent estimator)**.** Let $\{d(n)\}_{n=1}^{\infty}$ be a sequence of positive integers and let $\hat{\beta} : \bigcup_{n=1}^{\infty} \mathbb{R}^n \times \mathbb{R}^{n \times d(n)} \to \bigcup_{n=1}^{n} \mathbb{R}^{d(n)}$ be an estimator. Let $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ be a sequence of (possibly random) matrices and let $\{\boldsymbol{\eta}_n\}_{n=1}^{\infty}$ be a sequence of (possibly random) vectors such that $\forall n \in \mathbb{N}$, $\boldsymbol{X}_n$ has dimensions $n \times d(n)$ and $\boldsymbol{\eta}_n$ has dimension n.

We say that estimator $\hat{\beta}$ is *consistent* for $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ and $\{\boldsymbol{\eta}_n\}_{n=1}^{\infty}$ if there exists a sequence of positive numbers $\{\varepsilon_n\}_{n=1}^{n}$ such that $\lim_{n\to\infty} \varepsilon_n = 0$ and for all $n \in \mathbb{N}$,

$$\sup_{\beta^* \in \mathbb{R}^{d(n)}} \mathbb{P}\big(\tfrac{1}{n}\|\boldsymbol{X}_n\hat{\beta}\big(\boldsymbol{X}_n\beta^* + \boldsymbol{\eta}_n\,, \boldsymbol{X}_n\big) - \boldsymbol{X}_n\beta^*\|^2 \geqslant \varepsilon_n\big) \leqslant \varepsilon_n\,.$$

Theorem 1.1 implies that if sequences $d(n)$ and $\alpha(n)$ satisfy $d(n)/\alpha^2(n) \leqslant o(n)$ and $d(n) \to \infty$, then Huber loss estimator is consistent for a sequence $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ of standard Gaussian matrices $\boldsymbol{X}_n \sim N(0,1)^{n \times d(n)}$ and every sequence of vectors $\{\eta_n\}_{n=1}^{\infty}$ such that each $\eta_n \in \mathbb{R}^n$ is independent of $\boldsymbol{X}_n$ and has at least $\alpha(n) \cdot n$ entries of magnitude at most 1.

Similarly, Theorem 1.2 implies that if sequences $d(n)$ and $\alpha(n)$ satisfy $d(n)/\alpha^2(n) \leqslant o(n)$ and $d(n) \to \infty$, then Huber loss estimator is consistent for each sequence $\{X_n\}_{n=1}^{\infty}$ of matrices $X_n \in \mathbb{R}^{n \times d(n)}$ whose column span is $\big(\omega(d(n)/\alpha^2(n)), \Omega(1)\big)$-spread and every

sequence of n-dimensional random vectors $\{\boldsymbol{\eta}_n\}_{n=1}^{\infty}$ such that each $\boldsymbol{\eta}_n$ is independent of X_n and has mutually independent, symmetrically distributed entries whose magnitude does not exceed 1 with probability at least $\alpha(n)$.

Note that the algorithm from Theorem 1.3 requires some bound on $\|\beta^*\|$. So formally we cannot say that the estimator that is computed by this algorithm is consistent. However, if in Definition C.2 we replace supremum over $\mathbb{R}^{d(n)}$ by supremum over some ball in $\mathbb{R}^{d(n)}$ centered at zero (say, of radius n^{100}), then we can say that this estimator is consistent. More precisely, it is consistent (according to modified definition with sup over ball of radius n^{100}) for sequence $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ of standard Gaussian matrices $\boldsymbol{X}_n \sim N(0,1)^{n\times d(n)}$ and every sequence of vectors $\{\eta_n\}_{n=1}^{\infty}$ such that each $\eta_n \in \mathbb{R}^n$ is independent of $\boldsymbol{X}_n$ and has at least $\alpha(n)\cdot n$ entries of magnitude at most 1, if $n \gtrsim d(n)\log^2(d(n))/\alpha^2(n)$ and $d(n)\to\infty$.

To show that Huber loss minimizer is an estimator, we need the following fact:

Fact C.3. *[Jen69] For $d, N \in \mathbb{N}$, let $\Theta \subset \mathbb{R}^d$ be compact and let $\mathcal{M} \subseteq \mathbb{R}^N$ be a Borel set. Let $f : \mathcal{M}\times\Theta \to \mathbb{R}$ be a function such that for each $\theta\in\Theta$, $f(x,\theta)$ is a Borel function of x and for each $x\in\mathcal{M}$, $f(x,\theta)$ is a continuous function of θ. Then there exists a Borel function $\hat{\theta}:\mathcal{M}\to\Theta$ such that for all $x\in\mathcal{M}$,*

$$f(x,\hat{\theta}(x)) = \min_{\theta\in\Theta} f(x,\theta)\,.$$

The following lemma shows that Huber loss minimizer is an estimator.

Lemma C.4. *Let $\{d(n)\}_{n=1}^{\infty}$ be a sequence of positive integers. There exists an estimator $\hat{\beta}$ such that for each $n\in\mathbb{N}$ and for all $y\in\mathbb{R}^n$ and $X\in\mathbb{R}^{n\times d(n)}$,*

$$\sum_{i=1}^{n}\Phi\big(\big(X\hat{\beta}(y,X)-y\big)_i\big) = \min_{\beta\in\mathbb{R}^{d(n)}}\sum_{i=1}^{n}\Phi\big(\big(X\beta-y\big)_i\big)\,.$$

Proof. For $i\in\mathbb{N}$ denote

$$\mathcal{M}_n^i = \left\{(y,X)\in\mathbb{R}^n\times\mathbb{R}^{n\times d(n)} \;\middle|\; \|y\|\leqslant 2^i\,,\sigma_{\min}(X)\geqslant 2^{-i}\right\},$$

where $\sigma_{\min}(X)$ is the smallest positive singular value of X. Note that for all $(y,X)\in\mathcal{M}_n^i$, there exists a minimizer of Huber loss function at (y,X) in the ball $\left\{\beta\in\mathbb{R}^{d(n)} \;\middle|\; \|\beta\|\leqslant 2^{2i}n^{10}\right\}$. By Fact C.3, there exists a Borel measurable Huber loss minimizer $\hat{\beta}_n^i(y,X)$ on $\mathcal{M}_n^i$.

Denote $\mathcal{M}_n^0 = \left\{(y,X)\in\mathbb{R}^n\times\mathbb{R}^{n\times d(n)} \;\middle|\; X=0\right\}$ and let $\hat{\beta}_n^0(y,X)=0$ for all $(y,X)\in\mathcal{M}_n^0$. Note that $\bigcup_{i=0}^{\infty}\mathcal{M}_n^i = \mathbb{R}^n\times\mathbb{R}^{n\times d(n)}$. For $(y,X)\in\mathbb{R}^n\times\mathbb{R}^{n\times d(n)}$, define $\hat{\beta}_n(y,X)=\hat{\beta}_n^i(y,X)$, where i is a minimal index such that $(y,X)\in\mathcal{M}_n^i$. Then for each Borel set $\mathcal{B}\subseteq\mathbb{R}^{d(n)}$,

$$(\hat{\beta}_n)^{-1}(\mathcal{B}) = \bigcup_{i=0}^{\infty}(\hat{\beta}_n)^{-1}(\mathcal{B})\cap\left(\mathcal{M}_n^i\setminus\mathcal{M}_n^{i-1}\right) = \bigcup_{i=0}^{\infty}\left(\hat{\beta}_n^i\right)^{-1}(\mathcal{B})\cap\left(\mathcal{M}_n^i\setminus\mathcal{M}_n^{i-1}\right).$$

Hence $\hat{\beta}_n$ is a Borel function. Now for each $n\in\mathbb{N}$ and for all $y\in\mathbb{R}^n$ and $X\in\mathbb{R}^{n\times d(n)}$ define an estimator $\hat{\beta}(y,X)=\hat{\beta}_n(y,X)$. □

D Concentration of measure

This section contains some technical results needed for the proofs of Theorem 1.2 and Theorem 1.3. We start by proving a concentration bound for the empirical median.

Fact D.1 ([Ver18]). *Let $0 < \varepsilon < 1$. Let $\mathcal{B} = \{v \in \mathbb{R}^n \mid \|v\| \leqslant 1\}$. Then $\mathcal{B}$ has an ε-net of size $\left(\frac{2+\varepsilon}{\varepsilon}\right)^n$. That is, there exists a set $\mathcal{N}_\varepsilon \subseteq \mathcal{B}$ of size at most $\left(\frac{2+\varepsilon}{\varepsilon}\right)^n$ such that for any vector $u \in \mathcal{B}$ there exists some $v \in \mathcal{N}_\varepsilon$ such that $\|v - u\| \leqslant \varepsilon$.*

Fact D.2 (Chernoff's inequality, [Ver18]). *Let $\zeta_1, \ldots, \zeta_n$ be independent Bernoulli random variables such that $\mathbb{P}(\zeta_i = 1) = \mathbb{P}(\zeta_i = 0) = p$. Then for every $\Delta > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n \zeta_i \geqslant pn(1+\Delta)\right) \leqslant \left(\frac{e^{-\Delta}}{(1+\Delta)^{1+\Delta}}\right)^{pn}.$$

and for every $\Delta \in (0,1)$,

$$\mathbb{P}\left(\sum_{i=1}^n \zeta_i \leqslant pn(1-\Delta)\right) \leqslant \left(\frac{e^{-\Delta}}{(1-\Delta)^{1-\Delta}}\right)^{pn}.$$

Fact D.3 (Hoeffding's inequality, [Wai19]). *Let $z_1, \ldots, z_n$ be mutually independent random variables such that for each $i \in [n]$, z_i is supported on $[-c_i, c_i]$ for some $c_i \geqslant 0$. Then for all $t \geqslant 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n (z_i - \mathbb{E}\, z_i)\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n c_i^2}\right).$$

Fact D.4 (Bernstein's inequality [Wai19]). *Let $z_1, \ldots, z_n$ be mutually independent random variables such that for each $i \in [n]$, z_i is supported on $[-B, B]$ for some $B \geqslant 0$. Then for all $t \geqslant 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n (z_i - \mathbb{E}\, z_i) \geqslant t\right) \leqslant \exp\left(-\frac{t^2}{2\sum_{i=1}^n \mathbb{E}\, z_i^2 + \frac{2Bt}{3}}\right).$$

Lemma D.5 (Restate of Lemma 4.6). *Let $\mathcal{S} \subseteq [n]$ be a set of size γn and let $z_1, \ldots, z_n \in \mathbb{R}$ be mutually independent random variables satisfying*

1. *For all $i \in [n]$, $\mathbb{P}(z_i \geqslant 0) = \mathbb{P}(z_i \leqslant 0)$.*
2. *For some $\varepsilon \geqslant 0$, for all $i \in \mathcal{S}$, $\mathbb{P}(z_i \in [0, \varepsilon]) = \mathbb{P}(z_i \in [-\varepsilon, 0]) \geqslant q$.*

Then with probability at least $1 - 2\exp\{-\Omega(q^2\gamma^2 n)\}$ the median $\hat{z}$ satisfies

$$|\hat{z}| \leqslant \varepsilon\,.$$

Proof. Let $\mathcal{Z} = \{z_1, \dots, z_n\}$. Consider the following set:

$$\mathcal{A} := \{z \in \mathcal{Z} \mid |z| \leqslant \varepsilon\}.$$

Denote $\mathcal{Z}^+ = \mathcal{Z} \cap \mathbb{R}_{\geqslant 0}$, $\mathcal{A}^+ = \mathcal{A} \cap \mathbb{R}_{\geqslant 0}$, $\mathcal{Z}^- = \mathcal{Z} \cap \mathbb{R}_{\leqslant 0}$, $\mathcal{A}^- = \mathcal{A} \cap \mathbb{R}_{\leqslant 0}$. Applying Chernoff bound for $\gamma_1, \gamma_2, \gamma_3 \in (0,1)$,

$$\mathbb{P}\left(\left|\mathcal{Z}^+\right| \leqslant \left(\frac{1}{2} - \gamma_1\right)n\right) \leqslant \exp\left\{-\frac{\gamma_1^2 \cdot n}{10}\right\},$$

$$\mathbb{P}\left(|\mathcal{A}| \leqslant (1-\gamma_2) \cdot q \cdot |\mathcal{S}|\right) \leqslant \exp\left\{-\frac{\gamma_2^2 \cdot q \cdot |\mathcal{S}|}{10}\right\},$$

$$\mathbb{P}\left(\left|\mathcal{A}^+\right| \leqslant \left(\frac{1}{2} - \gamma_3\right) \cdot |\mathcal{A}| \,\middle|\, |\mathcal{A}|\right) \leqslant \exp\left\{-\frac{\gamma_3^2 \cdot |\mathcal{A}|}{10}\right\}.$$

Similar bounds hold for $\mathcal{Z}^-, \mathcal{A}^-$.

Now, the median is in $\mathcal{A}$ if $|\mathcal{Z}^-| + \left|\mathcal{A}^+\right| \geqslant n/2$ and $\left|\mathcal{Z}^+\right| + |\mathcal{A}^-| \geqslant n/2$. It is enough to prove one of the two inequalities, the proof for the other is analogous. A union bound then concludes the proof.

So for $\gamma_2 = \gamma_3 = \frac{1}{4}$, with probability at least $1 - \exp\left\{-\frac{\gamma_1^2 \cdot n}{10}\right\} - 2\exp\left\{-\Omega\left(q \cdot |\mathcal{S}|\right)\right\}$,

$$|\mathcal{Z}^-| + \left|\mathcal{A}^+\right| \geqslant \left(\frac{1}{2} - \gamma_1\right)n + \frac{q \cdot |\mathcal{S}|}{10}.$$

it follows that $|\mathcal{Z}^-| + \left|\mathcal{A}^+\right| \geqslant n/2$ for

$$\gamma_1 \leqslant \frac{q \cdot |\mathcal{S}|}{10n}.$$

□

Lemma D.6. *Let V be an m-dimensional vector subspace of $\mathbb{R}^n$. Let $\mathcal{B} \subseteq \{v \in V \mid \|v\| \leqslant R\}$ for some $R \geqslant 1$.*

Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function such that for all $y \in \mathbb{R}$ and $|x| \leqslant R$, $|g(x,y)| \leqslant C|x|$ for some $C \geqslant 1$ and for any $|\Delta x| \leqslant 1$, $|g(x+\Delta x, y) - g(x,y)| \leqslant K|\Delta x|$ for some $K \geqslant 1$.

Let $\boldsymbol{w} \in \mathbb{R}^n$ be a random vector such that $\boldsymbol{w}_1, \dots, \boldsymbol{w}_n$ are mutually independent. For any $N \geqslant n$, with probability at least $1 - N^{-m}$, for all $v \in \mathcal{B}$,

$$\left|\sum_{i=1}^{n}\left(g(v_i, \boldsymbol{w}_i) - \mathop{\mathbb{E}}_{\boldsymbol{w}} g(v_i, \boldsymbol{w}_i)\right)\right| \leqslant 10C\sqrt{m\ln(RKN)} \cdot \|v\| + 1/N.$$

Proof. Consider some $v \in \mathbb{R}^n$. Since $|g_i(v_i, \boldsymbol{w}_i)| \leqslant C|v_i|$, by Hoeffding's inequality,

$$\left|\sum_{i=1}^{n}\left(g(v_i, \boldsymbol{w}_i) - \mathop{\mathbb{E}}_{\boldsymbol{w}} g(v_i, \boldsymbol{w}_i)\right)\right| \leqslant \tau C\|v\|$$

with probability $1 - 2\exp(-\tau^2/2)$.

Let $N \geqslant n$ and $\varepsilon = \frac{1}{2KnN}$. Denote by $\mathcal{N}_\varepsilon$ some ε-net in $\mathcal{B}$ such that $|\mathcal{N}_\varepsilon| \leqslant \left(6\frac{R}{\varepsilon}\right)^m$. By union bound, for any $v \in \mathcal{N}_\varepsilon$,

$$\left|\sum_{i=1}^n \Big(g(v_i, \boldsymbol{w}_i) - \mathop{\mathbb{E}}_{\boldsymbol{w}} g(v_i, \boldsymbol{w}_i)\Big)\right| \leqslant 10C\sqrt{m\ln(RKN)} \cdot \|v\|$$

with probability at least $1 - N^{-m}$.

Consider arbitrary $\Delta v \in V$ such that $\|\Delta v\| \leqslant \varepsilon$. For any $v \in \mathcal{N}_\varepsilon$ and $w \in \mathbb{R}^n$,

$$\left|\sum_{i=1}^n \big(g(v_i + \Delta v_i, w_i) - g(v_i, w_i)\big)\right| \leqslant \sum_{i=1}^n \big|g(v_i + \Delta v_i, w_i) - g(v_i, w_i)\big| \leqslant K\sum_{i=1}^n |\Delta v_i| \leqslant \frac{1}{2N}\,.$$

Hence

$$\left|\mathop{\mathbb{E}}_{\boldsymbol{w}} \sum_{i=1}^n \big(g(v_i + \Delta v_i, \boldsymbol{w}_i) - g(v_i, \boldsymbol{w}_i)\big)\right| \leqslant \mathop{\mathbb{E}}_{\boldsymbol{w}}\left|\sum_{i=1}^n \big(g(v_i + \Delta v_i, \boldsymbol{w}_i) - g(v_i, \boldsymbol{w}_i)\big)\right| \leqslant \frac{1}{2N}\,,$$

and

$$\left|\sum_{i=1}^n \Big(g(v_i, \boldsymbol{w}_i) - \mathop{\mathbb{E}}_{\boldsymbol{w}} g(v_i, \boldsymbol{w}_i)\Big) - \sum_{i=1}^n \Big(g(v_i + \Delta v_i, w_i) - \mathop{\mathbb{E}}_{\boldsymbol{w}} g(\Delta v_i, \boldsymbol{w}_i)\Big)\right| \leqslant 1/N\,.$$

Therefore, with probability $1 - N^{-m}$, for any $v \in \mathcal{B}$,

$$\left|\sum_{i=1}^n \Big(g(v_i, \boldsymbol{w}_i) - \mathop{\mathbb{E}}_{\boldsymbol{w}} g(v_i, \boldsymbol{w}_i)\Big)\right| \leqslant 10C\sqrt{m\ln(RKN)} \cdot \|v\| + 1/N\,.$$

□

Lemma D.7. *Let $\zeta_1, \dots, \zeta_n$ be i.i.d. Bernoulli random variables such that $\mathbb{P}(\zeta_i = 1) = 1 - \mathbb{P}(\zeta_i = 0) = m/n$ for some integer $m \leqslant n$. Denote $\boldsymbol{\mathcal{S}}_1 = \{i \in [n] \mid \zeta_i = 1\}$. Let $\boldsymbol{\mathcal{S}}_2 \subseteq [n]$ be a random set chosen uniformly from all subsets of $[n]$ of size exactly m.*

Let P be an arbitrary property of subsets of $[n]$. If $\boldsymbol{\mathcal{S}}_1$ satisfies P with probability at least $1 - \varepsilon$ (for some $0 \leqslant \varepsilon \leqslant 1$), then $\boldsymbol{\mathcal{S}}_2$ satisfies P with probability at least $1 - 2\sqrt{n}\,\varepsilon$.

Proof. If $m = 0$ or $m = n$, then $\boldsymbol{\mathcal{S}}_1 = \boldsymbol{\mathcal{S}}_2$ with probability 1. So it is enough to consider the case $0 < m < n$. By Stirling's approximation, for any integer $k \geqslant 1$,

$$\sqrt{2\pi k} \cdot \frac{k^k}{e^k} \leqslant k! \leqslant \sqrt{2\pi k} \cdot \frac{k^k}{e^{k-1/(12k)}} \leqslant 1.1 \cdot \sqrt{2\pi k} \cdot \frac{k^k}{e^k}\,.$$

Hence

$$\mathbb{P}(|\boldsymbol{\mathcal{S}}_1| = m) = \binom{n}{m}\left(\frac{m}{n}\right)^m\left(\frac{n-m}{n}\right)^{n-m} \geqslant \frac{\sqrt{n}}{1.1^2 \cdot \sqrt{2\pi m(n-m)}} \geqslant \frac{1}{2\sqrt{n}}\,.$$

Therefore,

$$\mathbb{P}(\boldsymbol{S}_2 \not\in P) = \mathbb{P}(\boldsymbol{S}_1 \not\in P \mid |\boldsymbol{S}_1| = m) \leqslant \frac{\mathbb{P}(\boldsymbol{S}_1 \not\in P)}{\mathbb{P}(|\boldsymbol{S}_1| = m)} \leqslant 2\sqrt{n}\,\varepsilon\,.$$

□

Fact D.8 (Covariance estimation of Gaussian vectors, [Wai19])**.** *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ be iid $\boldsymbol{x}_i \sim N(0, \Sigma)$ for some positive definite $\Sigma \in \mathbb{R}^d$. Then, with probability at least $1 - \delta$,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} - \Sigma \right\| \leqslant O\left(\sqrt{\frac{d + \log(1/\delta)}{n}} + \frac{d + \log(1/\delta)}{n} \right) \cdot \|\Sigma\|\,.$$

Lemma D.9. *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ be iid $\boldsymbol{x}_i \sim N(0, \Sigma)$ for some positive definite $\Sigma \in \mathbb{R}^d$. Let $k \in [d]$. Then, with probability at least $1 - \varepsilon$, for any k-sparse unit vector $v \in \mathbb{R}^d$,*

$$v^\mathsf{T} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} - \Sigma \right) v \leqslant O\left(\sqrt{\frac{k \log d + \log(1/\varepsilon)}{n}} + \frac{k \log d + \log(1/\varepsilon)}{n} \right) \cdot \|\Sigma\|\,.$$

Proof. If $d = 1$, $1 - d^{-k} = 0$ and the statement is true, so assume $d > 1$. Consider some set $\mathcal{S} \subseteq [d]$ of size at most k. By Fact D.8, with probability at least $1 - \delta$, for any k-sparse unit vector v with support $\mathcal{S}$,

$$v^\mathsf{T} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} - \Sigma \right) v \leqslant O\left(\sqrt{\frac{k + \log(1/\delta)}{n}} + \frac{k + \log(1/\delta)}{n} \right) \cdot \|\Sigma\|\,.$$

Since there are at most $\exp(2k \ln d)$ subsets of $[d]$ of size at most k, the lemma follows from union bound with $\delta = \varepsilon \exp(-3k \ln d)$. □

Fact D.10 (Chi-squared tail bounds, [LM00])**.** *Let $X \sim \chi^2_m$ (that is, a squared norm of standard m-dimensional Gaussian vector). Then for all $x > 0$*

$$\mathbb{P}\left(X - m \geqslant 2x + 2\sqrt{mx} \right) \leqslant e^{-x}$$
$$\mathbb{P}\left(m - X \geqslant 2\sqrt{xm} \right) \leqslant e^{-x}$$

Fact D.11 (Singular values of Gaussian matrix, [Wai19])**.** *Let $W \sim N(0,1)^{n \times d}$, and assume $n \geqslant d$. Then for each $t \geqslant 0$*

$$\mathbb{P}\left(\sigma_{\max}(W) \geqslant \sqrt{n} + \sqrt{d} + \sqrt{t} \right) \leqslant \exp(-t/2)$$

and

$$\mathbb{P}\left(\sigma_{\min}(W) \leqslant \sqrt{n} - \sqrt{d} - \sqrt{t} \right) \leqslant \exp(-t/2)\,,$$

where $\sigma_{\max}(W)$ and $\sigma_{\min}(W)$ are the largest and the smallest singular values of W.

Fact D.12 (k-sparse norm of a Gaussian matrix)**.** *Let $W \sim N(0,1)^{n\times d}$ be a Gaussian matrix. Let $1 \leqslant k \leqslant n$. Then for every $\delta > 0$ with probability at least $1-\delta$,*

$$\max_{\substack{u\in\mathbb{R}^d \\ \|u\|=1}} \max_{\substack{k\text{-sparse } v\in\mathbb{R}^n \\ \|v\|=1}} v^\mathsf{T} W u \leqslant \sqrt{d} + \sqrt{k} + \sqrt{2k\ln\left(\frac{en}{k}\right)} + \sqrt{2\ln(1/\delta)}\,.$$

Proof. Let v be some k-sparse unit vector that maximizes the value, and let $S(v)$ be the set of nonzero coordinates of v. Consider some fixed (independend of W) unit k-sparse vector $x \in \mathbb{R}^n$ and the set $S(x)$ of nonzero coordinates of x. If we remove from W all the rows with indices not from $S(x)$, we get an $k \times d$ Gaussian matrix $W_{S(x)}$. By Fact D.11, norm of this matrix is bounded by $\sqrt{d} + \sqrt{k} + \sqrt{t}$ with probability at least $\exp(-t/2)$. Number of all subsets $S \subseteq [n]$ of size k is $\binom{n}{k} \leqslant \left(\frac{en}{k}\right)^k$. By union bound, the probability that the norm of $W_{S(v)}$ is greater than $\sqrt{d} + \sqrt{k} + \sqrt{t}$ is at most

$$\binom{n}{k} \cdot \exp(-t/2) \leqslant \exp(k\ln(en/k) - t/2)\,.$$

Taking $t = 2k\ln(en/k) + 2\log(1/\delta)$, we get the desired bound. □

E Spreadness notions of subspaces

Lemma E.1. *Let $v \in \mathbb{R}^n$ be a vector with $\frac{1}{n}\|v\|^2 = 1$. Suppose*

$$\tfrac{1}{n}\sum_{i=1}^{n} [\![v_i^2 \leqslant 1/\delta]\!] \cdot v_i^2 \geqslant \kappa\,.$$

Then, $\frac{1}{n}\|v_S\|^2 \geqslant \kappa/2$ for every subset $S \subseteq [n]$ with $|S| \geqslant (1-\delta\kappa/2)n$.

The above lemma is tight in the sense that there are vectors v that satisfy the premise but have $\|v_S\| = 0$ for a subset $S \subseteq [n]$ of size $|S| = (1-\delta k)n$.

Proof. Let $T \subseteq [n]$ consist of the $\delta\kappa/2 \cdot n$ entries of v. Let $w_i = [\![v_i^2 \leqslant 1/\delta]\!] \cdot v_i^2$. Then, $\frac{1}{n}\|w_T\|^2 \leqslant 1/\delta \cdot \delta\kappa/2 \leqslant \kappa/2$. Thus, $\frac{1}{n}\|v_S\|^2 \geqslant \frac{1}{n}\|w\|^2 - \frac{1}{n}\|w_T\|^2 \geqslant \kappa/2$. □

Lemma E.2. *Let $v \in \mathbb{R}^n$ be a vector with $\frac{1}{n}\|v\|^2 = 1$. Suppose $\frac{1}{n}\|v_S\|^2 \geqslant \kappa$ for every subset $S \subseteq [n]$ with $|S| \geqslant (1-\delta)n$. Then*

$$\tfrac{1}{n}\sum_{i=1}^{n} [\![v_i^2 \leqslant 1/\delta]\!] \cdot v_i^2 \geqslant \kappa\,.$$

Proof. The number of entries satisfying $v_i^2 \leqslant 1/\delta$ is at least $(1-\delta)n$. □

Lemma E.3. *Let $V \subseteq \mathbb{R}^n$ be a vector subspace. Assume that for some $\rho, R \in (0,1)$, for all $v \in V$,*

$$\sum_{i=1}^{n} \left[\!\left[v_i^2 \leqslant \tfrac{1}{R^2 n} \|v\|^2 \right]\!\right] \cdot v_i^2 \geqslant \rho^2 ||v||^2 .$$

Then V is $\left(\frac{\rho^2}{4} R^2 n, \frac{\rho}{2}\right)$-spread.

Proof. Let $m = \frac{\rho^2}{4} R^2 n$. For vector $v \in V$ with $||v||^2 = m$, the set $M = \{i \in [n] \mid v_i^2 > 1\}$ has size at most m, so its complement is a disjoint union of three sets: the set S of $n - m$ smallest entries of v, the set $M' \subseteq [n] \setminus S$ of entries of magnitude $\leqslant \rho/2$, and the set of entries $M'' \subseteq [n] \setminus (S \cup M')$ of magnitude between $\rho/2$ and 1. Note that since $|M'| \leqslant m$, $\sum_{i \in M'} v_i^2 \leqslant \frac{1}{4}\rho^2 m = \frac{1}{4}\rho^2 ||v||^2$.

Now consider $w = \frac{2}{\rho} v$. The set N of entries of w of magnitude at most one is a subset of $S \cup M'$. By our assumption, $\sum_{w_i^2 \leqslant 1} w_i^2 \geqslant \rho^2 ||w||^2$. Hence

$$\sum_{i \in S} w_i^2 \geqslant \sum_{w_i^2 \leqslant 1} w_i^2 - \sum_{i \in M'} v_i^2 \geqslant \frac{3}{4}\rho^2 ||w||^2$$

Since this inequality is scale invariant, V is $(m, \sqrt{3/4} \cdot \rho)$-spread. □

Fact E.4. *[GLR10] Let $n \in \mathbb{N}$ and let V be a vector subspace of $\mathbb{R}^n$. Define $\Delta(V) = \sup_{\substack{v \in V \\ \|v\|=1}} \sqrt{n}/\|v\|_1$.*

Then

1. *If V is (m, ρ)-spread, then $\Delta(V) \leqslant \frac{1}{\rho^2}\sqrt{n/m}$.*
2. *V is $\left(\frac{n}{2\Delta(V)^2}, \frac{1}{4\Delta(V)}\right)$-spread.*

The lemma below relates $\|\cdot\|_1$ and $\|\cdot\|$ of vectors satisfying specific sparsity constraints.

Lemma E.5. *Let $m \in [n]$, $\mathcal{A} \subset [n]$, $\gamma_1 > 0$ and $\gamma_2 > 0$. Let $v \in \mathbb{R}^n$ be a vector such that*

$$\sum_{i \in \mathcal{A}} v_i^2 \leqslant \gamma_1^2 \|v\|^2 .$$

and for any set $\mathcal{M} \subset [n]$ of size m,

$$\sum_{i \in \mathcal{M}} v_i^2 \leqslant \gamma_2^2 \|v\|^2 .$$

Then

$$\sum_{i \in [n] \setminus \mathcal{A}} |v_i| \geqslant \frac{1 - \gamma_1^2 - \gamma_2^2}{\gamma_2} \sqrt{m}\, \|v\| .$$

Proof. Let $\mathcal{M}$ be the set of m largest coordinates of v (by absolute value). Since the inequality $\sum_{i\in[n]\setminus S}|v_i| \geqslant \frac{1-\gamma_1^2-\gamma_2^2}{\gamma_2}\sqrt{m}\,\|v\|$ is scale invariant, assume without loss of generality that for all $i \in M$, $|v_i| \geqslant 1$ and for all $i \in [n] \setminus \mathcal{M}$, $|v_i| \leqslant 1$. Then

$$\|v\|^2 \leqslant \sum_{i\in\mathcal{M}} v_i^2 + \sum_{i\in\mathcal{A}} v_i^2 + \sum_{i\in[n]\setminus(\mathcal{A}\cup\mathcal{M})} v_i^2 \leqslant \left(\gamma_2^2+\gamma_1^2\right)\|v\|^2 + \sum_{i\in[n]\setminus(\mathcal{A}\cup\mathcal{M})} v_i^2\,.$$

hence

$$\left(1-\gamma_2^2-\gamma_1^2\right)\|v\|^2 \leqslant \sum_{i\in[n]\setminus(\mathcal{A}\cup\mathcal{M})} v_i^2 \leqslant \sum_{i\in[n]\setminus(\mathcal{A}\cup\mathcal{M})} |v_i| \leqslant \sum_{i\in[n]\setminus\mathcal{A}} |v_i|\,.$$

Note that

$$\left(1-\gamma^2-\gamma_1^2\right)\|v\|^2 \geqslant \left(\frac{1-\gamma_2^2-\gamma_1^2}{\gamma_2^2}\right)\sum_{i\in\mathcal{M}} v_i^2 \geqslant \left(\frac{1-\gamma_2^2-\gamma_1^2}{\gamma_2^2}\right)m\,.$$

Therefore,

$$\left(\sum_{i\in[n]\setminus\mathcal{A}} |v_i|\right)^2 \geqslant \frac{\left(1-\gamma_2^2-\gamma_1^2\right)^2}{\gamma_2^2}\cdot m\|v\|^2\,.$$

□

Lemma E.6. *Suppose that vector $v \in \mathbb{R}^n$ satisfies the following property: for any $\mathcal{S} \subseteq [n]$,*

$$\sum_{i\in\mathcal{S}} |v_i| \leqslant |\mathcal{S}|\cdot\ln\left(\frac{en}{|\mathcal{S}|}\right). \tag{E.1}$$

Then

$$\|v\| \leqslant \sqrt{6n}\,.$$

Proof. Note that the set of vectors which satisfy Eq. (E.1) is compact. Hence there exists a vector v in this set with maximal $\|v\|$. Without loss of generality we can assume that the entries of v are nonnegative and sorted in descending order. Then for any $m \in [n]$,

$$\sum_{i=1}^{m} v_i \leqslant m\ln\left(\frac{en}{m}\right).$$

Assume that for some m the corresponding inequality is strict. Let's increase the last term v_m by small enough $\varepsilon > 0$. If there are no nonzero $v_{m'}$ for $m' > m$, all inequalities are still satisfied and $\|v\|$ becomes larger, which contradicts our choice of v. So there exists the smallest $m' > m$ such that $v_{m'} > 0$, and after decreasing $v_{m'}$ by $\varepsilon < v_{m'}$ all inequalities are still satisfied. $\|v\|$ increases after this operation:

$$(v_m+\varepsilon)^2 + (v_{m'}-\varepsilon)^2 = v_m^2 + v_{m'}^2 + 2\varepsilon(v_m - v_{m'}) + \varepsilon^2 > v_m^2 + v_{m'}^2\,.$$

Therefore, there are no strict inequalities. Hence $v_1 = \ln(en)$ and for all $m > 1$,

$$v_m = m \ln\Big(\frac{en}{m}\Big) - (m-1)\ln\Big(\frac{en}{m-1}\Big) = m\ln(1-1/m) + \ln\Big(\frac{en}{m-1}\Big) \leqslant \ln\Big(\frac{en}{m-1}\Big).$$

Since for any decreasing function $f : [1, n] \to \mathbb{R}$, $\sum_{j=2}^{n} f(j) \leqslant \int_1^n f(x)dx$,

$$\|v\|^2 \leqslant \ln^2(en) + \sum_{j=1}^{n-1} \ln^2\Big(\frac{en}{j}\Big) \leqslant 2\ln^2(en) + \int_1^n \ln^2\Big(\frac{en}{x}\Big)dx\,.$$

Note that

$$\int \ln^2\Big(\frac{en}{x}\Big)dx = 2x + 2x\ln\Big(\frac{en}{x}\Big) + x\ln^2\Big(\frac{en}{x}\Big).$$

Hence

$$\int_1^n \ln^2\Big(\frac{en}{x}\Big)dx = 5n - \ln^2(en) - 2\ln(en) - 2 \leqslant 5n - \ln^2(en)\,,$$

and we get the desired bound. □

Lemma E.7. *Suppose that vectors $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^n$ satisfy the following properties: for some $t_1 \geqslant 0$ and $t_2 \geqslant 0$, for any $\mathcal{S} \subseteq [n]$,*

$$\sum_{i\in\mathcal{S}} |v_i| \leqslant t_1 + |\mathcal{S}| \cdot \ln\Big(\frac{en}{|\mathcal{S}|}\Big). \tag{E.2}$$

and

$$\sum_{i\in\mathcal{S}} |w_i| \leqslant t_2 + |\mathcal{S}| \cdot \ln\Big(\frac{en}{|\mathcal{S}|}\Big). \tag{E.3}$$

Then

$$|\langle v, w\rangle| \leqslant t_1 t_2 + (t_1 + t_2)\ln(en) + 6n\,.$$

Proof. Note that the set of pairs of vectors which satisfy these properties is compact. Hence there exist vectors v, w that satisfy these properties such that $|\langle v, w\rangle|$ is maximal. Without loss of generality we can assume that the entries of v and w are nonnegative and sorted in descending order. Moreover, if some entry v_i of v is zero, we can increase $|\langle v, w\rangle|$ by assigning some small positive value to it without violating conditions on v (if $w_i = 0$, we can also assign some positive value to it without violating conditions on w). Hence we can assume that all entries of v and w are strictly positive.

Now assume that for some m the corresponding inequality for v with the set $[m]$ is strict. Let's increase v_m by small enough $\varepsilon > 0$ and decrease v_{m+1} by ε. This operation does not decrease $|\langle v, w\rangle|$:

$$(v_m + \varepsilon)w_m + (v_{m-1} - \varepsilon)w_{m-1} = v_m w_m + v_{m-1}w_{m-1} + \varepsilon(w_m - w_{m-1}) \geqslant 0 \tag{E.4}$$

Moreover, if $w_m = w_{m-1}$, the inequality for w with a set $[m]$ is strict, so by adding ε to w_m and subtracting ε from w_{m-1} we can make w_m and w_{m-1} different without violating

constraints on w and without decreasing $|\langle v, w\rangle|$. Hence without loss of generality we can assume that all v_i are different from each other and all w_i are different from each other. Now, by Eq. (E.4), there are no strict inequalities (otherwise there would be a contradiction). Hence $v_1 = t_1 + \ln(en)$, $w_1 = t_2 + \ln(en)$ and for all $m > 1$,

$$v_m = w_m = m \ln\left(\frac{en}{m}\right) - (m-1)\ln\left(\frac{en}{m-1}\right) = m\ln(1-1/m) + \ln\left(\frac{en}{m-1}\right) \leqslant \ln\left(\frac{en}{m-1}\right).$$

Since $v - t_1 e_1$ satisfies conditions of Lemma E.6,

$$|\langle v, w\rangle| \leqslant v_1 w_1 + \|v - t_1 e_1\|^2 - \ln^2(en) \leqslant t_1 t_2 + (t_1 + t_2)\ln(en) + 6n\,.$$

□