High Dimensional Robust Sparse Regression

Liu Liu **Yanyao Shen** **Tianyang Li** **Constantine Caramanis**
The University of Texas at Austin

Abstract

We provide a novel – and to the best of our knowledge, the first – algorithm for high dimensional sparse regression with constant fraction of corruptions in explanatory and/or response variables. Our algorithm recovers the true sparse parameters with sub-linear sample complexity, in the presence of a constant fraction of arbitrary corruptions. Our main contribution is a robust variant of Iterative Hard Thresholding. Using this, we provide accurate estimators: when the covariance matrix in sparse regression is identity, our error guarantee is near information-theoretically optimal. We then deal with robust sparse regression with unknown structured covariance matrix. We propose a filtering algorithm which consists of a novel randomized outlier removal technique for robust sparse mean estimation that may be of interest in its own right: the filtering algorithm is flexible enough to deal with unknown covariance. Also, it is orderwise more efficient computationally than the ellipsoid algorithm. Using sub-linear sample complexity, our algorithm achieves the best known (and first) error guarantee. We demonstrate the effectiveness on large-scale sparse regression problems with arbitrary corruptions.

1 Introduction

Learning in the presence of arbitrarily (even adversarially) corrupted outliers in the training data has a long history in Robust Statistics (Huber, 2011; Hampel et al., 2011; Tukey, 1975), and has recently received much renewed attention. The high dimensional setting poses particular challenges as outlier removal via preprocessing is essentially impossible when the number of variables scales with the number of samples. We propose a computationally efficient estimator for outlier-robust sparse regression that has near-optimal sample complexity, and is the first algorithm resilient to a constant fraction of arbitrary outliers with corrupted covariates and/or response variables. Unless we specifically mention otherwise, all future mentions of outliers mean corruptions in covariates and/or response variables.

We assume that the authentic samples are independent and identically distributed (i.i.d.) drawn from an uncorrupted distribution P, where P represents the linear model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \xi_i$, where $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, and $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is the true parameter (see Section 1.3 for complete details and definitions). To model the corruptions, the adversary can choose an arbitrary ϵ-fraction of the authentic samples, and replace them with arbitrary values. We refer to the observations after corruption as ϵ-corrupted samples (Definition 1.1). This corruption model allows the adversary to select an ϵ-fraction of authentic samples to delete and corrupt, hence it is stronger than Huber's ϵ-contamination model Huber (1964), where the adversary independently corrupts each sample with probability ϵ.

Outlier-robust regression is a classical problem within robust statistics (e.g., Rousseeuw and Leroy (2005)), yet even in the low-dimensional setting, efficient algorithms robust to corruption in the covariates have proved elusive, until recent breakthroughs in Prasad et al. (2018); Diakonikolas et al. (2019a) and Klivans et al. (2018), which built on important results in Robust Mean Estimation Diakonikolas et al. (2016); Lai et al. (2016) and Sums of Squares Barak and Steurer (2016), respectively.

In the sparse setting, the parameter β^* we seek to recover is also k-sparse, and a key goal is to provide recovery guarantees with sample complexity scaling with k, and *sublinearly* with d. Without outliers, by now classical results (e.g., Donoho (2006)) show that $n = \Omega(k \log d)$ samples from a i.i.d sub-Gaussian distribution are enough to give recovery guarantees on β^* with and without additive noise. These strong assump-

Proceedings of the 23$^{\mathrm{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

tions on the probabilistic distribution are necessary, since in the worst case, sparse recovery is known to be NP-hard Bandeira et al. (2013); Zhang et al. (2014).

Sparsity recovery with a constant fraction of arbitrary corruption is fundamentally hard. For instance, to the best of our knowledge, there's no previous work can provide exact recovery for sparse linear equations with arbitrary corruption in polynomial time. In contrast, a simple exhaustive search can easily enumerate the samples and recover the sparse parameter in exponential time.

In this work, we seek to give an efficient, sample-complexity optimal algorithm that recovers β^* to within accuracy depending on ϵ (the fraction of outliers). In the case of no additive noise, we are interested in algorithms that can guarantee exact recovery, independent of ϵ.

1.1 Related work

The last 10 years have seen a resurgence in interest in robust statistics, including the problem of resilience to outliers in the data. Important problems attacked have included PCA (Klivans et al., 2009; Xu et al., 2012, 2013; Lai et al., 2016; Diakonikolas et al., 2016), and more recently robust regression (as in this paper) (Prasad et al., 2018; Diakonikolas et al., 2019a,b; Klivans et al., 2018) and robust mean estimation (Diakonikolas et al., 2016; Lai et al., 2016; Balakrishnan et al., 2017a), among others. We focus now on the recent work most related to the present paper.

Robust regression. Earlier work in robust regression considers corruption only in the output, and shows that algorithms nearly as efficient as for regression without outliers succeeds in parameter recovery, even with a constant fraction of outliers (Li, 2013; Nguyen and Tran, 2013; Bhatia et al., 2015, 2017; Karmalkar and Price, 2018). Yet these algorithms (and their analysis) focus on corruption in y, and do not seem to extend to the setting of corrupted covariates – the setting of this work. In the low dimensional setting, there has been remarkable recent progress. The work in Klivans et al. (2018) shows that the Sum of Squares (SOS) based semidefinite hierarchy can be used for solving robust regression. Essentially concurrent to the SOS work, (Chen et al., 2017; Holland and Ikeda, 2017; Prasad et al., 2018; Diakonikolas et al., 2019a) use robust gradient descent for empirical risk minimization, by using robust mean estimation as a subroutine to compute robust gradients at each iteration. Diakonikolas et al. (2019b) uses filtering algorithm Diakonikolas et al. (2016) for robust regression. Computationally, these latter works scale better than the algorithms in Klivans et al. (2018), as although the Sum of Squares SDP framework gives polynomial time algorithms, they are often not practical (Hopkins et al., 2016).

Much less appears to be known in the high-dimensional regime. Exponential time algorithm, such as Gao (2017); Johnson and Preparata (1978), optimizes Tukey depth Tukey (1975); Chen et al. (2018). Their results reveal that handling a constant fraction of outliers ($\epsilon = const.$) is actually minimax-optimal. Work in Chen et al. (2013) first provided a polynomial time algorithm for this problem. They show that replacing the standard inner product in Matching Pursuit with a trimmed version, one can recover from an ϵ-fraction of outliers, with $\epsilon = O(1/\sqrt{k})$. Very recently, Liu et al. (2019) considered more general sparsity constrained M-estimation by using a trimmed estimator in each step of gradient descent, yet the robustness guarantee $\epsilon = O(1/\sqrt{k})$ is still sub-optimal. Another approach follows as a byproduct of a recent algorithm for robust sparse mean estimation, in Balakrishnan et al. (2017a). However, their error guarantee scales with $\|\boldsymbol{\beta}^*\|_2$, and moreover, does not provide *exact recovery* in the adversarial corruption case without stochastic noise (i.e., noise variance $\sigma^2 = 0$). We note that this is an inevitable consequence of their approach, as they directly use sparse mean estimation on $\{y_i \boldsymbol{x}_i\}$, rather than considering Maximum Likelihood Estimation.

Robust mean estimation. The idea in (Prasad et al., 2018; Diakonikolas et al., 2019a,b) is to leverage recent breakthroughs in robust mean estimation. Very recently, (Diakonikolas et al., 2016; Lai et al., 2016) provided the first robust mean estimation algorithms that can handle a constant fraction of outliers (though Lai et al. (2016) incurs a small (logarithmic) dependence in the dimension). Following their work, Balakrishnan et al. (2017a) extended the ellipsoid algorithm from Diakonikolas et al. (2016) to robust sparse mean estimation in high dimensions. They show that k-sparse mean estimation in $\mathbb{R}^d$ with a constant fraction of outliers can be done with $n = \Omega\left(k^2 \log\left(d\right)\right)$ samples. The k^2 term appears to be necessary, as $n = \Omega(k^2)$ follows from an oracle-based lower bound Diakonikolas et al. (2017).

1.2 Main contributions

- Our result is a robust variant of Iterative Hard Thresholding (IHT) Blumensath and Davies (2009). We provide a deterministic stability result showing that IHT works with any robust sparse mean estimation algorithm. We show our robust IHT does not accumulate the errors of a (any) robust sparse mean estimation subroutine for computing the gradient. Specifically, robust IHT produces a final solution whose error is orderwise the same as the error guaranteed by an single use of the robust mean estimation subroutine. We refer to Definition 2.1 and Theorem 2.1 for the precise statement. Thus our result can be viewed as a meta-theorem that can be coupled with any robust sparse mean estimator.

- Coupling robust IHT with a robust sparse mean estimation subroutine based on a version of the ellipsoid algorithm given and analyzed in Balakrishnan et al. (2017a), our results Corollary 3.1 show that given ϵ-corrupted sparse regression samples with identity covariance, we recover β^* within additive error $O(\sigma\epsilon)$ (which is minimax optimal Gao (2017)). The proof of the ellipsoid algorithm's performance in Balakrishnan et al. (2017a) hinges on obtaining an upper bound on the sparse operator norm (their Lemmas A.2 and A.3). As we show (see Appendix B), the statement of Lemma A.3 seems to be incorrect, and the general approach of upper bounding the sparse operator norm may not work. Nevertheless, the algorithm performance they claim is correct, as we show through a different avenue (see Lemma D.3 in Appendix D.3).

  Using this ellipsoid algorithm, In particular, we obtain exact recovery if either the fraction of outliers goes to zero (this is just ordinary sparse regression), or in the presence of a constant fraction of outliers but with the additive noise term going to zero (this is the case of robust sparse linear equations). To the best of our knowledge, this is the first result that shows exact recovery for robust sparse linear equations with a constant fraction of outliers. This is the content of Section 3.

- For robust sparse regression with *unknown covariance matrix*, we consider the wide class of sparse covariance matrices Bickel and Levina (2008). We then prove a result that may be of interest in its own right: we provide a novel robust sparse mean estimation algorithm that is based on a filtering algorithm for sequentially screening and removing potential outliers. We show that the filtering algorithm is flexible enough to deal with unknown covariance, whereas the ellipsoid algorithm cannot. It also runs a factor of $O(d^2)$ faster than the ellipsoid algorithm. If the covariance matrix is sufficiently sparse, our filtering algorithm gives a robust sparse mean estimation algorithm, that can then be coupled with our meta-theorem. Together, these two guarantee recovery of β^* within an additive error of $O(\sigma\sqrt{\epsilon})$. In the case of unknown covariance, this is the best (and in fact, only) result we are aware of for robust sparse regression. We note that it can be applied to the case of known and identity covariance, though it is weaker than the optimal results we obtain using the computationally more expensive ellipsoid algorithm. Nevertheless, in both cases (unknown sparse, or known identity) the result is strong enough to guarantee exact recovery when either σ or ϵ goes to zero. We demonstrate the practical effectiveness of our filtering algorithm in Section 6. This is the content of Section 4 and Section 5.

1.3 Setup, Notation and Outline

In this subsection, we formally define the corruption model and the sparse regression model. We first introduce the ϵ-corrupted samples described above:

Definition 1.1 (ϵ-corrupted samples)**.** *Let $\{\boldsymbol{z}_i, i \in \mathcal{G}\}$ be i.i.d. observations follow from a distribution P. The ϵ-corrupted samples $\{\boldsymbol{z}_i, i \in \mathcal{S}\}$ are generated by the following process: an adversary chooses an arbitrary ϵ-fraction of the samples in $\mathcal{G}$ and modifies them with arbitrary values. After the corruption, we use $\mathcal{S}$ to denote the observations, and use $\mathcal{B} = \mathcal{S} \setminus \mathcal{G}$ to denote the corruptions.*

The parameter ϵ represents the fraction of outliers. Throughout, we assume that it is a (small) constant, *independent of dimension or other problem parameters.* Furthermore, we assume that the distribution P is the standard Gaussian-design AWGN linear model.

Model 1.1. *The observations $\{\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i), i \in \mathcal{G}\}$ follow from the linear model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \xi_i$, where $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is the model parameter, and assumed to be k-sparse. We assume that $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ and $\xi_i \sim \mathcal{N}\left(0, \sigma^2\right)$, where $\boldsymbol{\Sigma}$ is the normalized covariance matrix with $\boldsymbol{\Sigma}_{jj} \leq 1$ for all $j \in [d]$. We denote μ_α as the smallest eigenvalue of $\boldsymbol{\Sigma}$, and μ_β as its largest eigenvalue. They are assumed to be universal constants in this paper, and we denote the constant $c_\kappa = \mu_\beta/\mu_\alpha$.*

As in Balakrishnan et al. (2017a), we pre-process by removing "obvious" outliers; we henceforth assume that all authentic and corrupted points are within a radius bounded by a polynomial in n, d and $1/\epsilon$.

Notation. We denote the hard thresholding operator of sparsity k' by $\mathsf{P}_{k'}$. We define the k-sparse operator norm as $\|M\|_{\widetilde{k},\mathrm{op}} = \max_{\|\boldsymbol{v}\|_2=1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} |\boldsymbol{v}^\top M \boldsymbol{v}|$, where M is not required to be positive-semidefinite (p.s.d.). We use trace inner produce $\langle A, B\rangle$ to denote $\mathrm{Tr}\left(A^\top B\right)$. We use $\mathbb{E}_{i \in_u \mathcal{S}}$ to denote the expectation operator obtained by the uniform distribution over all samples i in a set $\mathcal{S}$. Finally, we use the notation $\widetilde{O}(\cdot)$ to hide the dependency on $\mathrm{poly}\log(1/\epsilon)$, and $\widetilde{\Omega}(\cdot)$ to hide the dependency on $\mathrm{poly}\log(k)$ in our bounds.

2 Hard thresholding with robust gradient estimation

In this section, we present our method of using robust sparse gradient updates in IHT. We then show sta-

Algorithm 1 Robust sparse regression with RSGE

```
1:  Input: Samples $\{y_i, \boldsymbol{x}_i\}_{i=1}^N$, RSGE subroutine.
2:  Output: The estimation $\widehat{\boldsymbol{\beta}}$.
3:  Parameters: Hard thresholding parameter $k'$.
4:  Split samples into $T$ subsets of size $n$. $\boldsymbol{\beta}^0 = \mathbf{0}$.
5:  for $t = 0$ to $T-1$, do
6:      At current $\boldsymbol{\beta}^t$, calculate all gradients for current $n$ samples: $\boldsymbol{g}_i^t = \boldsymbol{x}_i\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}^t - y_i\right)$, $i \in [n]$.
7:      For $\{\boldsymbol{g}_i^t\}_{i=1}^n$, We use a RSGE to get $\widehat{\boldsymbol{G}}^t$.
8:      Update the parameter: $\boldsymbol{\beta}^{t+1} = \mathsf{P}_{k'}\left(\boldsymbol{\beta}^t - \eta \widehat{\boldsymbol{G}}^t\right)$.
9:  end for
10: Output the estimation $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^T$.
```

tistical recovery guarantees given any accurate robust sparse gradient estimation, which is formally defined in Definition 2.1.

We define the notation for the stochastic gradient $\boldsymbol{g}_i$ corresponding to the i^{th} point $\boldsymbol{z}_i$, and the population gradient for $\boldsymbol{z}_i \sim P$ based on Model 1.1, $\boldsymbol{g}_i^t = \boldsymbol{x}_i\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}^t - y_i\right)$, and $\boldsymbol{G}^t = \mathbb{E}_{\boldsymbol{z}_i \sim P}\left(\boldsymbol{g}_i^t\right)$, where P is the distribution of the authentic points. Since $\mathbb{E}_{\boldsymbol{z}_i \sim P}\left(\boldsymbol{x}_i \boldsymbol{x}_i^\top\right) = \boldsymbol{\Sigma}$, the population mean of all authentic gradients is given by $\boldsymbol{G}^t = \mathbb{E}_{\boldsymbol{z}_i \sim P}\left(\boldsymbol{x}_i \boldsymbol{x}_i^\top\left(\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\right)\right) = \boldsymbol{\Sigma}(\boldsymbol{\beta}^t - \boldsymbol{\beta}^*)$.

In the *uncorrupted case* where all samples $\{\boldsymbol{z}_i, i \in \mathcal{G}\}$ follow from Model 1.1, a single iteration of IHT updates β^t via $\boldsymbol{\beta}^{t+1} = \mathsf{P}_{k'}(\boldsymbol{\beta}^t - \mathbb{E}_{i \in_u \mathcal{G}}\, \boldsymbol{g}_i^t)$. Here, the hard thresholding operator $\mathsf{P}_{k'}$ selects the k' largest elements in magnitude, and the parameter k' is proportional to k (specified in Theorem 2.1). However, given ϵ-corrupted samples $\{\boldsymbol{z}_i, i \in \mathcal{S}\}$ according to Definition 1.1, the IHT update based on empirical average of all gradient samples $\{\boldsymbol{g}_i, i \in \mathcal{S}\}$ can be arbitrarily bad.

The key goal in this paper is to find a robust estimate $\widehat{\boldsymbol{G}}^t$ to replace $\boldsymbol{G}^t$ in each step of IHT, with sample complexity *sub-linear* in the dimension d. For instance, we consider robust sparse regression with $\boldsymbol{\Sigma} = \boldsymbol{I}_d$. Then, $\boldsymbol{G}^t = \boldsymbol{\beta}^t - \boldsymbol{\beta}^*$ is guaranteed to be $(k'+k)$-sparse in each iteration of IHT. In this case, given ϵ-corrupted samples, we can use a robust sparse mean estimator to recover the unknown true $\boldsymbol{G}^t$ from $\{\boldsymbol{g}_i^t\}_{i=1}^{|\mathcal{S}|}$, with sub-linear sample complexity.

More generally, we propose Robust Sparse Gradient Estimator (RSGE) for gradient estimation given ϵ-corrupted samples, as defined in Definition 2.1, which guarantees that the deviation between the robust estimate $\widehat{\boldsymbol{G}}(\boldsymbol{\beta})$ and true $\boldsymbol{G}(\boldsymbol{\beta})$, with sample complexity $n \ll d$. For a fixed k-sparse parameter $\boldsymbol{\beta}$, we drop the superscript t without abuse of notation, and use $\boldsymbol{g}_i$ in place of $\boldsymbol{g}_i^t$, and $\boldsymbol{G}$ in place of $\boldsymbol{G}^t$; $\boldsymbol{G}(\boldsymbol{\beta})$ denotes the population gradient over the authentic samples' distribution P, at the point $\boldsymbol{\beta}$.

Definition 2.1 ($\psi(\epsilon)$-RSGE)**.** *Given $n(k,d,\epsilon,\nu)$ ϵ-corrupted samples $\{\boldsymbol{z}_i\}_{i=1}^n$ from Model 1.1, we call $\widehat{\boldsymbol{G}}(\boldsymbol{\beta})$ a $\psi(\epsilon)$-RSGE, if given $\{\boldsymbol{z}_i\}_{i=1}^n$, $\widehat{\boldsymbol{G}}(\boldsymbol{\beta})$ guarantees $\|\widehat{\boldsymbol{G}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta})\|_2^2 \leq \alpha(\epsilon)\|\boldsymbol{G}(\boldsymbol{\beta})\|_2^2 + \psi(\epsilon)$, with probability at least $1-\nu$.*

Here, we use $n(k,d,\epsilon,\nu)$ to denote the sample complexity as a function of (k,d,ϵ,ν), and note that the definition of RSGE does not require $\boldsymbol{\Sigma}$ to be identity matrix. The parameters $\alpha(\epsilon)$ and $\psi(\epsilon)$ will be specified by concrete robust sparse mean estimators in subsequent sections. Equipped with Definition 2.1, we propose Algorithm 1, which takes any RSGE as a subroutine in line 7, and runs a robust variant of IHT with the estimated sparse gradient $\widehat{\boldsymbol{G}}^t$ at each iteration in line 8.[1]

Global linear convergence and parameter recovery guarantees. In each single IHT update step, RSGE introduces a controlled amount of error. Theorem 2.1 gives a global linear convergence guarantee for Algorithm 1 by showing that IHT does not accumulate too much error. In particular, we are able to recover $\boldsymbol{\beta}^*$ within error $O(\sqrt{\psi(\epsilon)})$ given any $\psi(\epsilon)$-RSGE subroutine. We give the proof of Theorem 2.1 in Appendix A. The hyper-parameter $k' = c_\kappa^2 k$ guarantees global linear convergence of IHT when $c_\kappa > 1$ (when $\boldsymbol{\Sigma} \neq \boldsymbol{I}_d$). This setup has been used in Jain et al. (2014); Shen and Li (2017), and is proved to be necessary in Liu and Barber (2018). Note that Theorem 2.1 is a deterministic stability result in nature, and we obtain probabilistic results by certifying the RSGE condition.

Theorem 2.1 (Meta-theorem)**.** *Suppose we observe $N(k,d,\epsilon,\nu)$ ϵ-corrupted samples from Model 1.1. Algorithm 1, with $\psi(\epsilon)$-RSGE defined in Definition 2.1, with step size $\eta = 1/\mu_\beta$ outputs $\widehat{\boldsymbol{\beta}}$, such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O(\sqrt{\psi(\epsilon)})$, with probability at least $1-\nu$, by setting $k' = c_\kappa^2 k$ and $T = \Theta(\log(\|\boldsymbol{\beta}^*\|_2/\sqrt{\psi(\epsilon)}))$. The sample complexity is $N(k,d,\epsilon,\nu) = n(k,d,\epsilon,\nu/T)\,T$.*

3 Robust sparse regression with near-optimal guarantee

In this section, we provide near optimal statistical guarantee for robust sparse regression when the covariance matrix is identity. Under the assumption $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, Balakrishnan et al. (2017a) proposes a robust sparse regression estimator based on robust sparse

[1] Our results require sample splitting to maintain independence between subsequent iterations, though we believe this is an artifact of our analysis. Similar approach has been used in Balakrishnan et al. (2017b); Prasad et al. (2018) for theoretical analysis. We do not use sample splitting technique in the experiments.

Algorithm 2 Separation oracle for robust sparse estimation Balakrishnan et al. (2017a)

```
1: Input: Weights from the previous iteration $\{w_i, i \in \mathcal{S}\}$, gradient samples $\{\boldsymbol{g}_i, i \in \mathcal{S}\}$.
2: Output: Weight $\{w_i', i \in \mathcal{S}\}$
3: Parameters: Hard thresholding parameter $\widetilde{k}$, parameter $\rho_{\mathrm{sep}}$.
4: Compute the weighted sample mean $\widetilde{\boldsymbol{G}} = \sum_{i \in \mathcal{S}} w_i \boldsymbol{g}_i$, and $\widehat{\boldsymbol{G}} = \mathsf{P}_{2\widetilde{k}}(\widetilde{\boldsymbol{G}})$.
5: Compute the weighted sample covariance matrix $\widehat{\boldsymbol{\Sigma}} = \sum_{i \in \mathcal{S}} w_i \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)^\top$.
6: Let $\lambda^*$ be the optimal value, and $\boldsymbol{H}^*$ be the corresponding solution of the following program
       $\max_{\boldsymbol{H}} \operatorname{Tr}\left(\left(\widehat{\boldsymbol{\Sigma}} - F\left(\widehat{\boldsymbol{G}}\right)\right) \cdot \boldsymbol{H}\right)$,
       subject to $\boldsymbol{H} \succcurlyeq 0, \|\boldsymbol{H}\|_{1,1} \le \widetilde{k}, \operatorname{Tr}(\boldsymbol{H}) = 1$.
7: if $\lambda^* \le \rho_{\mathrm{sep}}$, then return "Yes".
8: return The hyperplane: $\ell(w') = \left\langle \left( \sum_{i \in \mathcal{S}} w_i' (\boldsymbol{g}_i - \widehat{\boldsymbol{G}})(\boldsymbol{g}_i - \widehat{\boldsymbol{G}})^\top - F(\widehat{\boldsymbol{G}}) \right), \boldsymbol{H}^* \right\rangle - \lambda^*$.
```

mean estimation on $\{y_i \boldsymbol{x}_i, i \in \mathcal{S}\}$, leveraging the fact that $\mathbb{E}_{\boldsymbol{z}_i \sim P}(y_i \boldsymbol{x}_i) = \boldsymbol{\beta}^*$. With sample complexity $N = \Omega\left(\frac{k^2 \log(d/\nu)}{\epsilon^2}\right)$, this algorithm produces a $\widetilde{\boldsymbol{\beta}}$ such that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 = \widetilde{O}(\epsilon^2(\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2))$, with probability at least $1 - \nu$. Using Theorem 2.1, we show that we can obtain significantly stronger statistical guarantees which are statistically optimal; in particular, our guarantees are independent of $\|\boldsymbol{\beta}^*\|_2$ and yield exact recovery when $\sigma = 0$.

3.1 RSGE via the ellipsoid algorithm

More specifically, the ellipsoid-based robust sparse mean estimation algorithm Balakrishnan et al. (2017a) deals with outliers by trying to optimize the set of weights $\{w_i, i \in \mathcal{S}\}$ on each of the samples in $\mathbb{R}^d$ – ideally outliers would receive lower weight and hence their impact would be minimized. Since the set of weights is convex, this can be approached using a separation oracle Algorithm 2. The Algorithm 2 depends on a convex relaxation of Sparse PCA, and the hard thresholding parameter is $\widetilde{k} = k' + k$, as the population mean of all authentic gradient samples $\boldsymbol{G}^t$ is guaranteed to be $(k' + k)$-sparse. In line 4 and 5, we calculate the weighted mean and covariance based on a hard thresholding operator. In line 6 of Algorithm 2, with each call to the relaxation of Sparse PCA, we obtain an optimal value, λ^*, and optimal solution, $\boldsymbol{H}^*$, to the problem:

$$\lambda^* = \max_{\boldsymbol{H}} \operatorname{Tr}\left(\left(\widehat{\boldsymbol{\Sigma}} - F\left(\widehat{\boldsymbol{G}}\right)\right) \cdot \boldsymbol{H}\right),$$
$$\text{subject to } \boldsymbol{H} \succcurlyeq 0, \|\boldsymbol{H}\|_{1,1} \le \widetilde{k}, \operatorname{Tr}(\boldsymbol{H}) = 1. \tag{1}$$

Here, $\widehat{\boldsymbol{G}}$, $\widehat{\boldsymbol{\Sigma}}$ are weighted first and second order moment estimates from ϵ-corrupted samples, and $F : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ is a function with closed-form $F(\widehat{\boldsymbol{G}}) = \|\widehat{\boldsymbol{G}}\|_2^2 \boldsymbol{I}_d + \widehat{\boldsymbol{G}}\widehat{\boldsymbol{G}}^\top + \sigma^2 \boldsymbol{I}_d$. Given the population mean $\boldsymbol{G}$, we have $F(\boldsymbol{G}) = \mathbb{E}_{\boldsymbol{z}_i \sim P}((\boldsymbol{g}_i - \boldsymbol{G})(\boldsymbol{g}_i - \boldsymbol{G})^\top)$, which calculates the underlying true covariance matrix. We provide more details about the calculation of $F(\cdot)$, as well as some smoothness properties, in Appendix C.

The key component in the separation oracle Algorithm 2 is to use convex relaxation of Sparse PCA eq. (1). This idea generalizes existing work on using PCA to detect outliers in low dimensional robust mean estimation Diakonikolas et al. (2016); Lai et al. (2016). To gain some intuition for eq. (1), if $\boldsymbol{g}_i$ is an outlier, then the optimal solution of eq. (1), $\boldsymbol{H}^*$, may detect the direction of this outlier. And this outlier will be down-weighted in the output of Algorithm 2 by the separating hyperplane. Finally, Algorithm 2 will terminate with $\lambda^* \le \rho_{\mathrm{sep}}$ (line 7) and output the robust sparse mean estimation of the gradients $\widehat{\boldsymbol{G}}$.

Indeed, the ellipsoid-algorithm-based robust sparse mean estimator gives a RSGE, which we can combine with Theorem 2.1 to obtain stronger results. We state these as Corollary 3.1. We note again that the analysis in Balakrishnan et al. (2017a) has a flaw. Their Lemma A.3 is incorrect, as our counterexample in Appendix B demonstrates. We provide a correct route of analysis in Lemma D.3 of Appendix D.

3.2 Near-optimal statistical guarantees

Corollary 3.1. *Suppose we observe $N(k, d, \epsilon, \nu)$ ϵ-corrupted samples from Model 1.1 with $\boldsymbol{\Sigma} = \boldsymbol{I}_d$. By setting $\widetilde{k} = k' + k$, if we use the ellipsoid algorithm for robust sparse gradient estimation with $\rho_{\mathrm{sep}} = \Theta\left(\epsilon\left(\|\boldsymbol{G}^t\|_2^2 + \sigma^2\right)\right)$, it requires $N(k, d, \epsilon, \nu) = \Omega\left(\frac{k^2 \log(dT/\nu)}{\epsilon^2}\right)T$ samples, and guarantees $\psi(\epsilon) = \widetilde{O}\left(\epsilon^2 \sigma^2\right)$. Hence, Algorithm 1 outputs $\widehat{\boldsymbol{\beta}}$, such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \widetilde{O}(\sigma\epsilon)$, with probability at least $1 - \nu$, by setting $T = \Theta\left(\log\left(\frac{\|\boldsymbol{\beta}^*\|_2}{\epsilon\sigma}\right)\right)$.*

Any one-dimensional robust variance estimate of the regression residual $\{r_i, i \in [n]\}$ in each iteration is sufficient for ρ_{sep}, where $r_i = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^t$. For a desired error level $\epsilon' \ge \epsilon$, we only require sample complexity $N(k, d, \epsilon, \nu) = \Omega\left(\frac{k^2 \log(dT/\nu)}{\epsilon'^2}\right)T$. Hence, we can achieve statistical error $\widetilde{O}\left(\sigma\left(\sqrt{k^2 \log(d)/N} \vee \epsilon\right)\right)$. Our error bound is nearly optimal compared to the information-theoretically optimal $O\left(\sigma\left(\sqrt{k \log(d)/N} \vee \epsilon\right)\right)$ in Gao (2017), as the k^2 term is necessary by an oracle-based SQ lower bound Diakonikolas et al. (2017).

Proof sketch of Corollary 3.1 The key to the proof relies on showing that λ^* controls the quality of the weights of the current iteration, i.e., small λ^* means good weights and thus a good current solution. Showing this relies on using λ^* to control $\widehat{\boldsymbol{\Sigma}} - F(\widehat{\boldsymbol{G}})$. Lemma A.3 in Balakrishnan et al. (2017a) claims that $\lambda^* \geq \|\widehat{\boldsymbol{\Sigma}} - F(\widehat{\boldsymbol{G}})\|_{\widetilde{\mathrm{k}},\mathrm{op}}$. As we show in Appendix B, however, this need not hold. This is because the trace norm maximization eq. (1) is *not* a valid convex relaxation for the $\widetilde{k}$-sparse operator norm when the term $\widehat{\boldsymbol{\Sigma}} - F(\widehat{\boldsymbol{G}})$ is not p.s.d. (which indeed it need not be). We provide a different line of analysis in Lemma D.3, essentially showing that even without the claimed (incorrect) bound, λ^* can still provide the control we need. With the corrected analysis for λ^*, the ellipsoid algorithm guarantees $\|\widehat{\boldsymbol{G}} - \boldsymbol{G}\|_2^2 = \widetilde{O}(\epsilon^2(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \sigma^2))$ with probability at least $1 - \nu$. Therefore, the algorithm provides an $\widetilde{O}\left(\epsilon^2\sigma^2\right)$-RSGE.

4 Robust sparse mean estimation via filtering

From a computational viewpoint, the time complexity of Algorithm 1 depends on the RSGE in each iterate. The time complexity of the ellipsoid algorithm is indeed polynomial in the dimension, but it requires $O\left(d^2\right)$ calls to a relaxation of Sparse PCA (Bubeck (2015)). In this section, we introduce a faster algorithm as a RSGE, which only requires $O\left(n\right)$ calls of Sparse PCA (recall that n only scales with $k^2 \log d$). Importantly, this RSGE is flexible enough to deal with unknown covariance matrix, yet the ellipsoid algorithm cannot. Before we move to the result for unknown covariance matrix in Section 5, we first introduce Algorithm 3 and analyze its performance when the covariance is identity. These supporting Lemmas will be later used in the unknown case.

Our proposed RSGE (Algorithm 3) attempts to remove one outlier at each iteration, as long as a good solution has not already been identified. It first estimates the gradient $\widehat{\boldsymbol{G}}$ by hard thresholding (line 4) and then estimates the corresponding sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ (line 5). By solving (a relaxation of) Sparse PCA, we obtain a scalar λ^* as well as a matrix $\boldsymbol{H}^*$. If λ^* is smaller than the predetermined threshold ρ_{sep}, we have a certificate that the effect of the outliers is well-controlled (specified in eq. (5)). Otherwise, we compute a score for each sample based on $\boldsymbol{H}^*$, and discard one of the samples according to a probability distribution where each sample's probability of being discarded is proportional to the score we have computed. [2]

Algorithm 3 RSGE via filtering

```
1: Input: A set $\mathcal{S}_{\mathrm{in}}$.
2: Output: A set $\mathcal{S}_{\mathrm{out}}$ or sparse mean vector $\widehat{\boldsymbol{G}}$.
3: Parameters: Hard thresholding parameter $\widetilde{k}$, parameter $\rho_{\mathrm{sep}}$.
4: Compute the sample mean $\widetilde{\boldsymbol{G}} = \mathbb{E}_{i \in_u \mathcal{S}_{\mathrm{in}}}\left(\boldsymbol{g}_i\right)$, and $\widehat{\boldsymbol{G}} = \mathsf{P}_{2\widetilde{k}}(\widetilde{\boldsymbol{G}})$.
5: Compute the sample covariance matrix
       $$\widehat{\boldsymbol{\Sigma}} = \mathbb{E}_{i \in_u \mathcal{S}_{\mathrm{in}}}\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)^\top.$$
6: Solve the following convex program:
       $$\max_{\boldsymbol{H}} \mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}} \cdot \boldsymbol{H}\right), \quad \text{subject to } \boldsymbol{H} \succcurlyeq 0, \|\boldsymbol{H}\|_{1,1} \leq \widetilde{k}, \mathrm{Tr}\left(\boldsymbol{H}\right) = 1. \quad (2)$$
   Let $\lambda^*$ be the optimal value, and $\boldsymbol{H}^*$ be the corresponding solution.
7: if $\lambda^* \leq \rho_{\mathrm{sep}}$ , then return with $\widehat{\boldsymbol{G}}$.
8: Calculate projection score for each $i \in \mathcal{S}_{\mathrm{in}}$:
       $$\tau_i = \mathrm{Tr}\left(\boldsymbol{H}^* \cdot \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)^\top\right). \quad (3)$$
9: Randomly remove a sample $r$ from $\mathcal{S}_{\mathrm{in}}$ according to
       $$\Pr\left(\boldsymbol{g}_i \text{ is removed}\right) = \frac{\tau_i}{\sum_{i \in \mathcal{S}_{\mathrm{in}}} \tau_i}. \quad (4)$$
10: return the set $\mathcal{S}_{\mathrm{out}} = \mathcal{S}_{\mathrm{in}} \setminus \{r\}$.
```

Algorithm 3 can be used for other robust sparse functional estimation problems (e.g., robust sparse mean estimation for $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{I_d}\right)$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ is k-sparse). To use Algorithm 3 as a RSGE given n gradient samples (denoted as $\mathcal{S}_{\mathrm{in}}$), we call Algorithm 3 repeatedly on $\mathcal{S}_{\mathrm{in}}$ and then on its output, $\mathcal{S}_{\mathrm{out}}$, until it returns a robust estimator $\widehat{\boldsymbol{G}}$. The next theorem provides guarantees on this iterative application of Algorithm 3.

Theorem 4.1. *Suppose we observe $n = \Omega\left(\frac{k^2 \log(d/\nu)}{\epsilon}\right)$ ϵ-corrupted samples from Model 1.1 with $\boldsymbol{\Sigma} = \boldsymbol{I}_d$. Let $\mathcal{S}_{\mathrm{in}}$ be an ϵ-corrupted set of gradient samples $\{\boldsymbol{g}_i^t\}_{i=1}^n$. By setting $\widetilde{k} = k' + k$, if we run Algorithm 3 iteratively with initial set $\mathcal{S}_{\mathrm{in}}$, and subsequently on $\mathcal{S}_{\mathrm{out}}$, and use $\rho_{\mathrm{sep}} = C_\gamma\left(\|\boldsymbol{G}^t\|_2^2 + \sigma^2\right)$, then this repeated use of Algorithm 3 will stop after at most $\frac{1.1\gamma}{\gamma-1}\epsilon n$ iterations, and output $\widehat{\boldsymbol{G}}^t$, such that $\|\widehat{\boldsymbol{G}}^t - \boldsymbol{G}^t\|_2^2 = \widetilde{O}\left(\epsilon\left(\|\boldsymbol{G}^t\|_2^2 + \sigma^2\right)\right)$, with probability at least $1 - \nu - \exp\left(-\Theta\left(\epsilon n\right)\right)$. Here, C_γ is a constant depending on γ, where $\gamma \geq 4$ is a constant.*

[2] Although we remove one sample in Algorithm 3, our theoretical analysis naturally extend to removing constant number of outliers. This speeds up the algorithm in practice, yet shares the same computational complexity

Thus, Theorem 4.1 shows that with high probability, Algorithm 3 provides a Robust Sparse Gradient Estimator where $\psi(\epsilon) = \widetilde{O}\left(\epsilon\sigma^2\right)$. For example, we can take $\nu = d^{-\Theta(1)}$. Combining now with Theorem 2.1, we obtain an error guarantee for robust sparse regression.

Corollary 4.1. *Suppose we observe $N(k, d, \epsilon, \nu)$ ϵ-corrupted samples from Model 1.1 with $\boldsymbol{\Sigma} = \boldsymbol{I}_d$. Under the same setting as Theorem 4.1, if we use Algorithm 3 for robust sparse gradient estimation, it requires $N(k, d, \epsilon, \nu) = \Omega\left(\frac{k^2 \log(dT/\nu)}{\epsilon}\right) T$ samples, and $T = \Theta\left(\log\left(\frac{\|\boldsymbol{\beta}^*\|_2}{\sigma\sqrt{\epsilon}}\right)\right)$, then we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \widetilde{O}\left(\sigma\sqrt{\epsilon}\right)$ with probability at least $1 - \nu - T\exp\left(-\Theta\left(\epsilon n\right)\right)$.*

Similar to Section 3, we can achieve statistical error $\widetilde{O}\big(\sigma\big(\sqrt{k^2\log(d)/N} \vee \sqrt{\epsilon}\big)\big)$. The scaling of ϵ in Corollary 4.1 is $\widetilde{O}\left(\sqrt{\epsilon}\right)$. These guarantees are worse than $\widetilde{O}\left(\epsilon\right)$ achieved by ellipsoid methods. Nevertheless, this result is strong enough to guarantee exact recovery when either σ or ϵ goes to zero. The simulation of robust estimation for the filtering algorithm is in Section 6.

The key step in Algorithm 3 is outlier removal eq. (4) based on the solution of Sparse PCA's convex relaxation eq. (2). We describe the outlier removal below, and then give the proofs in Appendix E and Appendix F.

Outlier removal guarantees in Algorithm 3. We denote samples in the input set $\mathcal{S}_{\text{in}}$ as $\boldsymbol{g}_i$. This input set $\mathcal{S}_{\text{in}}$ can be partitioned into two parts: $\mathcal{S}_{\text{good}} = \{i : i \in \mathcal{G} \text{ and } i \in \mathcal{S}_{\text{in}}\}$, and $\mathcal{S}_{\text{bad}} = \{i : i \in \mathcal{B} \text{ and } i \in \mathcal{S}_{\text{in}}\}$. Lemma 4.1 shows that Algorithm 3 can return a guaranteed gradient estimate, or the outlier removal step eq. (4) is likely to discard an outlier. The guarantee on the outlier removal step eq. (4) hinges on the fact that if $\sum_{i\in\mathcal{S}_{\text{good}}} \tau_i$ is less than $\sum_{i\in\mathcal{S}_{\text{bad}}} \tau_i$, we can show eq. (4) is likely to remove an outlier.

Lemma 4.1. *Suppose we observe $n = \Omega\big(\frac{k^2\log(d/\nu)}{\epsilon}\big)$ ϵ-corrupted samples from Model 1.1 with $\boldsymbol{\Sigma} = \boldsymbol{I}_d$. Let $\mathcal{S}_{\text{in}}$ be an ϵ-corrupted set $\{\boldsymbol{g}_i^t\}_{i=1}^n$. Algorithm 3 computes λ^* that satisfies*

$$\lambda^* \geq \max_{\|\boldsymbol{v}\|_2=1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} \boldsymbol{v}^\top \left(\mathbb{E}_{i\in_u \mathcal{S}_{\text{in}}}\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)^\top\right)\boldsymbol{v}. \tag{5}$$

If $\lambda^* \geq \rho_{\text{sep}} = C_\gamma\left(\|\boldsymbol{G}^t\|_2^2 + \sigma^2\right)$, then with probability at least $1-\nu$, we have $\sum_{i\in\mathcal{S}_{\text{good}}} \tau_i \leq \frac{1}{\gamma}\sum_{i\in\mathcal{S}_{\text{in}}} \tau_i$, where τ_i is defined in eq. (3), C_γ is a constant depending on γ, and $\gamma \geq 4$ is a constant.

The proofs are collected in Appendix E. In a nutshell, eq. (5) is a natural convex relaxation for the sparsity constraint $\{\boldsymbol{v} : \|\boldsymbol{v}\|_2 = 1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}\}$. On the other hand, when $\lambda^* \geq \rho_{\text{sep}}$, the contribution of $\sum_{i\in\mathcal{S}_{\text{good}}} \tau_i$ is relatively small, which can be obtained through concentration inequalities for the samples in $\mathcal{S}_{\text{good}}$. Based on Lemma 4.1, if $\lambda^* \leq \rho_{\text{sep}}$, then the RHS of eq. (5) is bounded, leading to the error guarantee of $\|\widehat{\boldsymbol{G}}^t - \boldsymbol{G}^t\|_2^2$. On the other hand, if $\lambda^* \geq \rho_{\text{sep}}$, we can show that eq. (4) is more likely to throw out samples of $\mathcal{S}_{\text{bad}}$ rather than $\mathcal{S}_{\text{good}}$. Iteratively applying Algorithm 3 on the remaining samples, we can remove those outliers with large effect, and keep the remaining outliers' effect well-controlled. This leads to the final bounds in Theorem 4.1.

5 Robust sparse regression with unknown covariance

In this section, we consider robust sparse regression with unknown covariance matrix $\boldsymbol{\Sigma}$, which has additional sparsity structure. Formally, we define the sparse covariance matrices as follows:

Model 5.1 (Sparse covariance matrices)**.** *In Model 1.1, the authentic covariates $\{\boldsymbol{x}_i, i \in \mathcal{G}\}$ are drawn from $\mathcal{N}(0, \boldsymbol{\Sigma})$. We assume that each row and column of $\boldsymbol{\Sigma}$ is r-sparse, but the positions of the non-zero entries are unknown.*

Model 5.1 is widely studied in high dimensional statistics Bickel and Levina (2008); El Karoui (2008); Wainwright (2019). Under Model 5.1, for the population gradient $\boldsymbol{G}^t = \mathbb{E}_P\left(\boldsymbol{x}_i\boldsymbol{x}_i^\top\left(\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\right)\right) = \boldsymbol{\Sigma}\omega^t$, where we use ω^t to denote the $(k'+k)$-sparse vector $\boldsymbol{\beta}^t - \boldsymbol{\beta}^*$, we can guarantee the $\|\boldsymbol{G}^t\|_0 = \|\boldsymbol{\Sigma}\omega^t\|_0 \leq r(k'+k)$. Hence, we can use the filtering algorithm (Algorithm 3) with $\widetilde{k} = r(k'+k)$ as a RSGE for robust sparse regression with unknown $\boldsymbol{\Sigma}$. When the covariance is unknown, we cannot evaluate $F(\cdot)$ a priori, thus the ellipsoid algorithm is not applicable to this case. And we provide error guarantees as follows.

Theorem 5.1. *Suppose we observe $N(k, d, \epsilon, \nu)$ ϵ-corrupted samples from Model 1.1, where the covariates $\boldsymbol{x}_i$'s follow from Model 5.1. If we use Algorithm 3 for robust sparse gradient estimation, it requires $\widetilde{\Omega}\left(\frac{r^2k^2\log(dT/\nu)}{\epsilon}\right) T$ samples, and $T = \Theta\left(\log\left(\frac{\|\boldsymbol{\beta}^*\|_2}{\sigma\sqrt{\epsilon}}\right)\right)$, then, we have$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \widetilde{O}\left(\sigma\sqrt{\epsilon}\right)$, with probability at least $1 - \nu - T\exp\left(-\Theta\left(\epsilon n\right)\right)$.*

The proof of Theorem 5.1 is collected in Appendix G, and the main technique hinges on previous analysis for the identity covariance case (Theorem 4.1 and Lemma 4.1). In the case of unknown covariance, this is the best (and in fact, only) recovery guarantee we are aware of for robust sparse regression. We show the performance of robust estimation using our filtering

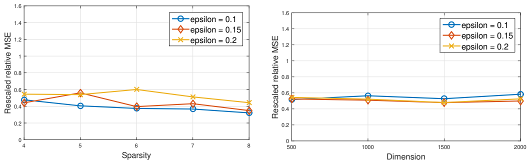

Figure 1: Simulations for Algorithm 3 showing the dependence of relative MSE on sparsity and dimension. For each parameter, we choose corresponding sample complexity $n \propto k^2 \log(d)/\epsilon$. Different curves for $\epsilon \in \{0.1, 0.15, 0.2\}$ are the average of 15 trials. Consistent with the theory, the rescaled relative MSE's are nearly independent of sparsity and dimension. Furthermore, by rescaling for different ϵ, three curves have the same magnitude.

algorithm with unknown covariance in Section 6, and we observe same linear convergence as Section 4.

6 Numerical results

We provide the complete details for our experiment setup and experiments on wall-clock time in Appendix.

Robust sparse mean estimation. We first demonstrate the performance of Algorithm 3 for robust sparse mean estimation, and then move to Algorithm 1 for robust sparse regression. For the robust sparse gradient estimation, we generate samples through $\boldsymbol{g}_i = \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{G} - \boldsymbol{x}_i\xi_i$, where the unknown true mean $\boldsymbol{G}$ is k-sparse. The authentic $\boldsymbol{x}_i$'s are generated from $\mathcal{N}(0, \boldsymbol{I}_d)$. We set $\sigma = 0$, since the main part of the error in robust sparse mean estimation is $\boldsymbol{G}$. We plot the relative MSE of parameter recovery, defined as $\|\widehat{\boldsymbol{G}} - \boldsymbol{G}\|_2^2 / \|\boldsymbol{G}\|_2^2$, with respect to different sparsities and dimensions.

Parameter error vs. sparsity k. We fix the dimension to be $d = 50$. We solve the trace norm maximization in Algorithm 3 using CVX Grant et al. (2008). We solve robust sparse gradient estimation under different levels of outlier fraction ϵ and different sparsity values k.

Parameter error vs. dimension d. We fix $k = 5$. We use a Sparse PCA solver from d'Aspremont et al. (2008) which is much more efficient for higher dimensions. We run robust sparse gradient estimation Algorithm 3 under different levels of outlier fraction ϵ and different dimensions d.

For each parameter, the corresponding number of samples required for the authentic data is $n \propto k^2 \log(d)/\epsilon$ according to Theorem 4.1. Therefore, we add $\epsilon n/(1-\epsilon)$ outliers (so that the outliers are an ϵ-fraction of the total samples), and then run Algorithm 3. According to Theorem 4.1, the rescaled relative MSE: $\|\widehat{\boldsymbol{G}} - \boldsymbol{G}\|_2^2 / (\epsilon\|\boldsymbol{G}\|_2^2)$ should be independent of the parameters $\{\epsilon, k, d\}$. We plot this in Figure 1, and these plots validate our theorem on the sample complexity in robust sparse mean estimation problems.

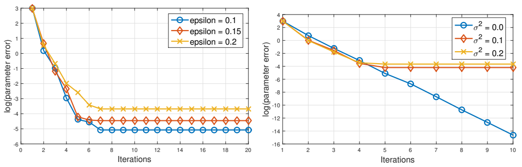

Figure 2: Empirical illustration of the linear convergence of $\log(\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2^2)$ vs. iteration counts in the Algorithm 1. In all cases, we fix $k = 5$, $d = 500$, and choose the sample complexity $n \propto 1/\epsilon$. The left plot considers different ϵ with fixed $\sigma^2 = 0.1$. The right plot considers different σ^2 with fixed $\epsilon = 0.1$. As expected, the convergence is linear, and flatten out at the level of the final error.

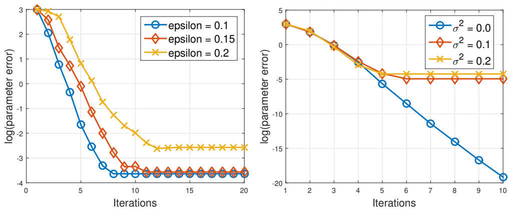

Figure 3: When the unknown covariance matrix is a Toeplitz matrix with a decay $\boldsymbol{\Sigma}_{ij} = \exp^{-(i-j)^2}$, and the other settings are the same as Figure 2. We observe similar linear convergence as Figure 2.

Robust sparse regression. We use Algorithm 1 for robust sparse regression. Similarly as in robust sparse mean estimation, we use Algorithm 3 as our Robust Sparse Gradient Estimator, and leverage the Sparse PCA solver from d'Aspremont et al. (2008). In the simulation, we fix $d = 500$, and $k = 5$, hence the corresponding sample complexity is $n \propto 1/\epsilon$. We do not use the sample splitting technique in the simulations.

To show the performance of Algorithm 1 under different settings, we use different levels of ϵ and σ in Figure 2, and track the parameter error $\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2^2$ of Algorithm 1 in each iteration. Consistent with the theory, the algorithm displays linear convergence, and the error curves flatten out at the level of the final error. Furthermore, Algorithm 1 achieves machine precision when $\sigma^2 = 0$ in the right plot of Figure 2.

We then study the empirical performance of robust sparse regression with unknown covariance matrix $\boldsymbol{\Sigma}$ following from Model 5.1. We use the same experimental setup as in identity covariance case, but modify the covariance matrix to be a Toeplitz matrix with a decay $\boldsymbol{\Sigma}_{ij} = \exp^{-(i-j)^2}$. Under this setting, the covariance matrix is sparse, thus follows from Model 5.1. Figure 3 indicates that we have nearly the same performance as the $\boldsymbol{\Sigma} = \boldsymbol{I}_d$ case.

7 Acknowledgments

This work was partially funded by NSF grants 1609279, 1646522, and 1704778, and the authors would like to thank Simon S. Du for helpful discussions.

References

Balakrishnan, S., Du, S. S., Li, J., and Singh, A. (2017a). Computationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 2017 Conference on Learning Theory*.

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017b). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120.

Bandeira, A. S., Dobriban, E., Mixon, D. G., and Sawin, W. F. (2013). Certifying the restricted isometry property is hard. *IEEE Transactions on Information Theory*, 59(6):3448–3450.

Barak, B. and Steurer, D. (2016). Proofs, beliefs, and algorithms through the lens of sum-of-squares. *Course notes: http://www. sumofsquares. org/public/index. html*.

Bhatia, K., Jain, P., and Kar, P. (2015). Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729.

Bhatia, K., Jain, P., and Kar, P. (2017). Consistent robust regression. In *Advances in Neural Information Processing Systems*, pages 2107–2116.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Blumensath, T. and Davies, M. E. (2009). Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

Chen, M., Gao, C., and Ren, Z. (2018). Robust covariance and scatter matrix estimation under hubers contamination model. *Ann. Statist.*, 46(5):1932–1960.

Chen, Y., Caramanis, C., and Mannor, S. (2013). Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pages 774–782.

Chen, Y., Su, L., and Xu, J. (2017). Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):44.

d'Aspremont, A., Bach, F., and Ghaoui, L. E. (2008). Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9(Jul):1269–1294.

d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., and Lanckriet, G. R. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Steinhardt, J., and Stewart, A. (2019a). Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2016). Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 655–664. IEEE.

Diakonikolas, I., Kane, D. M., and Stewart, A. (2017). Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 73–84. IEEE.

Diakonikolas, I., Kong, W., and Stewart, A. (2019b). Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM.

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

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, 36(6):2717–2756.

Gao, C. (2017). Robust regression via mutivariate regression depth. *arXiv preprint arXiv:1702.04656*.

Grant, M., Boyd, S., and Ye, Y. (2008). Cvx: Matlab software for disciplined convex programming.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons.

Holland, M. J. and Ikeda, K. (2017). Efficient learning with robust gradient descent. *arXiv preprint arXiv:1706.00182*.

Hopkins, S. B., Schramm, T., Shi, J., and Steurer, D. (2016). Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual*

ACM symposium on Theory of Computing, pages 178–191. ACM.

Huber, P. J. (1964). Robust estimation of a location parameter. *The annals of mathematical statistics*, pages 73–101.

Huber, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer.

Jain, P., Tewari, A., and Kar, P. (2014). On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693.

Johnson, D. S. and Preparata, F. P. (1978). The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107.

Karmalkar, S. and Price, E. (2018). Compressed sensing with adversarial sparse noise via l1 regression. In *2nd Symposium on Simplicity in Algorithms (SOSA 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Klivans, A., Kothari, P. K., and Meka, R. (2018). Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430.

Klivans, A. R., Long, P. M., and Servedio, R. A. (2009). Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(Dec):2715–2740.

Lai, K. A., Rao, A. B., and Vempala, S. (2016). Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 665–674. IEEE.

Li, X. (2013). Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37(1):73–99.

Liu, H. and Barber, R. F. (2018). Between hard and soft thresholding: optimal iterative thresholding algorithms. *arXiv preprint arXiv:1804.08841*.

Liu, L., Li, T., and Caramanis, C. (2019). High dimensional robust m-estimation: Arbitrary corruption and heavy tails. *arXiv preprint arXiv:1901.08237*.

Nguyen, N. H. and Tran, T. D. (2013). Exact recoverability from dense corrupted observations via l1-minimization. *IEEE transactions on information theory*, 59(4):2017–2035.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. (2018). Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*.

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

Shen, J. and Li, P. (2017). A tight bound of hard thresholding. *The Journal of Machine Learning Research*, 18(1):7650–7691.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151.

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

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Advances in neural information processing systems*, pages 2670–2678.

Wainwright, M. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.

Xu, H., Caramanis, C., and Mannor, S. (2013). Outlier-robust PCA: the high-dimensional case. *IEEE transactions on information theory*, 59(1):546–572.

Xu, H., Caramanis, C., and Sanghavi, S. (2012). Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064.

Zhang, Y., Wainwright, M. J., and Jordan, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948.