High Dimensional Robust Sparse Regression

Liu Liu
liuliu@utexas.edu

Yanyao Shen
shenyanyao@utexas.edu

Tianyang Li
lty@cs.utexas.edu

Constantine Caramanis
constantine@utexas.edu

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 [27, 23, 42], 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 [26], where the adversary independently corrupts each sample with probability ϵ.

Outlier-robust regression is a classical problem within robust statistics (e.g., [39]), yet even in the low-dimensional setting, efficient algorithms robust to corruption in the covariates have proved elusive, until recent breakthroughs in [38, 15] and [31], which built on important results in Robust Mean Estimation [13, 33] and Sums of Squares [4], 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., [18]) 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 assumptions on the probabilistic distribution are necessary, since in the worst case, sparse recovery is known to be NP-hard [3, 48].

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 [32, 47, 46, 33, 13], and more recently robust regression (as in this paper) [38, 15, 17, 31] and robust mean estimation [13, 33, 1], 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 [34, 37, 5, 6, 30]. 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 [31] shows that the Sum of Squares (SOS) based semidefinite hierarchy can be used for solving robust regression. Essentially concurrent to the SOS work, [12, 24, 38, 15] use robust gradient descent for empirical risk minimization, by using robust mean estimation as a subroutine to compute robust gradients at each iteration. [17] uses filtering algorithm [13] for robust regression. Computationally, these latter works scale better than the algorithms in [31], as although the Sum of Squares SDP framework gives polynomial time algorithms, they are often not practical [25].

Much less appears to be known in the high-dimensional regime. Exponential time algorithm, such as [21, 29], optimizes Tukey depth [42, 10]. Their results reveal that handling a constant fraction of outliers ($\epsilon = const.$) is actually minimax-optimal. Work in [11] 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, [36] 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 [1]. 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 [38, 15, 17] is to leverage recent breakthroughs in robust mean estimation. Very recently, [13, 33] provided the first robust mean estimation algorithms that can handle a constant fraction of outliers (though [33] incurs a small (logarithmic) dependence in the dimension). Following their work, [1] extended the ellipsoid algorithm from [13] 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 $n = \Omega\left(k^2 \log(d)\right)$ samples. The k^2 term appears to be necessary, as $n = \Omega(k^2)$ follows from an oracle-based lower bound [16].

1.2 Main contributions

- Our result is a robust variant of Iterative Hard Thresholding (IHT) [8]. 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 [1], 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 [21]). The proof of the ellipsoid algorithm's performance in [1] 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 [7]. 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 Appendix H. 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* $\xi_i \sim \mathcal{N}(0, \sigma^2)$, *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 [1], 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 $\|M\|_{\widetilde{\mathrm{k}},\mathrm{op}} = \max_{\|\boldsymbol{v}\|_2=1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} |\boldsymbol{v}^\top M \boldsymbol{v}|$, 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 statistical 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]

[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 [2, 38] for theoretical analysis. We do not use sample splitting technique in the experiments.

Algorithm 1 Robust sparse regression with RSGE

Input: Data samples $\{y_i, \boldsymbol{x}_i\}_{i=1}^N$, RSGE subroutine.
Output: The estimation $\widehat{\boldsymbol{\beta}}$.
Parameters: Hard thresholding parameter k'.

Split samples into T subsets of size n. Initialize with $\boldsymbol{\beta}^0 = \mathbf{0}$.
for $t = 0$ to $T-1$, **do**
  At current $\boldsymbol{\beta}^t$, calculate all gradients for current n samples: $\boldsymbol{g}_i^t = \boldsymbol{x}_i\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}^t - y_i\right)$, $i \in [n]$.
  The initial input set is $\{\boldsymbol{g}_i^t\}_{i=1}^n$. We use a RSGE to get $\widehat{\boldsymbol{G}}^t$.
  Update the parameter: $\boldsymbol{\beta}^{t+1} = \mathsf{P}_{k'}\left(\boldsymbol{\beta}^t - \eta \widehat{\boldsymbol{G}}^t\right)$.
end for
Output the estimation $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^T$.

2.1 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 [28, 40], and is proved to be necessary in [35]. 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 $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, [1] proposes a robust sparse regression estimator based on robust sparse 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 $N = \Omega\big(\frac{k^2 \log(d/\nu)}{\epsilon^2}\big)$, 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 [1] 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:

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

Algorithm 2 Separation oracle for robust sparse estimation [1]

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 ρ_{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: Solve: $\max_{\boldsymbol{H}} \operatorname{Tr}\left(\left(\widehat{\boldsymbol{\Sigma}} - F\left(\widehat{\boldsymbol{G}}\right)\right)\cdot \boldsymbol{H}\right)$, subject to $\boldsymbol{H} \succcurlyeq 0, \|\boldsymbol{H}\|_{1,1} \le \widetilde{k}, \operatorname{Tr}(\boldsymbol{H}) = 1$.
Let λ^* be the optimal value, and $\boldsymbol{H}^*$ be the corresponding solution.
7: **if** $\lambda^* \le \rho_{\text{sep}}$ **, then return** "Yes".
8: **return** The hyperplane: $\ell(w') = \left\langle \left(\sum_{i\in\mathcal{S}} w_i'(\boldsymbol{g}_i - \widehat{\boldsymbol{G}})(\boldsymbol{g}_i - \widehat{\boldsymbol{G}})^\top - F(\widehat{\boldsymbol{G}})\right), \boldsymbol{H}^*\right\rangle - \lambda^*$.

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

$$F(\widehat{\boldsymbol{G}}) = \|\widehat{\boldsymbol{G}}\|_2^2 \boldsymbol{I}_d + \widehat{\boldsymbol{G}}\widehat{\boldsymbol{G}}^\top + \sigma^2 \boldsymbol{I}_d. \tag{2}$$

For eq. (2), 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 [13, 33]. 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 $\lambda^* \le \rho_{\text{sep}}$ (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 [1] 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 $\rho_{\text{sep}} = \Theta\big(\epsilon\big(\|\boldsymbol{G}^t\|_2^2 + \sigma^2\big)\big)$, it requires $N(k, d, \epsilon, \nu) = \Omega\big(\frac{k^2\log(dT/\nu)}{\epsilon^2}\big)T$ 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 $T = \Theta\left(\log\left(\frac{\|\boldsymbol{\beta}^*\|_2}{\epsilon\sigma}\right)\right)$.*

For a desired error level $\epsilon' \ge \epsilon$, we only require sample complexity $N(k, d, \epsilon, \nu) = \Omega\big(\frac{k^2\log(dT/\nu)}{\epsilon'^2}\big)T$. Hence, we can achieve statistical error $\widetilde{O}\big(\sigma\big(\sqrt{k^2\log(d)/N} \vee \epsilon\big)\big)$. Our error bound is nearly optimal compared to the information-theoretically optimal $O\big(\sigma\big(\sqrt{k\log(d)/N} \vee \epsilon\big)\big)$ in [21], as the k^2 term is necessary by an oracle-based SQ lower bound [16].

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 [1] claims that $\lambda^* \ge \|\widehat{\boldsymbol{\Sigma}} - F(\widehat{\boldsymbol{G}})\|_{\widetilde{k},\text{op}}$. 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

Algorithm 3 RSGE via filtering

1: **Input:** A set $\mathcal{S}_{\text{in}}$.
2: **Output:** A set $\mathcal{S}_{\text{out}}$ or sparse mean vector $\widehat{\boldsymbol{G}}$.
3: **Parameters:** Hard thresholding parameter $\widetilde{k}$, parameter ρ_{sep}.

4: Compute the sample mean $\widetilde{\boldsymbol{G}} = \mathbb{E}_{i\in_u \mathcal{S}_{\text{in}}}\left(\boldsymbol{g}_i\right)$, and $\widehat{\boldsymbol{G}} = \mathsf{P}_{2\widetilde{k}}(\widetilde{\boldsymbol{G}})$.
5: Compute the sample covariance matrix $\widehat{\boldsymbol{\Sigma}} = \mathbb{E}_{i\in_u \mathcal{S}_{\text{in}}}\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)^\top$.
6: Solve the following convex program:

$$\max_{\boldsymbol{H}} \operatorname{Tr}\left(\widehat{\boldsymbol{\Sigma}} \cdot \boldsymbol{H}\right), \quad \text{subject to } \boldsymbol{H} \succcurlyeq 0, \|\boldsymbol{H}\|_{1,1} \leq \widetilde{k}, \operatorname{Tr}(\boldsymbol{H}) = 1. \tag{3}$$

Let λ^* be the optimal value, and $\boldsymbol{H}^*$ be the corresponding solution.
7: **if** $\lambda^* \leq \rho_{\text{sep}}$ **, then return** with $\widehat{\boldsymbol{G}}$.
8: Calculate projection score for each $i \in \mathcal{S}_{\text{in}}$:

$$\tau_i = \operatorname{Tr}\Big(\boldsymbol{H}^* \cdot \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)^\top\Big).$$

9: Randomly remove a sample r from $\mathcal{S}_{\text{in}}$ according to

$$\Pr\left(\boldsymbol{g}_i \text{ is removed}\right) = \frac{\tau_i}{\sum_{i\in\mathcal{S}_{\text{in}}} \tau_i}. \tag{4}$$

10: **return** the set $\mathcal{S}_{\text{out}} = \mathcal{S}_{\text{in}} \setminus \{r\}$.

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 ([9]). In this section, we introduce a faster algorithm as a RSGE, which only requires $O(n)$ 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 can be used for other robust sparse

[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

functional estimation problems (e.g., robust sparse mean estimation for $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I_d})$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ is k-sparse). To use Algorithm 3 as a RSGE given n gradient samples (denoted as $\mathcal{S}_{\text{in}}$), we call Algorithm 3 repeatedly on $\mathcal{S}_{\text{in}}$ and then on its output, $\mathcal{S}_{\text{out}}$, 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 $\mathcal{S}_{\text{in}}$ 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 $\mathcal{S}_{\text{in}}$, and subsequently on $\mathcal{S}_{\text{out}}$, and use $\rho_{\text{sep}} = C_\gamma\left(\|\boldsymbol{G}^t\|_2^2 + \sigma^2\right)$, [3] 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.*

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 $\widetilde{O}\left(\sigma\left(\sqrt{k^2 \log(d)/N} \vee \sqrt{\epsilon}\right)\right)$. 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 Appendix H.

The key step in Algorithm 3 is outlier removal eq. (4) based on the solution of Sparse PCA's convex relaxation eq. (3). 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 $n = \Omega\left(\frac{k^2 \log(d/\nu)}{\epsilon}\right)$ ϵ-corrupted samples from Model 1.1 with $\boldsymbol{\Sigma} = \boldsymbol{I}_d$. Let $\mathcal{S}_{\text{in}}$ 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 line 8, 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.

[3] Similar to [13, 14, 15, 1], our results seem to require this side information.

5 Robust sparse regression with unknown covariance

In this section, we consider robust sparse regression with unknown covariance matrix $\mathbf{\Sigma}$, 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 $\mathcal{N}(0, \mathbf{\Sigma})$. We assume that each row and column of $\mathbf{\Sigma}$ is r-sparse, but the positions of the non-zero entries are unknown.*

Model 5.1 is widely studied in high dimensional statistics [7, 20, 45]. Under Model 5.1, for the population gradient $\boldsymbol{G}^t = \mathbb{E}_P\left(\boldsymbol{x}_i \boldsymbol{x}_i^\top (\boldsymbol{\beta}^t - \boldsymbol{\beta}^*)\right) = \mathbf{\Sigma}\omega^t$, where we use ω^t to denote the $(k'+k)$-sparse vector $\boldsymbol{\beta}^t - \boldsymbol{\beta}^*$, we can guarantee the $\|\boldsymbol{G}^t\|_0 = \|\mathbf{\Sigma}\omega^t\|_0 \le r(k'+k)$. Hence, we can use the filtering algorithm (Algorithm 3) with $\widetilde{k} = r(k'+k)$ as a RSGE for robust sparse regression with unknown $\mathbf{\Sigma}$. 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 $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \widetilde{O}(\sigma\sqrt{\epsilon})$, with probability at least $1 - \nu - T\exp(-\Theta(\epsilon n))$.*

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

6 Acknowledgments

The authors would like to thank Simon S. Du for helpful discussions.

References

[1] Sivaraman Balakrishnan, Simon S. Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.

[2] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[3] Afonso S. Bandeira, Edgar Dobriban, Dustin G. Mixon, and William F. Sawin. Certifying the restricted isometry property is hard. *IEEE Transactions on Information Theory*, 59(6):3448–3450, 2013.

[4] Boaz Barak and David Steurer. Proofs, beliefs, and algorithms through the lens of sum-of-squares. *Course notes: http://www. sumofsquares. org/public/index. html*, 2016.

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

[6] Kush Bhatia, Prateek Jain, and Purushottam Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems*, pages 2107–2116, 2017.

[7] Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[8] Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

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

[10] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under hubers contamination model. *Ann. Statist.*, 46(5):1932–1960, 10 2018.

[11] Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pages 774–782, 2013.

[12] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):44, 2017.

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

[14] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008, 2017.

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

[16] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. 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, 2017.

[17] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. 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, 2019.

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

[19] Alexandre dAspremont, Francis Bach, and Laurent El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9(Jul):1269–1294, 2008.

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

[21] Chao Gao. Robust regression via mutivariate regression depth. *arXiv preprint arXiv:1702.04656*, 2017.

[22] Michael Grant, Stephen Boyd, and Yinyu Ye. Cvx: Matlab software for disciplined convex programming, 2008.

[23] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

[24] Matthew J Holland and Kazushi Ikeda. Efficient learning with robust gradient descent. *arXiv preprint arXiv:1706.00182*, 2017.

[25] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. 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, 2016.

[26] Peter J Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, pages 73–101, 1964.

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

[28] Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

[29] David S Johnson and Franco P Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107, 1978.

[30] Sushrut Karmalkar and Eric Price. Compressed sensing with adversarial sparse noise via l1 regression. *arXiv preprint arXiv:1809.08055*, 2018.

[31] Adam Klivans, Pravesh K. Kothari, and Raghu Meka. Efficient Algorithms for Outlier-Robust Regression. *arXiv preprint arXiv:1803.03241*, 2018.

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

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

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

[35] Haoyang Liu and Rina Foygel Barber. Between hard and soft thresholding: optimal iterative thresholding algorithms. *arXiv preprint arXiv:1804.08841*, 2018.

[36] Liu Liu, Tianyang Li, and Constantine Caramanis. High dimensional robust estimation of sparse models via trimmed hard thresholding. *arXiv preprint arXiv:1901.08237*, 2019.

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

[38] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

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

[40] Jie Shen and Ping Li. A tight bound of hard thresholding. *The Journal of Machine Learning Research*, 18(1):7650–7691, 2017.

[41] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

[42] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

[43] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[44] Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Advances in neural information processing systems*, pages 2670–2678, 2013.

[45] Martin Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

[46] Huan Xu, Constantine Caramanis, and Shie Mannor. Outlier-robust PCA: the high-dimensional case. *IEEE transactions on information theory*, 59(1):546–572, 2013.

[47] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012.

[48] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948, 2014.

Notation. In proofs, we let $\otimes$ denote the Kronecker product, and for a vector $\boldsymbol{u}$, we denote the outer product by $\boldsymbol{u}^{\otimes 2} = \boldsymbol{u}\boldsymbol{u}^\top$. We define the infinity norm for a matrix M as $\|M\|_\infty = \max_{i,j} |M_{ij}|$. Given index set $\mathcal{J}$, $\boldsymbol{v}^{\mathcal{J}}$ is the vector restricted to indices $\mathcal{J}$. Similarly, $M^{\mathcal{J}\mathcal{J}}$ is the sub-matrix on indices $\mathcal{J} \times \mathcal{J}$. we use $\{C_j\}_{j=0}^3$ to denote constants that are independent of dimension, but whose value can change from line to line.

A Proofs for the meta-theorem

In this section, we prove the global linear convergence guarantee given the Definition 2.1. In each iteration of Algorithm 1, we use $\widehat{\boldsymbol{G}}^t$ to update

$$\boldsymbol{\beta}^{t+1} = \mathsf{P}_{k'}\left(\boldsymbol{\beta}^t - \eta \widehat{\boldsymbol{G}}^t\right),$$

where $\eta = 1/\mu_\beta$ is a fixed step size. Given the condition $\|\widehat{\boldsymbol{G}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta})\|_2^2 \le \alpha(\epsilon)\|\boldsymbol{G}(\boldsymbol{\beta})\|_2^2 + \psi(\epsilon)$ in RSGE's definition, we show that Algorithm 1 linearly converges to a neighborhood around $\boldsymbol{\beta}^*$ with error at most $O(\sqrt{\psi(\epsilon)})$.

First, we introduce a supporting Lemma from [40], which bounds the distance between $\mathsf{P}_{k'}(\boldsymbol{\beta}^t - \eta\widehat{\boldsymbol{G}}^t)$ and $\boldsymbol{\beta}^*$ in each iteration of Algorithm 1.

Lemma A.1 (Theorem 1 in [40]). *Let $\boldsymbol{z} \in \mathbb{R}^d$ be an arbitrary vector and $\boldsymbol{\beta}^* \in \mathbb{R}^d$ be any k-sparse signal. For any $k' \ge k$, we have the following bound:*

$$\|\mathsf{P}_{k'}(\boldsymbol{z}) - \boldsymbol{\beta}^*\|_2 \le \sqrt{\zeta}\|\boldsymbol{z} - \boldsymbol{\beta}^*\|_2, \quad \zeta = 1 + \frac{\rho + \sqrt{(4+\rho)\rho}}{2}, \quad \rho = \frac{\min\{k, d-k'\}}{k' - k + \min\{k, d-k'\}}.$$

We choose the hard thresholding parameter $k' = kc_\kappa^2 \ll d$, hence $\rho = 1/c_\kappa^2$.

Theorem A.1 (Theorem 2.1). *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*

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2 = O\left(\sqrt{\psi(\epsilon)}\right),$$

with probability at least $1 - \nu$, by setting $k' = c_\kappa^2 k$ and $T = \Theta\left(\log\left(\|\boldsymbol{\beta}^*\|_2/\sqrt{\psi(\epsilon)}\right)\right)$. The sample complexity is $N(k, d, \epsilon, \nu) = n(k, d, \epsilon, \nu/T)\, T$.

Proof. By splitting N samples into T sets (each set has sample size n), Algorithm 1 collects a fresh batch of samples with size $n(k, d, \epsilon, \nu/T)$ at each iteration $t \in [T]$. Definition 2.1 shows that for the fixed gradient expectation $\boldsymbol{G}^t$, the estimate for the gradient $\boldsymbol{G}^t$ satisfies:

$$\left\|\widehat{\boldsymbol{G}}^t - \boldsymbol{G}^t\right\|_2^2 \le \alpha(\epsilon)\left\|\boldsymbol{G}^t\right\|_2^2 + \psi(\epsilon) \tag{6}$$

with probability at least $1 - \nu/T$, where $\alpha(\epsilon)$ is determined by ϵ.

Letting $z^t = \boldsymbol{\beta}^t - \eta\widehat{\boldsymbol{G}}^t$, we study the t-th iteration of Algorithm 1. Based on Lemma A.1, we have

$$\left\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\right\|_2 \le \sqrt{\zeta}\left\|\boldsymbol{\beta}^t - \eta\widehat{\boldsymbol{G}} - \boldsymbol{\beta}^*\right\|_2$$

$$
\begin{aligned}
&= \sqrt{\zeta}\left\|\boldsymbol{\beta}^t - \eta\boldsymbol{G} - \boldsymbol{\beta}^* + \eta(\boldsymbol{G} - \widehat{\boldsymbol{G}})\right\|_2 \\
&\leq \sqrt{\zeta}\left\|\boldsymbol{\beta}^t - \eta\boldsymbol{G} - \boldsymbol{\beta}^*\right\|_2 + \sqrt{\zeta}\eta\left\|\boldsymbol{G} - \widehat{\boldsymbol{G}}\right\|_2 \\
&\overset{(i)}{\leq} \sqrt{\zeta}\left\|(\boldsymbol{I}_d - \eta\boldsymbol{\Sigma})(\boldsymbol{\beta}^t - \boldsymbol{\beta}^*)\right\|_2 + \sqrt{\zeta}\eta\sqrt{\alpha(\epsilon)\|\boldsymbol{G}\|_2^2 + \psi(\epsilon)} \\
&\overset{(ii)}{\leq} \sqrt{\zeta}\left\|(\boldsymbol{I}_d - \eta\boldsymbol{\Sigma})(\boldsymbol{\beta}^t - \boldsymbol{\beta}^*)\right\|_2 + \sqrt{\zeta}\eta\sqrt{\alpha(\epsilon)}\left\|\boldsymbol{\Sigma}(\boldsymbol{\beta}^t - \boldsymbol{\beta}^*)\right\|_2 + \sqrt{\zeta}\eta\sqrt{\psi(\epsilon)}
\end{aligned}
$$

where (i) follows from the theoretical guarantee of RSGE, and (ii) follows from the basic inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for non-negative a, b.

By setting $\eta = 1/\mu_\beta$, we have

$$
\begin{aligned}
\left\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\right\|_2 &\leq \sqrt{\zeta}\left\|(\boldsymbol{I}_d - \eta\boldsymbol{\Sigma})(\boldsymbol{\beta}^t - \boldsymbol{\beta}^*)\right\|_2 + \sqrt{\zeta}\eta\sqrt{\alpha(\epsilon)}\left\|\boldsymbol{\Sigma}(\boldsymbol{\beta}^t - \boldsymbol{\beta}^*)\right\|_2 + \sqrt{\zeta}\eta\sqrt{\psi(\epsilon)} \\
&\leq \sqrt{\zeta}(1 - \frac{1}{c_\kappa})\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\right\|_2 + \sqrt{\zeta}\sqrt{\alpha(\epsilon)}\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\right\|_2 + \sqrt{\zeta}\eta\sqrt{\psi(\epsilon)} \\
&\leq \sqrt{\zeta}(1 - \frac{1}{c_\kappa} + \sqrt{\alpha(\epsilon)})\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\right\|_2 + \sqrt{\zeta}\eta\sqrt{\psi(\epsilon)} \qquad (7)
\end{aligned}
$$

When ϵ is a small enough constant, we have $\sqrt{\alpha(\epsilon)} \leq \frac{1}{2c_\kappa}$, then

$$
\begin{aligned}
\sqrt{\zeta}(1 - \frac{1}{c_\kappa} + \sqrt{\alpha(\epsilon)}) &\leq \sqrt{\zeta}(1 - \frac{1}{2c_\kappa}) \\
&\leq \sqrt{1 + \frac{\rho + \sqrt{(4+\rho)\rho}}{2}}(1 - \frac{1}{2c_\kappa})
\end{aligned}
$$

Plugging in the parameter $\rho = 1/c_\kappa^2$ in Lemma A.1, we have

$$
\sqrt{\zeta}(1 - \frac{1}{c_\kappa} + \sqrt{\alpha(\epsilon)}) \leq 1 - \frac{1}{10c_\kappa}
$$

Together with eq. (7), we have the recursion

$$
\left\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\right\|_2 \leq \left(1 - \frac{1}{10c_\kappa}\right)\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\right\|_2 + \sqrt{\zeta}\eta\sqrt{\psi(\epsilon)}.
$$

By solving this recursion and using a union bound, we have

$$
\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\right\|_2 \leq \left(1 - \frac{1}{10c_\kappa}\right)^t\left\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^*\right\|_2 + \frac{\sqrt{\zeta}\eta\sqrt{\psi(\epsilon)}}{1 - \left(1 - \frac{1}{10c_\kappa}\right)} \leq (4\alpha(\epsilon))^t\left\|\boldsymbol{\beta}^*\right\|_2^2 + 10c_\kappa\sqrt{\zeta}\eta\sqrt{\psi(\epsilon)},
$$

with probability at least $1 - \nu$.

By the definition of c_κ and η, we have $\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2 = O\left(\frac{\sqrt{\psi(\epsilon)}}{\mu_\alpha}\right)$ □

B Correcting Lemma A.3 in [1]'s proof

A key part of the proof of the main theorem in [1] is to obtain an upper bound on the k-sparse operator norm. Specifically, their Lemmas A.2 and A.3 aim to show:

$$\lambda^* \geq \left\| \sum_{i=1}^{|\mathcal{S}|} w_i \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}(w) \right)^{\otimes 2} - F\left(\widehat{\boldsymbol{G}}(w) \right) \right\|_{\widetilde{\mathrm{k}},\mathrm{op}} \geq \frac{\left\| \mathsf{P}_{\widetilde{k}} \left(\widetilde{\Delta}(w) \right) \right\|_2^2}{5\epsilon}, \tag{8}$$

where $\widehat{\boldsymbol{G}}(w) = \mathsf{P}_{2\widetilde{k}} \left(\sum_{i=1}^{|\mathcal{S}|} w_i \boldsymbol{g}_i \right)$, $\widetilde{\Delta}(w) = \sum_{i=1}^{|\mathcal{S}|} w_i \boldsymbol{g}_i - \boldsymbol{G}$[4], and recall λ^* is the solution to the SDP as given in Algorithm 3.

Lemma A.3 asserts the first inequality above, and Lemma A.2 the second. As we show below, Lemma A.3 cannot be correct. Specifically, the issue is that the quantity inside the second term in eq. (8) may not be positive semidefinite. In this case, the convex optimization problem whose solution is λ^* is not a valid relaxation, and hence the λ^* they obtain need not a valid upper bound. Indeed, we give a simple example below that illustrates precisely this potential issue.

Fortunately, not all is lost – indeed, as our results imply, the main results in [1] is correct. The key is to show that while λ^* does not upper bound the sparse operator norm, it does, however, upper bound the quantity

$$\max_{\|\boldsymbol{v}\|_2 = 1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} \boldsymbol{v}^\top \left(\sum_{i=1}^{|\mathcal{S}|} w_i \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}(w) \right)^{\otimes 2} - F\left(\widehat{\boldsymbol{G}}(w) \right) \right) \boldsymbol{v}. \tag{9}$$

We show this in Appendix D. More specifically, in Lemma D.3, we replace the $\widetilde{k}$-sparse operator norm in the second term of eq. (8) by the term in eq. (9). We show this can be used to complete the proof in Appendix D.4.

We now provide a counterexample that shows the first inequality in (8) cannot hold. The main argument is that the convex relaxation for sparse PCA is a valid upper bound of the sparse operator norm only for positive semidefinite matrices. Specifically, denoting $\boldsymbol{E} = \widehat{\boldsymbol{\Sigma}}(w) - F(\widehat{\boldsymbol{G}}(w))$ as the matrix in eq. (9), [1] solves the following convex program:

$$\max_{\boldsymbol{H}} \mathrm{Tr}\left(\boldsymbol{E} \cdot \boldsymbol{H} \right), \quad \text{subject to } \boldsymbol{H} \succcurlyeq 0, \|\boldsymbol{H}\|_{1,1} \leq k, \mathrm{Tr}\left(\boldsymbol{H} \right) = 1.$$

Since $\widehat{\boldsymbol{\Sigma}}(w) - F(\widehat{\boldsymbol{G}}(w))$ is no longer a p.s.d. matrix, the trace maximization above may not be a valid convex relaxation, and thus not an upper bound. Let us consider a specific example, in robust sparse mean estimation for $\mathcal{N}(\mu, \boldsymbol{I}_d)$, where function $F(\cdot)$ is a fixed identity matrix $\boldsymbol{I}_d$. We choose $\widetilde{k} = 1$, $\mu = [1, 0]^\top$, and $d = 2$. Suppose we observe data to be $x_1 = [2.5, 0]^\top$, $x_2 = [0, 0]^\top$, and the weights for x_1 and x_2 are the same. Then, we can compute the following matrices as:

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1.5625 & 0 \\ 0 & 0 \end{bmatrix}, F = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{E} = \widehat{\boldsymbol{\Sigma}} - F = \begin{bmatrix} 0.5625 & 0 \\ 0 & -1 \end{bmatrix}.$$

[4]The $\{w_i\}$ are weights, and these are defined precisely in Section D, but are not required for the present discussion or counterexample.

It is clear that $\|\widehat{\boldsymbol{\Sigma}} - F\|_{\widetilde{\mathrm{k}},\mathrm{op}} = 1$. Solving the convex relaxation $\max_{\boldsymbol{H}} \operatorname{Tr}(\boldsymbol{E} \cdot \boldsymbol{H})$ or $\max_{\boldsymbol{H}} \operatorname{Tr}(\widehat{\boldsymbol{\Sigma}} \cdot \boldsymbol{H})$ gives answer $\boldsymbol{H}^* = \begin{bmatrix} 1 & 0; 0 & 0 \end{bmatrix}$ and the corresponding $\lambda^* = 0.5625$, which is clearly not an upper bound of $\|\widehat{\boldsymbol{\Sigma}} - F\|_{\widetilde{\mathrm{k}},\mathrm{op}}$. Hence $\lambda^* \geq \|\widehat{\boldsymbol{\Sigma}} - F\|_{\widetilde{\mathrm{k}},\mathrm{op}}$ cannot hold in general.

C Covariance smoothness properties in robust sparse mean estimation

When the covariance is identity, the ellipsoid algorithm requires a closed form expression of the true covariance function $F(\boldsymbol{G})$. Indeed, the ellipsoid-based robust sparse mean estimation algorithm uses the covariance structure given by $F(\cdot)$ to detect outliers. The accuracy of robust sparse mean estimation explicitly depends on the properties of $F(\boldsymbol{G})$. L_{cov} and L_{F} are two important properties of $F(\boldsymbol{G})$, related to its smoothness. We first provide a closed-form expression for F, and then define precisely smoothness parameters L_{cov} and L_{F}, and show how these can be controlled.

Closed form expression of $F(\boldsymbol{G})$.

Lemma C.1. *Suppose we observe i.i.d. samples $\{\boldsymbol{z}_i, i \in \mathcal{G}\}$ from the distribution P in Model 1.1 with $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, we have the covariance of gradient as*

$$\operatorname{Cov}(\boldsymbol{g}) = \mathbb{E}_{\boldsymbol{z}_i \sim P}\left((\boldsymbol{g}_i - \boldsymbol{G})(\boldsymbol{g}_i - \boldsymbol{G})^\top\right) = \|\boldsymbol{G}\|_2^2 \boldsymbol{I}_d + \boldsymbol{G}\boldsymbol{G}^\top + \sigma^2 \boldsymbol{I}_d.$$

Proof. Since $\boldsymbol{g}_i = \boldsymbol{x}_i\left(\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i\right)$, and $\boldsymbol{G} = \mathbb{E}_{\boldsymbol{z}_i \sim P}(\boldsymbol{g}_i)$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z}_i \sim P}\left((\boldsymbol{g}_i - \boldsymbol{G})(\boldsymbol{g}_i - \boldsymbol{G})^\top\right) &= \mathbb{E}_P\left(\left(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}_d\right)\boldsymbol{G}\boldsymbol{G}^\top\left(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}_d\right)\right) + \sigma^2 \boldsymbol{I}_d \\
&= \mathbb{E}_P\left(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{G}\boldsymbol{G}^\top \boldsymbol{x}\boldsymbol{x}^\top\right) - 2\,\mathbb{E}_P\left(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{G}\boldsymbol{G}^\top\right) + \boldsymbol{G}\boldsymbol{G}^\top + \sigma^2 \boldsymbol{I}_d,
\end{aligned}$$

where we drop i in $\boldsymbol{x}_i$ without abuse of notation.

Next, we apply the Stein-type Lemma [41] for $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, and a function $f(x)$ whose second derivative exists:

$$\mathbb{E}\left(f(x)\boldsymbol{x}\boldsymbol{x}^\top\right) = \mathbb{E}(f(x))\boldsymbol{I}_d + \mathbb{E}\left(\nabla^2 f(x)\right). \tag{10}$$

By eq. (10), we have

$$\operatorname{Cov}(\boldsymbol{g}) = \|\boldsymbol{G}\|_2^2 \boldsymbol{I}_d + \boldsymbol{G}\boldsymbol{G}^\top + \sigma^2 \boldsymbol{I}_d.$$

□

Smoothness properties of $\|F\|_{\mathrm{op}}$. We first assume

$$L_{\mathrm{cov}} = \max_{\|\boldsymbol{v}\|_2 = 1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} \left|\boldsymbol{v}^\top \operatorname{Cov}(\boldsymbol{g})\boldsymbol{v}\right|. \tag{11}$$

If we define the functional $F(\cdot)$, such that $F(\widehat{\boldsymbol{G}}) = \|\widehat{\boldsymbol{G}}\|_2^2 \boldsymbol{I}_d + \widehat{\boldsymbol{G}}\widehat{\boldsymbol{G}}^\top + \sigma^2 \boldsymbol{I}_d$, and $F(\boldsymbol{G}) = \|\boldsymbol{G}\|_2^2 \boldsymbol{I}_d + \boldsymbol{G}\boldsymbol{G}^\top + \sigma^2 \boldsymbol{I}_d$, then we assume that there exists L_{F} satisfying

$$\left\|F\left(\boldsymbol{G}\right) - F\left(\widehat{\boldsymbol{G}}\right)\right\|_{\mathrm{op}} \leq L_{\mathrm{F}}\left\|\boldsymbol{G} - \widehat{\boldsymbol{G}}\right\|_2 + C\left\|\boldsymbol{G} - \widehat{\boldsymbol{G}}\right\|_2^2, \tag{12}$$

where C is a universal constant.

Lemma C.2. *Under the same setting as Lemma C.1, we have*

$$L_{\mathrm{cov}} = 2\|\boldsymbol{G}\|_2^2 + \sigma^2, \text{ and } L_{\mathrm{F}} = 4\|\boldsymbol{G}\|_2.$$

Proof. L_{cov} is upper bounded by the top eigenvalue of $F\left(\boldsymbol{G}\right)$,

$$L_{\mathrm{cov}} \leq \left\|F\left(\boldsymbol{G}\right)\right\|_2 \leq 2\|\boldsymbol{G}\|_2^2 + \sigma^2.$$

For the L_{F} term, we have

$$\begin{aligned}
&\left\|F\left(\boldsymbol{G}\right) - F\left(\widehat{\boldsymbol{G}}\right)\right\|_{\mathrm{op}} \\
&= \left\|2\boldsymbol{G}^\top\left(\boldsymbol{G} - \widehat{\boldsymbol{G}}\right)\boldsymbol{I}_d - \left\|\boldsymbol{G} - \widehat{\boldsymbol{G}}\right\|_2^2 \boldsymbol{I}_d + \boldsymbol{G}\left(\boldsymbol{G} - \widehat{\boldsymbol{G}}\right)^\top + \left(\boldsymbol{G} - \widehat{\boldsymbol{G}}\right)\boldsymbol{G}^\top - \left(\boldsymbol{G} - \widehat{\boldsymbol{G}}\right)\left(\boldsymbol{G} - \widehat{\boldsymbol{G}}\right)^\top\right\|_{\mathrm{op}} \\
&\leq 4\|\boldsymbol{G}\|_2\left\|\boldsymbol{G} - \widehat{\boldsymbol{G}}\right\|_2 + 2\left\|\boldsymbol{G} - \widehat{\boldsymbol{G}}\right\|_2^2.
\end{aligned}$$

Therefore, we can choose $L_{\mathrm{F}} = 4\|\boldsymbol{G}\|_2$ and $C = 2$.

□

D Proofs for the ellipsoid algorithm in robust sparse regression

In this section, we prove guarantees for the ellipsoid algorithm in robust sparse regression. In the theoretical analysis of the ellipsoid algorithm, we use $\mathcal{S}_{\mathrm{in}}$ to denote the observations $\mathcal{S}$, which shares the same notations with Algorithm 3. We first give preliminary definitions of error terms defined on $\mathcal{S}_{\mathrm{good}}$ and $\mathcal{S}_{\mathrm{in}}$, and then prove Lemma D.1. Next, we prove concentration results for gradients of uncorrupted sparse linear regression in Lemma D.2. In Lemma D.3, we provide lower bounds for the $\widetilde{k}$-sparse largest eigenvalue defined in eq. (9). Finally, we prove Corollary 3.1 based on previous Lemmas in Appendix D.4.

D.1 Preliminary definitions and properties related to $\mathcal{S}_{\mathrm{good}}, \mathcal{S}_{\mathrm{bad}}$

Here, we state again the definitions of $\mathcal{S}_{\mathrm{good}}$, $\mathcal{S}_{\mathrm{bad}}$ and $\mathcal{S}_{\mathrm{in}}$. In Algorithm 3, we denote the input set as $\mathcal{S}_{\mathrm{in}}$, which can be partitioned into two parts: $\mathcal{S}_{\mathrm{good}} = \{i : i \in \mathcal{G} \text{ and } i \in \mathcal{S}_{\mathrm{in}}\}$, and $\mathcal{S}_{\mathrm{bad}} = \{i : i \in \mathcal{B} \text{ and } i \in \mathcal{S}_{\mathrm{in}}\}$. Note that $\mathcal{S}_{\mathrm{in}} = \mathcal{S}_{\mathrm{good}} \cup \mathcal{S}_{\mathrm{bad}}$, and $n = |\mathcal{S}_{\mathrm{in}}|$. For the convenience of our analysis, we

define the following error terms:

$$\begin{aligned}
\widetilde{\Delta}_{\mathcal{S}_{\text{good}}} &= \mathbb{E}_{i\in_u\mathcal{S}_{\text{good}}}(\boldsymbol{g}_i) - \boldsymbol{G},\\
\widehat{\Delta}_{\mathcal{S}_{\text{good}}} &= \mathsf{P}_{2\widetilde{k}}\left(\mathbb{E}_{i\in_u\mathcal{S}_{\text{good}}}(\boldsymbol{g}_i)\right) - \boldsymbol{G},\\
\widetilde{\Delta} &= \mathbb{E}_{i\in_u\mathcal{S}_{\text{in}}}(\boldsymbol{g}_i) - \boldsymbol{G},\\
\widehat{\Delta} &= \mathsf{P}_{2\widetilde{k}}\left(\mathbb{E}_{i\in_u\mathcal{S}_{\text{in}}}(\boldsymbol{g}_i)\right) - \boldsymbol{G}.
\end{aligned}$$

These error terms are defined under a uniform distribution over samples, whereas previous papers using ellipsoid algorithms consider a set of balanced weighted distribution. More specifically, the weights in our setting are defined as:

$$\widetilde{w}_i = \frac{1}{n}, \quad \forall i \in \mathcal{S}_{\text{good}} \cup \mathcal{S}_{\text{bad}}.$$

The balanced weighted distribution is defined to satisfy:

$$0 \le w_i \le \frac{1}{(1-2\epsilon)n}, \quad \forall i \in \mathcal{S}_{\text{good}} \cup \mathcal{S}_{\text{bad}}, \qquad \sum_{i\in\mathcal{S}_{\text{in}}} w_i = 1.$$

Notice that $\sum_{i\in\mathcal{S}_{\text{bad}}} \widetilde{w}_i = O(\epsilon)$, and $\sum_{i\in\mathcal{S}_{\text{bad}}} w_i = O\left(\frac{\epsilon}{1-2\epsilon}\right)$ with high probability, which intuitively says that both types of distributions have $O(\epsilon)$ weights over all bad samples. We are interested in considering uniform weighted samples since this formulation helps us analyze the filtering algorithm more conveniently, as we show in the following sections.

We restate the following Lemma which shows the connection of these different error terms.

Lemma D.1 (Lemma A.1 in [1])**.** *Suppose G is k-sparse. Then we have the following result:*

$$\frac{1}{5}\left\|\widehat{\Delta}\right\|_2 \le \left\|\mathsf{P}_k\left(\widetilde{\Delta}\right)\right\|_2 \le 4\left\|\widehat{\Delta}\right\|_2.$$

D.2 Concentration bounds for gradients in $\mathcal{S}_{\text{good}}$

We first prove concentration bounds for gradients for sparse linear regression in the uncorrupted case. The following is similar to Lemma D.1 in [1].

Lemma D.2. *Suppose we observe i.i.d. gradient samples $\{\boldsymbol{g}_i, i \in \mathcal{G}\}$ from Model 1.1 with $|\mathcal{G}| = \Omega\left(\frac{k\log(d/\nu)}{\epsilon^2}\right)$. Then, there is a $\delta = \widetilde{O}(\epsilon)$, such that with probability at least $1-\nu$, for any index subset $\mathcal{J} \subset [d]$, $|\mathcal{J}| \le \widetilde{k}$ and for any $\mathcal{G}' \subset \mathcal{G}$, $|\mathcal{G}'| \ge (1-2\epsilon)|\mathcal{G}|$, the following inequalities hold:*

$$\left\|\mathbb{E}_{i\in_u\mathcal{G}'}\left(\boldsymbol{g}_i^{\mathcal{J}}\right) - \boldsymbol{G}^{\mathcal{J}}\right\|_2 \le \delta\left(\|\boldsymbol{G}\|_2 + \sigma\right), \tag{13}$$

$$\left\|\mathbb{E}_{i\in_u\mathcal{G}'}\left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}}\right)^{\otimes 2} - F\left(\boldsymbol{G}\right)^{\mathcal{J}\mathcal{J}}\right\|_{\text{op}} \le \delta\left(\|\boldsymbol{G}\|_2^2 + \sigma^2\right). \tag{14}$$

Proof. The main difference from their Lemma D.1 is that we consider a uniform distribution over all samples instead of a balanced weighted distribution. Furthermore, eqs. (13) and (14) are the concentration inequalities for the mean and covariance of the collected gradient samples $\{\boldsymbol{g}_i, i \in \mathcal{G}\}$ in

the good set with the form:

$$\boldsymbol{g}_i = \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{G} - \boldsymbol{x}_i \xi_i,$$

which is equivalent to their Lemma D.1, where they consider $y_i \boldsymbol{x}_i = \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{x}_i \xi_i$. Therefore, by setting all weights to $\frac{1}{(1-2\epsilon)|\mathcal{G}|}$ in their Lemma D.1 we obtain the desired concentration properties. □

D.3 Relationship between the first and second moment of samples in $\mathcal{S}_{\text{in}}$

In this part, we show an important connection between the covariance deviation (the empirical covariance of $\mathcal{S}_{\text{in}}$ minus the true covariance of authentic data) and the mean deviation (the empirical mean of $\mathcal{S}_{\text{in}}$ minus the true mean of authentic data). When the mean deviation (in ℓ_2 sense) is large, the following Lemma implies that the covariance deviation must also be large. As a result, when the magnitude of the covariance deviation is large, the current set of samples (or the current weights of all samples) needs to be adjusted; when the magnitude of the covariance deviation is small, the average of current sample set (or the weighted sum of samples using current weights) provides a good enough estimate of the model parameter. Moreover, the same principle holds when we use an approximation of the true covariance, which can be efficiently estimated.

Unlike Lemma A.2 in [1], in eq. (17), eq. (18), we provide lower bounds for the $\widetilde{k}$-sparse largest eigenvalue (rigorous definition in eq. (20)), instead of the $\widetilde{k}$-sparse operator norm. As we discussed in Appendix B, λ^* is the convex relaxation of finding the $\widetilde{k}$-sparse largest eigenvalue (instead of the $\widetilde{k}$-sparse operator norm). In the statement of the following Lemma, for the purpose of consistency, we consider the uniform distribution of weights. However, the proof and results can be easily extended to the setting with the balanced distribution of weights. This is due to the similarity between the two types of weight representation, as discussed in Appendix D.1.

Lemma D.3. *Suppose $|\mathcal{S}_{\text{bad}}| \leq 2\epsilon |\mathcal{S}_{\text{in}}|$, $\delta = \Omega(\epsilon)$, and the gradient samples in $\mathcal{S}_{\text{good}}$ satisfy*

$$\left\| \mathsf{P}_{\widetilde{k}} \left(\widetilde{\Delta}_{\mathcal{S}_{\text{good}}} \right) \right\|_2 \leq c \left(\|\boldsymbol{G}\|_2 + \sigma \right) \delta, \tag{15}$$

$$\left\| \mathbb{E}_{i \in_u \mathcal{S}_{\text{good}}} \left(\boldsymbol{g}_i - \boldsymbol{G} \right)^{\otimes 2} - F(\boldsymbol{G}) \right\|_{\widetilde{\text{k}},\text{op}} \leq c \left(\|\boldsymbol{G}\|_2^2 + \sigma^2 \right) \delta, \tag{16}$$

where c is a constant. If $\left\| \mathsf{P}_{\widetilde{k}} \left(\widetilde{\Delta} \right) \right\|_2 \geq C_1 \left(\|\boldsymbol{G}\|_2 + \sigma \right) \delta$, where C_1 is a large constant, we have,

$$\max_{\|\boldsymbol{v}\|_2 = 1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} \boldsymbol{v}^\top \left(\mathbb{E}_{i \in_u \mathcal{S}_{\text{in}}} \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}} \right)^{\otimes 2} - F(\boldsymbol{G}) \right) \boldsymbol{v} \geq \frac{\left\| \mathsf{P}_{\widetilde{k}} \left(\widetilde{\Delta} \right) \right\|_2^2}{4\epsilon}, \tag{17}$$

$$\max_{\|\boldsymbol{v}\|_2 = 1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} \boldsymbol{v}^\top \left(\mathbb{E}_{i \in_u \mathcal{S}_{\text{in}}} \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}} \right)^{\otimes 2} - F\left(\widehat{\boldsymbol{G}} \right) \right) \boldsymbol{v} \geq \frac{\left\| \mathsf{P}_{\widetilde{k}} \left(\widetilde{\Delta} \right) \right\|_2^2}{5\epsilon}. \tag{18}$$

Proof. We focus on the $\widetilde{k}$-sparse largest eigenvalue (rigorous definition in eq. (20)), which is the correct route of analysis the convex relaxation of Sparse PCA.

Let $\mathcal{J} = \arg\max_{\mathcal{J}' \subset [d], |\mathcal{J}'| \leq \widetilde{k}} \left\| \widetilde{\Delta}^{\mathcal{J}'} \right\|_2$. Then $\widetilde{\Delta}^{\mathcal{J}} = \left\| \mathsf{P}_{\widetilde{k}} \left(\widetilde{\Delta} \right) \right\|_2 \geq C_1 \left(\|\boldsymbol{G}\|_2 + \sigma \right) \delta$ according to the

assumption. Using $|\mathcal{S}_{\text{in}}|$ to denote the size of $\mathcal{S}_{\text{in}}$, we have a lower bound for the sum over bad samples:

$$\begin{aligned}
\left\| \frac{1}{|\mathcal{S}_{\text{in}}|} \sum_{i \in \mathcal{S}_{\text{bad}}} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}} \right) \right\|_2 &= \left\| \widetilde{\Delta}^{\mathcal{J}} - \frac{1}{|\mathcal{S}_{\text{in}}|} \sum_{i \in \mathcal{S}_{\text{good}}} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}} \right) \right\|_2 \\
&\geq \left\| \widetilde{\Delta}^{\mathcal{J}} \right\|_2 - \left\| \frac{1}{|\mathcal{S}_{\text{in}}|} \sum_{i \in \mathcal{S}_{\text{good}}} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}} \right) \right\|_2 \\
&\overset{(i)}{\geq} \left\| \widetilde{\Delta}^{\mathcal{J}} \right\|_2 - c \left(\|\boldsymbol{G}\|_2 + \sigma \right) \delta \\
&\overset{(ii)}{\geq} \frac{\left\| \widetilde{\Delta}^{\mathcal{J}} \right\|_2}{1.1},
\end{aligned}$$

where (i) follows from eq. (15) and the assumptions; (ii) follows from that we choose C_1 large enough.

By p.s.d.-ness of covariance matrices, we have

$$\frac{1}{|\mathcal{S}_{\text{bad}}|} \sum_{i \in \mathcal{S}_{\text{bad}}} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}} \right) \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}} \right)^\top \succcurlyeq \left(\frac{1}{|\mathcal{S}_{\text{bad}}|} \sum_{i \in \mathcal{S}_{\text{bad}}} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}} \right) \right)^{\otimes 2}.$$

Therefore, because $|\mathcal{S}_{\text{bad}}| \leq 2\epsilon |\mathcal{S}_{\text{in}}|$, we have

$$\left\| \frac{1}{|\mathcal{S}_{\text{in}}|} \sum_{i \in \mathcal{S}_{\text{bad}}} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}} \right)^{\otimes 2} \right\|_{\text{op}} \geq \frac{\left\| \frac{1}{|\mathcal{S}_{\text{in}}|} \sum_{i \in \mathcal{S}_{\text{bad}}} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}} \right) \right\|_2^2}{2\epsilon} \geq \frac{\left\| \widetilde{\Delta}^{\mathcal{J}} \right\|_2^2}{2.5\epsilon}. \tag{19}$$

With a lower bound of this submatrix of the covariance matrix, we define a vector $\boldsymbol{v}_0 \in \mathbb{R}^{\widetilde{k}}$ as follows:

$$\boldsymbol{v}_0 = \underset{\|\boldsymbol{v}\|_2 = 1}{\arg\max} \, \boldsymbol{v}^\top \left(\sum_{i \in \mathcal{S}_{\text{bad}}} \frac{1}{|\mathcal{S}_{\text{in}}|} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}} \right)^{\otimes 2} \right) \boldsymbol{v}. \tag{20}$$

For this $\boldsymbol{v}_0$, we have

$$\begin{aligned}
&\boldsymbol{v}_0^\top \left(\frac{1}{|\mathcal{S}_{\text{in}}|} \sum_{i=1}^{|\mathcal{S}_{\text{in}}|} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}} \right)^{\otimes 2} - F\left(\boldsymbol{G} \right)^{\mathcal{J}\mathcal{J}} \right) \boldsymbol{v}_0 \\
&\geq \boldsymbol{v}_0^\top \left(\frac{1}{|\mathcal{S}_{\text{in}}|} \sum_{i \in \mathcal{S}_{\text{bad}}} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}} \right)^{\otimes 2} \right) \boldsymbol{v}_0 \\
&- \left\| \frac{1}{|\mathcal{S}_{\text{in}}|} \sum_{i \in \mathcal{S}_{\text{good}}} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}} \right)^{\otimes 2} - \frac{|\mathcal{S}_{\text{good}}|}{|\mathcal{S}_{\text{in}}|} F\left(\boldsymbol{G} \right)^{\mathcal{J}\mathcal{J}} \right\|_{\text{op}} - \left\| \frac{|\mathcal{S}_{\text{bad}}|}{|\mathcal{S}_{\text{in}}|} F\left(\boldsymbol{G} \right)^{\mathcal{J}\mathcal{J}} \right\|_{\text{op}} \\
&\overset{(i)}{\geq} \frac{\left\| \widetilde{\Delta}^{\mathcal{J}} \right\|^2}{2.5\epsilon} - c \left(\|\boldsymbol{G}\|_2^2 + \sigma^2 \right) \delta - 2\epsilon (\|\boldsymbol{G}\|_2^2 + \sigma^2) \\
&\overset{(ii)}{\geq} \frac{\left\| \widetilde{\Delta}^{\mathcal{J}} \right\|^2}{3\epsilon},
\end{aligned} \tag{21}$$

where (i) follows from eq. (16) and eq. (19); (ii) follows from the assumption that ϵ is sufficiently small.

Applying eq. (21) on our target $\mathbb{E}_{i\in_u \mathcal{S}_{\text{in}}}\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)^{\otimes 2} - F\left(\boldsymbol{G}\right)$, we have

$$
\begin{aligned}
&\boldsymbol{v}_0^\top \left(\frac{1}{|\mathcal{S}_{\text{in}}|} \sum_{i=1}^{|\mathcal{S}_{\text{in}}|} \left(\boldsymbol{g}_i^{\mathcal{J}} - \widehat{\boldsymbol{G}}^{\mathcal{J}}\right)^{\otimes 2} - F\left(\boldsymbol{G}\right)^{\mathcal{J}\mathcal{J}} \right) \boldsymbol{v}_0 \\
&= \boldsymbol{v}_0^\top \left(\frac{1}{|\mathcal{S}_{\text{in}}|} \sum_{i=1}^{|\mathcal{S}_{\text{in}}|} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}}\right)^{\otimes 2} - F\left(\boldsymbol{G}\right)^{\mathcal{J}\mathcal{J}} - \widehat{\Delta}^{\mathcal{J}} \left(\widetilde{\Delta}^{\mathcal{J}}\right)^\top - \widetilde{\Delta}^{\mathcal{J}} \left(\widehat{\Delta}^{\mathcal{J}}\right)^\top + \left(\widehat{\Delta}^{\mathcal{J}}\right)^{\otimes 2} \right) \boldsymbol{v}_0 \\
&\overset{(i)}{\geq} \boldsymbol{v}_0^\top \left(\frac{1}{|\mathcal{S}_{\text{in}}|} \sum_{i=1}^{|\mathcal{S}_{\text{in}}|} \left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}}\right)^{\otimes 2} - F\left(\boldsymbol{G}\right)^{\mathcal{J}\mathcal{J}} \right) \boldsymbol{v}_0 - 24 \left(\left\| \widetilde{\Delta}^{\mathcal{J}} \right\|_2^2 \right) \\
&\overset{(ii)}{\geq} \frac{\left\| \widetilde{\Delta}^{\mathcal{J}} \right\|_2^2}{4\epsilon},
\end{aligned} \tag{22}
$$

where (i) follows from Lemma D.1; (ii) follows from eq. (21) and ϵ is sufficiently small. By a construction $\boldsymbol{v} = (\boldsymbol{v}_0, \boldsymbol{0}_{d-\widetilde{k}})^\top$, it is easy to see that $\boldsymbol{v}_0$ provides a lower bound for the maximum of $\{\boldsymbol{v} : \|\boldsymbol{v}\|_2 = 1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}\}$ in eq. (17).

By eq. (22), we already know that

$$
\boldsymbol{v}_0^\top \left(\frac{1}{|\mathcal{S}_{\text{in}}|} \sum_{i=1}^{|\mathcal{S}_{\text{in}}|} \left(\boldsymbol{g}_i^{\mathcal{J}} - \widehat{\boldsymbol{G}}^{\mathcal{J}}\right)^{\otimes 2} - F\left(\boldsymbol{G}\right)^{\mathcal{J}\mathcal{J}} \right) \boldsymbol{v}_0 \geq \frac{\left\| \widetilde{\Delta}^{\mathcal{J}} \right\|_2^2}{4\epsilon}.
$$

By our assumptions on F, we have

$$
\begin{aligned}
\left\| F\left(\boldsymbol{G}\right) - F\left(\widehat{\boldsymbol{G}}\right) \right\|_{\widetilde{\mathrm{k}},\mathrm{op}} &\leq L_F \left\| \widehat{\Delta} \right\|_2 + C \left\| \widehat{\Delta} \right\|_2^2 \\
&\overset{(i)}{\leq} 5L_F \left\| \widetilde{\Delta}^{\mathcal{J}} \right\|_2 + 5C \left\| \widetilde{\Delta}^{\mathcal{J}} \right\|_2^2,
\end{aligned}
$$

where (i) follows from Lemma D.1. Since $\delta = \Omega\left(\epsilon\right)$, we obtain eq. (18) by using the triangle inequality. □

D.4 Proof of Corollary 3.1

Equipped with Lemma D.1, Lemma D.2 and Lemma D.3, we can now prove Corollary 3.1.

Corollary D.1 (Corollary 3.1). *Suppose we observe $N\left(k, d, \epsilon, \nu\right)$ 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 $\rho_{\text{sep}} = \Theta\big(\epsilon\big(\|\boldsymbol{G}^t\|_2^2 + \sigma^2\big)\big)$, it requires $N\left(k, d, \epsilon, \nu\right) = \Omega\big(\frac{k^2 \log(dT/\nu)}{\epsilon^2}\big)T$ samples, and guarantees $\psi\left(\epsilon\right) = \widetilde{O}\left(\epsilon^2\sigma^2\right)$. Hence, Algorithm 1 outputs $\widehat{\boldsymbol{\beta}}$, such that*

$$
\left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|_2 = \widetilde{O}\left(\sigma\epsilon\right),
$$

with probability at least $1 - \nu$, by setting $T = \Theta\left(\log\left(\frac{\|\boldsymbol{\beta}^*\|_2}{\epsilon\sigma}\right)\right)$.

Proof. We consider only the t-th iteration, and thus omit t in $\boldsymbol{g}_i^t$ and $\boldsymbol{G}^t$. The function $F(\boldsymbol{G})$ is given by $F(\boldsymbol{G}) = \|\boldsymbol{G}\|_2^2 \boldsymbol{I}_d + \boldsymbol{G}\boldsymbol{G}^\top + \sigma^2 \boldsymbol{I}_d$, as in Appendix C. The accuracy in robust sparse estimation on gradients depends on two parameters for $F(\boldsymbol{G})$: $L_{\text{cov}} = 2\|\boldsymbol{G}\|_2^2 + \sigma^2$, and $L_{\text{F}} = 4\|\boldsymbol{G}\|_2$, which are calculated in Appendix C.

Under the statistical model and the contamination model described in Theorem 2.1, we can set the parameters $\rho_{\text{sep}} = \Theta(\epsilon(\|\boldsymbol{G}^t\|_2^2 + \sigma^2))$ in Algorithm 2 by the calculation of L_{cov} and L_{F}

The ellipsoid algorithm considers all possible sample weights in a convex set and finds the optimal weight for each sample. The algorithm iteratively uses a separation oracle Algorithm 2, which solves the convex relaxation of Sparse PCA at each iteration:

$$\lambda^* = \max_{\boldsymbol{H}} \operatorname{Tr}\left(\left(\widehat{\boldsymbol{\Sigma}} - F\left(\widehat{\boldsymbol{G}}\right)\right) \cdot \boldsymbol{H}\right), \quad \text{subject to } \boldsymbol{H} \succcurlyeq 0, \|\boldsymbol{H}\|_{1,1} \leq \widetilde{k}, \operatorname{Tr}(\boldsymbol{H}) = 1. \tag{23}$$

To prove the Main Theorem (Theorem 3.1) in [1], the only modification is to replace the lower bound of λ^* in their Lemma A.3.

A weighted version of Lemma D.3 implies that if the mean deviation is large, then

$$\max_{\|\boldsymbol{v}\|_2=1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} \boldsymbol{v}^\top \left(\sum_{i=1}^{|\mathcal{S}_{\text{in}}|} w_i \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}(w)\right)^{\otimes 2} - F\left(\widehat{\boldsymbol{G}}(w)\right)\right) \boldsymbol{v} \geq \frac{\left\|\mathsf{P}_{\widetilde{k}}\left(\widetilde{\Delta}(w)\right)\right\|_2^2}{5\epsilon}, \tag{24}$$

where $\widehat{\boldsymbol{G}}(w) = \mathsf{P}_{2\widetilde{k}}\left(\sum_{i=1}^{|\mathcal{S}_{\text{in}}|} w_i \boldsymbol{g}_i\right)$, and $\widetilde{\Delta}(w) = \sum_{i=1}^{|\mathcal{S}_{\text{in}}|} w_i \boldsymbol{g}_i - \boldsymbol{G}$. Then, λ^* in the ellipsoid algorithm satisfies

$$\lambda^* \geq \max_{\|\boldsymbol{v}\|_2=1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} \boldsymbol{v}^\top \left(\sum_{i=1}^{|\mathcal{S}_{\text{in}}|} w_i \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}(w)\right)^{\otimes 2} - F\left(\widehat{\boldsymbol{G}}(w)\right)\right) \boldsymbol{v}, \tag{25}$$

since λ^* is the solution to the trace norm maximization eq. (23), which is the convex relaxation of finding the $\widetilde{k}$-sparse largest eigenvalue.

Combining eq. (24) and eq. (25), we have

$$\lambda^* \geq \frac{\left\|\mathsf{P}_{\widetilde{k}}\left(\widetilde{\Delta}(w)\right)\right\|_2^2}{5\epsilon}, \tag{26}$$

which recovers the correctness of the separation oracle in the ellipsoid algorithm, and their Main Theorem (Theorem 3.1).

Finally, the ellipsoid algorithm guarantees that, with sample complexity $\Omega\left(\frac{k^2 \log(d/\nu)}{\epsilon^2}\right)$, the estimate $\widehat{\boldsymbol{G}}$ satisfies

$$\left\|\widehat{\boldsymbol{G}} - \boldsymbol{G}\right\|_2^2 = \widetilde{O}\left(\epsilon^2 \left(L_{\text{F}}^2 + L_{\text{cov}}\right)\right) = \widetilde{O}\left(\epsilon^2 \left(\|\boldsymbol{G}\|_2^2 + \sigma^2\right)\right), \tag{27}$$

with probability at least $1 - \nu$. This exactly gives us a $\widetilde{O}(\epsilon^2 \sigma^2)$-RSGE. Hence, we can apply eq. (27) as the RSGE in Theorem 2.1 to prove Corollary 3.1. □

E Outlier removal guarantees in the filtering algorithm

In this section, we consider a single iteration of Algorithm 1, and prove Lemma 4.1 at the t-th step. For clarity, we omit the superscript t in both $\boldsymbol{g}_i^t$ and $\boldsymbol{G}^t$.

In order to show guarantees for Lemma 4.1, we leverage previous results Lemma D.2 and Lemma D.3. We state Lemma E.1 as a modification of Lemma D.2 by replacing ϵ by $\sqrt{\epsilon}$, using concentration results in Lemma D.2, and replacing ϵ by $\sqrt{\epsilon}$. We state Lemma E.2 as a modification of Lemma D.3 by replacing $\delta = \Omega(\epsilon)$ with $\delta = \Omega(\sqrt{\epsilon})$, since the results for $\delta = \Omega(\epsilon)$ implies the results for $\delta = \Omega(\sqrt{\epsilon})$.

The reason we modify the above is to prove guarantees for our computationally more efficient RSGE described in Algorithm 3. Our motivation for calculating the score for each sample according to $\tau_i = \mathrm{Tr}(\boldsymbol{H}^* \cdot (\boldsymbol{g}_i - \widehat{\boldsymbol{G}})(\boldsymbol{g}_i - \widehat{\boldsymbol{G}})^\top)$ is to make sure that all the scores τ_i are positive (notice that the scores calculated based on the original non-p.s.d matrix may be negative). Based on this, we show that the sum of scores over all bad samples is a large constant (> 1) times larger than the sum of scores over all good samples. When finding an upper bound for $\sum_{i\in\mathcal{S}_{\text{good}}} \tau_i$, we compromise an ϵ factor in the value of λ^*, which results in an $\sqrt{\epsilon}$ factor in the recovery guarantee.

As described above, we immediately have Lemma E.1 and Lemma E.2 given the proofs in Appendix D. Note that we still use the same definitions $\widetilde{\Delta}_{\mathcal{S}_{\text{good}}}$ and $\widetilde{\Delta}$ on set $\mathcal{S}_{\text{good}}$ and $\mathcal{S}_{\text{in}}$ respectively as in Appendix D.1.

Lemma E.1. *Suppose we observe i.i.d. gradient samples $\{\boldsymbol{g}_i, i \in \mathcal{G}\}$ from Model 1.1 with $|\mathcal{G}| = \Omega\left(\frac{k\log(d/\nu)}{\epsilon}\right)$. Then there is a $\delta = \widetilde{O}(\sqrt{\epsilon})$ that with probability at least $1-\nu$, we have for any subset $\mathcal{J} \subset [d]$, $|\mathcal{J}| \le \widetilde{k}$, and for any $\mathcal{G}' \subset \mathcal{G}$, $|\mathcal{G}'| \ge (1-2\epsilon)|\mathcal{G}|$, the following inequalities hold:*

$$\left\|\mathbb{E}_{i\in_u\mathcal{G}'}\left(\boldsymbol{g}_i^{\mathcal{J}}\right) - \boldsymbol{G}^{\mathcal{J}}\right\|_2 \le \delta\left(\|\boldsymbol{G}\|_2 + \sigma\right), \tag{28}$$

$$\left\|\mathbb{E}_{i\in_u\mathcal{G}'}\left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}}\right)^{\otimes 2} - F\left(\boldsymbol{G}\right)^{\mathcal{J}\mathcal{J}}\right\|_{\text{op}} \le \delta\left(\|\boldsymbol{G}\|_2^2 + \sigma^2\right). \tag{29}$$

Lemma E.2. *Suppose $|\mathcal{S}_{\text{bad}}| \le 2\epsilon|\mathcal{S}_{\text{in}}|$, $\delta = \Omega(\sqrt{\epsilon})$, and the gradient samples in $\mathcal{S}_{\text{good}}$ satisfy*

$$\left\|\mathsf{P}_{\widetilde{k}}\left(\widetilde{\Delta}_{\mathcal{S}_{\text{good}}}\right)\right\|_2 \le c\left(\|\boldsymbol{G}\|_2 + \sigma\right)\delta, \tag{30}$$

$$\left\|\mathbb{E}_{i\in_u\mathcal{S}_{\text{good}}}\left(\boldsymbol{g}_i - \boldsymbol{G}\right)^{\otimes 2} - F\left(\boldsymbol{G}\right)\right\|_{\widetilde{k},\text{op}} \le c\left(\|\boldsymbol{G}\|_2^2 + \sigma^2\right)\delta, \tag{31}$$

where c is a constant. If $\left\|\mathsf{P}_{\widetilde{k}}\left(\widetilde{\Delta}\right)\right\|_2 \ge C_1\left(\|\boldsymbol{G}\|_2 + \sigma\right)\delta$, where C_1 is a constant. Then we have,

$$\max_{\|\boldsymbol{v}\|_2=1,\|\boldsymbol{v}\|_0\le\widetilde{k}} \boldsymbol{v}^\top\left(\mathbb{E}_{i\in_u\mathcal{S}_{\text{in}}}\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)^{\otimes 2} - F\left(\boldsymbol{G}\right)\right)\boldsymbol{v} \ge \frac{\left\|\mathsf{P}_{\widetilde{k}}\left(\widetilde{\Delta}\right)\right\|_2^2}{4\epsilon}, \tag{32}$$

$$\max_{\|\boldsymbol{v}\|_2=1,\|\boldsymbol{v}\|_0\le\widetilde{k}} \boldsymbol{v}^\top\left(\mathbb{E}_{i\in_u\mathcal{S}_{\text{in}}}\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)^{\otimes 2} - F\left(\widehat{\boldsymbol{G}}\right)\right)\boldsymbol{v} \ge \frac{\left\|\mathsf{P}_{\widetilde{k}}\left(\widetilde{\Delta}\right)\right\|_2^2}{5\epsilon}. \tag{33}$$

By Lemma E.1, eq. (30) and eq. (31) in Lemma E.2 are satisfied, provided that we have $|\mathcal{G}| = \Omega\left(\frac{k\log(d/\nu)}{\epsilon}\right)$. Now, equipped with Lemma E.1 and Lemma E.2, the effect of good samples can be controlled by concentration inequalities. Based on these, we are ready to prove Lemma 4.1.

Lemma E.3 (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 of gradient samples $\{\boldsymbol{g}_i^t\}_{i=1}^n$. Algorithm 3 computes λ^* that satisfies*

$$\lambda^* \geq \max_{\|\boldsymbol{v}\|_2=1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} \boldsymbol{v}^\top \left(\mathbb{E}_{i \in_u \mathcal{S}_{\text{in}}} \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}} \right)^{\otimes 2} \right) \boldsymbol{v}. \tag{34}$$

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

$$\sum_{i \in \mathcal{S}_{\text{good}}} \tau_i \leq \tfrac{1}{\gamma} \sum_{i \in \mathcal{S}_{\text{in}}} \tau_i, \tag{35}$$

where τ_i is defined in line 10, C_γ is a constant depending on γ, and $\gamma \geq 4$ is a constant.

Proof. Since λ^* is the solution of the convex relaxation of Sparse PCA, we have

$$\begin{aligned}
\lambda^* =& \operatorname{Tr}\left(H^* \cdot \left(\mathbb{E}_{i \in_u \mathcal{S}_{\text{in}}} \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}} \right)^{\otimes 2} \right) \right) \\
\geq & \max_{\|\boldsymbol{v}\|_2=1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} \boldsymbol{v}^\top \left(\mathbb{E}_{i \in_u \mathcal{S}_{\text{in}}} \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}} \right)^{\otimes 2} \right) \boldsymbol{v}.
\end{aligned}$$

By Theorem A.1 in [1], we have

$$\begin{aligned}
&\operatorname{Tr}\left(H^* \cdot \left(\mathbb{E}_{i \in_u \mathcal{S}_{\text{good}}} \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}} \right)^{\otimes 2} - F\left(\widehat{\boldsymbol{G}} \right) \right) \right) \\
&\leq C \left(\left\| \widehat{\Delta} \right\|_2^2 + \left(L_{\text{F}} + \widetilde{k} \left\| \widetilde{\Delta}_{\mathcal{S}_{\text{good}}} \right\|_\infty \right) \left\| \widehat{\Delta} \right\|_2 + \widetilde{k} \left\| \mathbb{E}_{i \in_u \mathcal{S}_{\text{good}}} \left(g_i - \boldsymbol{G} \right)^{\otimes 2} - F\left(\boldsymbol{G} \right) \right\|_\infty \right),
\end{aligned} \tag{36}$$

where C is a constant. Noticing that $\left\| \widetilde{\Delta}_{\mathcal{S}_{\text{good}}} \right\|_\infty$ and $\left\| \mathbb{E}_{i \in_u \mathcal{S}_{\text{good}}} \left(g_i - \boldsymbol{G} \right)^{\otimes 2} - F\left(\boldsymbol{G} \right) \right\|_\infty$ are unrelated to $\widehat{\boldsymbol{G}}$ and only defined on $\mathcal{S}_{\text{good}}$, [1] shows concentration bounds for these two terms, when $n = \Omega\left(\frac{\widetilde{k}^2 \log(d/\nu)}{\epsilon} \right)$. Specifically, it showed that with probability at least $1 - \nu$, we have

$$\left\| \widetilde{\Delta}_{\mathcal{S}_{\text{good}}} \right\|_\infty \leq C_1 \left(L_{\text{F}} + \sqrt{L_{\text{cov}}} \right) \sqrt{\epsilon} / \widetilde{k} \tag{37}$$

$$\left\| \mathbb{E}_{i \in_u \mathcal{S}_{\text{good}}} \left(g_i - \boldsymbol{G} \right)^{\otimes 2} - F\left(\boldsymbol{G} \right) \right\|_\infty \leq C_1 \left(L_{\text{F}}^2 + L_{\text{cov}} \right) \sqrt{\epsilon} / \widetilde{k} \tag{38}$$

Now, we focus on the LHS of eq. (35), the sum of scores of points in $\mathcal{S}_{\text{good}}$. By definition, we have

$$\begin{aligned}
&\mathbb{E}_{i \in_u \mathcal{S}_{\text{good}}} \tau_i \\
&= \operatorname{Tr}\left(H^* \cdot \left(\mathbb{E}_{i \in_u \mathcal{S}_{\text{good}}} \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}} \right)^{\otimes 2} \right) \right) \\
&= \operatorname{Tr}\left(H^* \cdot \left(\mathbb{E}_{i \in_u \mathcal{S}_{\text{good}}} \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}} \right)^{\otimes 2} - F\left(\widehat{\boldsymbol{G}} \right) \right) \right) + \operatorname{Tr}\left(H^* F\left(\widehat{\boldsymbol{G}} \right) \right) \\
&\overset{(i)}{\leq} C \left(\left\| \widehat{\Delta} \right\|_2^2 + \left(L_{\text{F}} + \widetilde{k} \left\| \widetilde{\Delta}_{\mathcal{S}_{\text{good}}} \right\|_\infty \right) \left\| \widehat{\Delta} \right\|_2 + \widetilde{k} \left\| \mathbb{E}_{i \in_u \mathcal{S}_{\text{good}}} \left(g_i - \boldsymbol{G} \right)^{\otimes 2} - F\left(\boldsymbol{G} \right) \right\|_\infty \right)
\end{aligned}$$

$$+ \operatorname{Tr}\left(H^* F\left(\widehat{\boldsymbol{G}}\right)\right),$$

where (i) follows from eq. (36).

To bound the RHS above, we first bound $\operatorname{Tr}\left(H^* \cdot F\left(\widehat{\boldsymbol{G}}\right)\right)$. Because of the constraint of the SDP given in eq. (3), H^* belongs to the Fantope $\mathcal{F}^1$ [44], and thus for any matrix A, we have $\operatorname{Tr}\left(A \cdot H^*\right) \leq \|A\|_{\mathrm{op}}$.

Thus, we have

$$\begin{aligned}
\operatorname{Tr}\left(H^* \cdot F\left(\widehat{\boldsymbol{G}}\right)\right) &= \operatorname{Tr}\left(H^* \cdot F\left(\boldsymbol{G}\right)\right) + \operatorname{Tr}\left(H^* * \left(F\left(\widehat{\boldsymbol{G}}\right) - F\left(\boldsymbol{G}\right)\right)\right) \\
&\leq \|F\left(\boldsymbol{G}\right)\|_{\mathrm{op}} + \left\|F\left(\widehat{\boldsymbol{G}}\right) - F\left(\boldsymbol{G}\right)\right\|_{\mathrm{op}} \\
&\overset{(i)}{\leq} C_1\left(\|\boldsymbol{G}\|_2^2 + \sigma^2\right) + \left\|F\left(\widehat{\boldsymbol{G}}\right) - F\left(\boldsymbol{G}\right)\right\|_{\mathrm{op}} \\
&\overset{(ii)}{\leq} C_1\left(\|\boldsymbol{G}\|_2^2 + \sigma^2\right) + L_{\mathrm{F}}\left\|\widehat{\Delta}\right\|_2 + C_2\left\|\widehat{\Delta}\right\|_2^2, \qquad (39)
\end{aligned}$$

where (i) follows from the expression of $F\left(G\right)$ in Appendix C; (ii) from the smoothness of $F\left(G\right)$.

By plugging in the concentration guarantees eq. (37) and combining eq. (39), we have

$$\begin{aligned}
&\mathbb{E}_{i\in_u \mathcal{S}_{\mathrm{good}}} \tau_i \\
&\leq C_2\left(\left(L_{\mathrm{F}}^2 + L_{\mathrm{cov}}\right)\sqrt{\epsilon} + \left(\left(L_{\mathrm{F}} + \sqrt{L_{\mathrm{cov}}}\right)\sqrt{\epsilon} + L_{\mathrm{F}}\right)\left\|\widehat{\Delta}\right\|_2 + \left\|\widehat{\Delta}\right\|_2^2\right) + C_1\left(\|\boldsymbol{G}\|_2^2 + \sigma^2\right) \\
&\overset{(i)}{\leq} C_2\left(\|\boldsymbol{G}\|_2\left\|\widehat{\Delta}\right\|_2 + \left\|\widehat{\Delta}\right\|_2^2\right) + C_1\left(\|\boldsymbol{G}\|_2^2 + \sigma^2\right) \\
&\leq C_1\left(\|\boldsymbol{G}\|_2\left\|\widehat{\Delta}\right\|_2 + \left\|\widehat{\Delta}\right\|_2^2 + \|\boldsymbol{G}\|_2^2 + \sigma^2\right), \qquad (40)
\end{aligned}$$

where (i) follows from the fact that ϵ is sufficiently small.

On the other hand, we know that: $\mathbb{E}_{i\in_u \mathcal{S}_{\mathrm{in}}} \tau_i = \lambda^*$.

Now, under the condition $\lambda^* \geq \rho_{\mathrm{sep}} = \Theta\left(\|\boldsymbol{G}\|_2^2 + \sigma^2\right)$, we consider two cases separately. By separating two cases, we can always show λ^* is very large, and the contribution from good samples is limited.

First, if $\|\widehat{\Delta}\|_2^2 \geq \Theta\left(\|\boldsymbol{G}\|_2^2 + \sigma^2\right)$, then in eq. (40), we have

$$\|\widehat{\Delta}\|_2^2 \gtrsim \|\boldsymbol{G}\|_2\|\widehat{\Delta}\|_2 \gtrsim \|\boldsymbol{G}\|_2^2, \quad \text{and } \|\widehat{\Delta}\|_2^2 \gtrsim \sigma^2.$$

Thus, we only need to compare λ^* and $\|\widehat{\Delta}\|_2^2$. By Lemma E.2, we have

$$\begin{aligned}
\mathbb{E}_{i\in_u \mathcal{S}_{\mathrm{in}}} \tau_i = \lambda^* &\geq \max_{\|\boldsymbol{v}\|_2=1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} \boldsymbol{v}^\top \left(\mathbb{E}_{i\in_u \mathcal{S}_{\mathrm{in}}}\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)^{\otimes 2}\right)\boldsymbol{v} \\
&\geq \max_{\|\boldsymbol{v}\|_2=1, \|\boldsymbol{v}\|_0 \leq \widetilde{k}} \boldsymbol{v}^\top \left(\mathbb{E}_{i\in_u \mathcal{S}_{\mathrm{in}}}\left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}}\right)^{\otimes 2} - F\left(\widehat{\boldsymbol{G}}\right)\right)\boldsymbol{v} \\
&\geq \frac{\left\|\widehat{\Delta}\right\|_2^2}{\epsilon}.
\end{aligned}$$

Hence, by eq. (40), we have $\mathbb{E}_{i\in_u\mathcal{S}_{\text{in}}}\tau_i \geq \gamma\,\mathbb{E}_{i\in_u\mathcal{S}_{\text{good}}}\tau_i$, where $\gamma \geq 4$ is a constant.

Second, if $\|\widehat{\Delta}\|_2^2 \leq \Theta\left(\|\boldsymbol{G}\|_2^2+\sigma^2\right)$, then in eq. (40), we have

$$\|\boldsymbol{G}\|_2^2 \gtrsim \|\boldsymbol{G}\|_2\|\widehat{\Delta}\|_2 \gtrsim \|\widehat{\Delta}\|_2^2, \quad \text{or } \sigma^2 \gtrsim \|\widehat{\Delta}\|_2^2.$$

Thus, we only need to compare λ^* and $\max\left(\|\boldsymbol{G}\|_2^2,\sigma^2\right)$. Since $\lambda^* \geq C_\gamma\left(\|\boldsymbol{G}\|_2^2+\sigma^2\right)$ by the condition of Lemma 4.1, we still have $\mathbb{E}_{i\in_u\mathcal{S}_{\text{in}}}\tau_i \geq \gamma\,\mathbb{E}_{i\in_u\mathcal{S}_{\text{good}}}\tau_i$, where $\gamma \geq 4$ is a constant.

Combing all of above, and setting $\rho_{\text{sep}} = C_\gamma\left(\|\boldsymbol{G}\|_2^2+\sigma^2\right)$, we have

$$\sum_{i\in\mathcal{S}_{\text{in}}}\tau_i = |\mathcal{S}_{\text{in}}|\,\mathbb{E}_{i\in_u\mathcal{S}_{\text{in}}}\tau_i \geq \gamma|\mathcal{S}_{\text{good}}|\,\mathbb{E}_{i\in_u\mathcal{S}_{\text{good}}}\tau_i = \gamma\sum_{i\in\mathcal{S}_{\text{good}}}\tau_i.$$

□

F RSGE via the filtering algorithm

In this section, we still consider the t-th iteration of Algorithm 1 and prove Theorem 4.1 on t. We omit t in $\boldsymbol{g}_i^t$ and $\boldsymbol{G}^t$.

In the case of $\lambda^* \geq C_\gamma\left(\|\boldsymbol{G}\|_2^2+\sigma^2\right)$, Algorithm 3 iteratively removes one sample according to the probability distribution eq. (4). We denote the steps of this outlier removal procedure as $l = 1, 2, \cdots, n$. The first step of proving Theorem 4.1 is to show we can remove a corrupted samples with high probability at each step, which is a result by Lemma 4.1.

Intuitively, if all subsequent steps are i.i.d., we can expect Algorithm 3 to remove outliers within around ϵn iterations, with exponentially high probability. However, the subsequent steps in Algorithm 3 are not independent. To circumvent this challenge we appeal to a martingale argument.

F.1 Supermartingale construction

Let $\mathcal{F}^l$ be the filtration generated by the set of events until iteration l of Algorithm 3. We define the corresponding set $\mathcal{S}_{\text{in}}^l$, $\mathcal{S}_{\text{good}}^l$ and $\mathcal{S}_{\text{bad}}^l$ at the step l. We have that $\mathcal{S}_{\text{in}}^l, \mathcal{S}_{\text{good}}^l, \mathcal{S}_{\text{bad}}^l \in \mathcal{F}^l$, and $|\mathcal{S}_{\text{in}}^l| = n - l$.

We denote a good event $\mathcal{E}^l$ at step l as

$$\sum_{i\in\mathcal{S}_{\text{bad}}^l}\tau_i \leq (\gamma-1)\sum_{i\in\mathcal{S}_{\text{good}}^l}\tau_i.$$

Then, by the definition of Algorithm 3 and Lemma 4.1, if $\lambda^* \geq C_\gamma\left(\|\boldsymbol{G}\|_2^2+\sigma^2\right)$, $\mathcal{E}^l$ is not true; if $\mathcal{E}^l$ is true, then Algorithm 3 will return a $\widehat{\boldsymbol{G}}$.

In Lemma F.1, we show that at any step l when $\mathcal{E}^l$ is not true, the random outlier removal procedure removes a corrupted sample with probability at least $(\gamma-1)/\gamma$.

Lemma F.1. *In each subsequent step l, if $\mathcal{E}^l$ is not true, then we can remove one remaining outlier*

from $\mathcal{S}_{\text{in}}^l$ *with probability at least* $(\gamma-1)/\gamma$*:*

$$\Pr\left(\textit{one sample from } \mathcal{S}_{\text{bad}}^l \textit{ is removed } |\mathcal{F}_l\right) \geq \frac{\gamma-1}{\gamma}.$$

Proof of Lemma F.1. When $\lambda^* \geq C_\gamma\left(\|\boldsymbol{G}\|_2^2+\sigma^2\right)$, Lemma 4.1 implies

$$\sum_{i\in\mathcal{S}_{\text{bad}}^l}\tau_i \geq (\gamma-1)\sum_{i\in\mathcal{S}_{\text{good}}^l}\tau_i.$$

Then we randomly remove a sample r from $\mathcal{S}_{\text{in}}$ according to

$$\Pr\left(\boldsymbol{g}_i \text{ is removed } |\mathcal{F}_l\right) = \frac{\tau_i}{\sum_{i\in\mathcal{S}_{\text{in}}^l}\tau_i}.$$

Finally,

$$\Pr\left(\text{one sample from } \mathcal{S}_{\text{bad}}^l \text{ is removed } |\mathcal{F}_l\right) = \sum_{i\in\mathcal{S}_{\text{bad}}^l}\frac{\tau_i}{\sum_{i\in\mathcal{S}_{\text{in}}^l}\tau_i} \geq \frac{\gamma-1}{\gamma}.$$

□

Since subsequent steps for applying Algorithm 3 on $\mathcal{S}_{\text{in}}$ are not independent, we need martingale arguments to show the total iterations of applying Algorithm 3 is limited.

We use the martingale technique in [46], by defining T: $T=\min\{l : \mathcal{E}^l \text{ is true}\}$. Based on T, we define a random variable:

$$Y^l = \begin{cases} |\mathcal{S}_{\text{bad}}^{T-1}| + \frac{\gamma-1}{\gamma}(T-1), & \text{if } l \geq T \\ |\mathcal{S}_{\text{bad}}^{l}| + \frac{\gamma-1}{\gamma}l, & \text{if } l < T \end{cases}$$

Lemma F.2 (Lemma 1 in [46])**.** $\{Y^l, \mathcal{F}^l\}$ *is a supermartingale.*

Now, equipped with Lemma F.1 and Lemma F.2, we are ready to prove Theorem 4.1.

F.2 Proof of Theorem 4.1

Theorem F.1 (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* $\mathcal{S}_{\text{in}}$ *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* $\mathcal{S}_{\text{in}}$*, and subsequently on* $\mathcal{S}_{\text{out}}$*, and use* $\rho_{\text{sep}}=C_\gamma\left(\|\boldsymbol{G}^t\|_2^2+\sigma^2\right)$*, 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*

$$\left\|\widehat{\boldsymbol{G}}^t-\boldsymbol{G}^t\right\|_2^2 = \widetilde{O}\left(\epsilon\left(\left\|\boldsymbol{G}^t\right\|_2^2+\sigma^2\right)\right),$$

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.*

Proof. We analyze Algorithm 3 by discussing a series of $\{\mathcal{E}^l\}$.

If $\mathcal{E}^l$ is true, then $\lambda^* \leq \rho_{\text{sep}} = C_\gamma \left(\|\boldsymbol{G}\|_2^2 + \sigma^2 \right)$. By Lemma E.2, we have

$$\lambda^* \geq \max_{\|\boldsymbol{v}\|_2=1, \|\boldsymbol{v}\|_0 \leq k} \boldsymbol{v}^\top \left(\mathbb{E}_{i \in_u \mathcal{S}_{\text{in}}} \left(\boldsymbol{g}_i - \widehat{\boldsymbol{G}} \right)^{\otimes 2} - F\left(\widehat{\boldsymbol{G}} \right) \right) \boldsymbol{v} \geq \frac{\left\| \mathsf{P}_{\widetilde{k}} \left(\widetilde{\Delta}_S \right) \right\|_2^2}{5\epsilon}.$$

Plugging in $\lambda^* \leq C_\gamma \left(\|\boldsymbol{G}\|_2^2 + \sigma^2 \right)$, we have

$$\frac{1}{5} \left\| \widehat{\Delta}_S \right\|_2^2 \overset{(i)}{\leq} \left\| \mathsf{P}_{\widetilde{k}} \left(\widetilde{\Delta}_S \right) \right\|_2^2 \leq 5\epsilon\lambda^* \leq O\left(\epsilon \left(\|\boldsymbol{G}\|_2^2 + \sigma^2 \right) \right),$$

where (i) follows from Lemma D.1. Hence, when $\mathcal{E}^l$ is true, Algorithm 3 can return a $\widehat{\boldsymbol{G}}$, such that $\left\| \widehat{\boldsymbol{G}} - \boldsymbol{G} \right\|_2^2 \leq O\left(\epsilon \left(\|\boldsymbol{G}\|_2^2 + \sigma^2 \right) \right)$.

Then, we only need to show $\bigcup_{l=1}^{L} \mathcal{E}^l$ is true, where $L = \frac{1.1\gamma}{\gamma - 1}\epsilon n$, with high probability. That said, we need to upper bound the probability

$$\Pr\left(\bigcap_{l=1}^{L} \overline{\mathcal{E}^l} \right) = \Pr\left(T \geq L \right) \leq \Pr\left(Y^L \geq \frac{\gamma - 1}{\gamma} L \right) = \Pr\left(Y^L \geq 1.1\epsilon n \right). \tag{41}$$

Then, we can construct the martingale difference according to [46]. Let $D^l = Y^l - Y^{l-1}$, where $Y^0 = \epsilon n$, and

$$\bar{D}^l = D^l - \mathbb{E}\left(D^l | D^1, \cdots, D^{l-1} \right).$$

Thus $\{\bar{D}^l\}$ is a martingale difference process, and $\mathbb{E}\left(D^l | D^1, \cdots, D^{l-1} \right) \leq 0$, since $\{Y^l\}$ is a super-martingale. Now, eq. (41) can be viewed as a bound for the sum of the associated martingale difference sequence.

$$Y^l - Y^0 = \sum_{j=1}^{l} D^j = \sum_{j=1}^{l} \bar{D}^j + \sum_{j=1}^{l} \mathbb{E}\left(D^j | D^1, \cdots, D^{j-1} \right) \leq \sum_{j=1}^{l} \bar{D}^j.$$

Since we only remove one example from the set $\mathcal{S}_{\text{in}}^l$, we can guarantee $|D^l| \leq 1$ and $|\bar{D}^l| \leq 2$. For these bounded random variables, by applying the Azuma-Hoeffding inequality, we have

$$\begin{aligned} \Pr\left(Y^L \geq 1.1\epsilon n \right) &\leq \Pr\left(\sum_{l=1}^{L} \bar{D}^l \geq 0.1\epsilon n \right) \\ &\leq \exp\left(\frac{-\left(0.1\epsilon n\right)^2}{8L} \right). \end{aligned}$$

Plugging in $L = \frac{1.1\gamma}{\gamma - 1}\epsilon n$, this probability is upper bounded by $\exp\left(-\Theta\left(\epsilon n\right)\right)$.

Notice that $L = \frac{1.1\gamma}{\gamma - 1}\epsilon n \leq 1.5\epsilon n$, by setting $\gamma \geq 4$. Hence, from $l = 1$ to L, we always have $|\mathcal{S}_{\text{bad}}^l| \leq 2\epsilon |\mathcal{S}_{\text{in}}^l|$. Then Lemma E.1 and Lemma E.2 hold and Lemma 4.1 is still valid.

Combining all of the above, we have proven that, with exponentially high probability, Algorithm 3

returns a $\widehat{\boldsymbol{G}}$ satisfying $\left\|\widehat{\boldsymbol{G}}-\boldsymbol{G}\right\|_2^2 \leq O\left(\epsilon\left(\|\boldsymbol{G}\|_2^2+\sigma^2\right)\right)$, within $\frac{1.1\gamma}{\gamma-1}\epsilon n$ iterations.

□

G Robust sparse regression with unknown covariance

In this section, we prove the guarantees for RSGE when the covariance matrix $\boldsymbol{\Sigma}$ is unknown, but each row and column is sparse. In this case, the population mean of all authentic gradients $\boldsymbol{G}^t$ can be calculated as

$$\boldsymbol{G}^t=\mathbb{E}_P\left(\boldsymbol{g}_i^t\right)=\mathbb{E}_P\left(\boldsymbol{x}_i\boldsymbol{x}_i^\top\left(\boldsymbol{\beta}^t-\boldsymbol{\beta}^*\right)\right)=\boldsymbol{\Sigma}\omega^t.$$

Therefore, $\boldsymbol{G}^t=\boldsymbol{\Sigma}\omega^t$ is guaranteed to be $r(k'+k)$ sparse. And we use the filtering algorithm (Algorithm 3) with $\widetilde{k}=r(k'+k)$ as a RSGE.

First, we derive the functional $F\left(\boldsymbol{G}\right)$ with general covariance matrix $\boldsymbol{\Sigma}$, and compute the corresponding $L_{\mathrm{F}}, L_{\mathrm{cov}}$, which has been defined in eq. (11) and eq. (12) for the case $\boldsymbol{\Sigma}=\boldsymbol{I}_d$ in Appendix C.

Lemma G.1. *Suppose we observe i.i.d. samples $\{\boldsymbol{z}_i, i\in\mathcal{G}\}$ from the distribution P in Model 1.1 with an unknown $\boldsymbol{\Sigma}$, we have the covariance of gradient as*

$$\mathrm{Cov}(\boldsymbol{g}):=\mathbb{E}_{\boldsymbol{z}_i\sim P}\left(\left(\boldsymbol{g}_i-\boldsymbol{G}\right)\left(\boldsymbol{g}_i-\boldsymbol{G}\right)^\top\right)=\boldsymbol{\Sigma}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{G}\right\|_2^2+\boldsymbol{G}\boldsymbol{G}^\top+\sigma^2\boldsymbol{\Sigma}.$$

Proof. As in the Model 1.1, we draw $\boldsymbol{x}$ from Gaussian distribution $\mathcal{N}(0,\boldsymbol{\Sigma})$, the expression of $F(\cdot)$ is given by

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{g}) &= \mathbb{E}\left(\left(\boldsymbol{g}_i-\boldsymbol{G}\right)\left(\boldsymbol{g}_i-\boldsymbol{G}\right)^\top\right)\\
&= \mathbb{E}\left(\left(\boldsymbol{x}\boldsymbol{x}^\top-\boldsymbol{\Sigma}\right)\omega\omega^\top\left(\boldsymbol{x}\boldsymbol{x}^\top-\boldsymbol{\Sigma}\right)\right)+\sigma^2\boldsymbol{\Sigma}\\
&\overset{(i)}{=} \mathbb{E}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^\top-\boldsymbol{I}_d\right)\boldsymbol{\Sigma}^{\frac{1}{2}}\omega\omega^\top\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^\top-\boldsymbol{I}_d\right)\boldsymbol{\Sigma}^{\frac{1}{2}}\right)+\sigma^2\boldsymbol{\Sigma}\\
&\overset{(ii)}{=} \boldsymbol{\Sigma}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{G}\right\|_2^2+\boldsymbol{G}\boldsymbol{G}^\top+\sigma^2\boldsymbol{\Sigma}.
\end{aligned}$$

where (i) follows from the re-parameterization $\boldsymbol{x}=\boldsymbol{\Sigma}^{\frac{1}{2}}\widetilde{\boldsymbol{x}}$, where $\widetilde{\boldsymbol{x}}\sim\mathcal{N}(0,\boldsymbol{I}_d)$, and (ii) follows from the Stein-type Lemma as in Appendix C. □

By Lemma G.1, we define the functional $F(\boldsymbol{G})=\boldsymbol{\Sigma}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{G}\right\|_2^2+\boldsymbol{G}\boldsymbol{G}^\top+\sigma^2\boldsymbol{\Sigma}$. In Algorithm 3, we do not need to evaluate $F(\cdot)$, but our analysis requires upper bounds for two parameters of $F(\cdot)$ – $L_{\mathrm{cov}}, L_{\mathrm{F}}$ – to control tail bounds. Under the same setting as Lemma G.1, we use similar bounds as Appendix C, based on assumptions in Model 1.1. Hence, we have $L_{\mathrm{cov}}=\Theta(\|\boldsymbol{G}\|_2^2+\sigma^2)$, and $L_{\mathrm{F}}=\Theta(\|\boldsymbol{G}\|_2)$.

Next, we show concentration bounds (Lemma G.2) similar to Lemma E.1, which controls deviation of empirical mean and covariance for all samples in the good set $\mathcal{G}$.

Lemma G.2. *Suppose we observe i.i.d. gradient samples $\{\boldsymbol{g}_i, i\in\mathcal{G}\}$ from Model 1.1 with $|\mathcal{G}|=\widetilde{\Omega}\left(\frac{\widetilde{k}\log(d/\nu)}{\epsilon}\right)$. Then, there is a $\delta=\widetilde{O}\left(\sqrt{\epsilon}\right)$, such that with probability at least $1-\nu$, for any index*

subset $\mathcal{J} \subset [d]$, $|\mathcal{J}| \leq \widetilde{k}$ *and for any* $\mathcal{G}' \subset \mathcal{G}$, $|\mathcal{G}'| \geq (1-2\epsilon)|\mathcal{G}|$, *we have*

$$\left\|\mathbb{E}_{i\in_u\mathcal{G}'}\left(\boldsymbol{g}_i^{\mathcal{J}}\right) - \boldsymbol{G}^{\mathcal{J}}\right\|_2 \leq \delta\left(\|\boldsymbol{G}\|_2 + \sigma\right), \tag{42}$$

$$\left\|\mathbb{E}_{i\in_u\mathcal{G}'}\left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}}\right)^{\otimes 2} - F\left(\boldsymbol{G}\right)^{\mathcal{J}\mathcal{J}}\right\|_{\text{op}} \leq \delta\left(\|\boldsymbol{G}\|_2^2 + \sigma^2\right). \tag{43}$$

Proof. We prove the concentration inequality for the covariance eq. (43), the bound for mean eq. (42) is similar. For any index subset $\mathcal{J} \subset [d]$, $|\mathcal{J}| \leq \widetilde{k}$, we can expand eq. (43) as follows,

$$\begin{aligned}
&\mathbb{E}_{i\in_u\mathcal{G}'}\left(\boldsymbol{g}_i^{\mathcal{J}} - \boldsymbol{G}^{\mathcal{J}}\right)^{\otimes 2} - F\left(\boldsymbol{G}\right)^{\mathcal{J}\mathcal{J}} \\
&= \mathbb{E}_{i\in_u\mathcal{G}'}\left(\boldsymbol{x}^{\mathcal{J}}\boldsymbol{x}^\top\omega\omega^\top\boldsymbol{x}(\boldsymbol{x}^{\mathcal{J}})^\top\right) - \left(\boldsymbol{\Sigma}^{\mathcal{J}\mathcal{J}}\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\omega\right\|_2^2 + 2\boldsymbol{G}^{\mathcal{J}}(\boldsymbol{G}^{\mathcal{J}})^\top\right) && (44)\\
&- \mathbb{E}_{i\in_u\mathcal{G}'}\left(\boldsymbol{x}\boldsymbol{x}^\top\omega\omega^\top\boldsymbol{\Sigma}\right)^{\mathcal{J}\mathcal{J}} + \boldsymbol{G}^{\mathcal{J}}(\boldsymbol{G}^{\mathcal{J}})^\top && (45)\\
&+ \mathbb{E}_{i\in_u\mathcal{G}'}\xi_i^2\boldsymbol{x}^{\mathcal{J}}(\boldsymbol{x}^{\mathcal{J}})^\top - \sigma^2\boldsymbol{\Sigma}^{\mathcal{J}\mathcal{J}} && (46)
\end{aligned}$$

Here, we prove the concentration inequality for eq. (44), and the other two terms can be bounded by the same technique. It is sufficient to prove an upper bound for the operator norm as follows

$$\left\|\mathbb{E}_{i\in_u\mathcal{G}'}\boldsymbol{x}^{\mathcal{J}}(\boldsymbol{x}^{\mathcal{J}})^\top\omega^{\mathcal{J}}(\omega^{\mathcal{J}})^\top\boldsymbol{x}^{\mathcal{J}}(\boldsymbol{x}^{\mathcal{J}})^\top - \left(\boldsymbol{\Sigma}^{\mathcal{J}\mathcal{J}}\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\omega\right\|_2^2 + 2\boldsymbol{G}^{\mathcal{J}}(\boldsymbol{G}^{\mathcal{J}})^\top\right)\right\|_{\text{op}} \leq \delta\|\boldsymbol{G}\|_2^2, \tag{47}$$

where $\boldsymbol{x}$ is drawn from a Gaussian distribution $\mathcal{N}(0, \boldsymbol{\Sigma})$. Note that the index subset $\mathcal{J}$ reduce the matrix to $\mathbb{R}^{|\mathcal{J}|\times|\mathcal{J}|}$. For the concentration bounds of covariance matrix estimation eq. (47), we have a near identical argument as Lemma 4.5 of [13], by replacing Theorem 5.50 with Theorem 5.44 in [43].

This establishes eq. (47) with sample complexity $n = \widetilde{\Omega}\left(\frac{\widetilde{k}\log(1/\nu)}{\epsilon}\right)$, with probability at least $1-\nu$. Next, we take a union bound over all possible subsets $\mathcal{J} \subset [d]$, and this gives concentration results for the covariance eq. (43). Hence we have proved the concentration results for the gradient under the assumption that $\boldsymbol{\Sigma}$ is row/column sparse. □

Based on Lemma G.2, we have Theorem 5.1, which guarantees the recovery of $\boldsymbol{\beta}^*$ in robust sparse regression with unknown covariance as defined in Model 5.1.

Corollary G.1 (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*

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2 = \widetilde{O}\left(\sigma\sqrt{\epsilon}\right), \tag{48}$$

with probability at least $1 - \nu - T\exp\left(-\Theta\left(\epsilon n\right)\right)$.

Proof. With the concentration result Lemma G.2 in hand, the remaining parts share the same theoretical analysis as Appendix E and Appendix F, by replacing $(k'+k)^2$ with $r^2(k'+k)^2 = \Theta(r^2k^2)$. Hence, we have a result similar to Corollary 4.1, with sample complexity $\widetilde{\Omega}\left(\frac{r^2k^2\log(dT/\nu)}{\epsilon}\right)$. And this yields Theorem 5.1. □

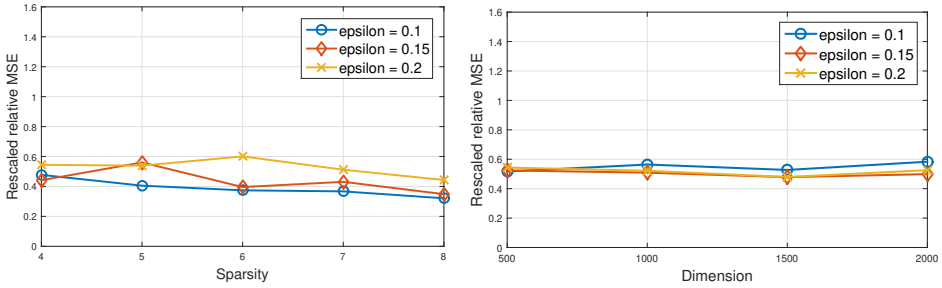

Figure 1: Simulations for Algorithm 3 showing the dependence of relative MSE on sparsity and dimension. For each parameter, we choose corresponding sample complexity $n \propto k^2 \log(d)/\epsilon$. Different curves for $\epsilon \in \{0.1, 0.15, 0.2\}$ are the average of 15 trials. Consistent with 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.

H Numerical results

H.1 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 $\boldsymbol{G}$. Each entry of $\boldsymbol{G}$ is either $+1$ or -1, hence $\|\boldsymbol{G}\|_2^2 = k$.

The outliers are specially designed: the norm of the outliers is $\|\boldsymbol{G}\|_2$, and the directions are orthogonal to $\boldsymbol{G}$. Through this construction, outliers cannot be easily removed by simple pruning, and the directions of outliers can cause large effects on the estimation of $\boldsymbol{G}$. 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 [22]. 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 [19] 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.

H.2 Robust sparse regression with identity covariance

We use Algorithm 1 for robust sparse regression. Similarly as in Appendix H.1, we use Algorithm 3 as our Robust Sparse Gradient Estimator, and leverage the Sparse PCA solver from [19]. In the simulation, we fix $d = 500$, and $k = 5$, hence the corresponding sample complexity is $n \propto 1/\epsilon$. However,

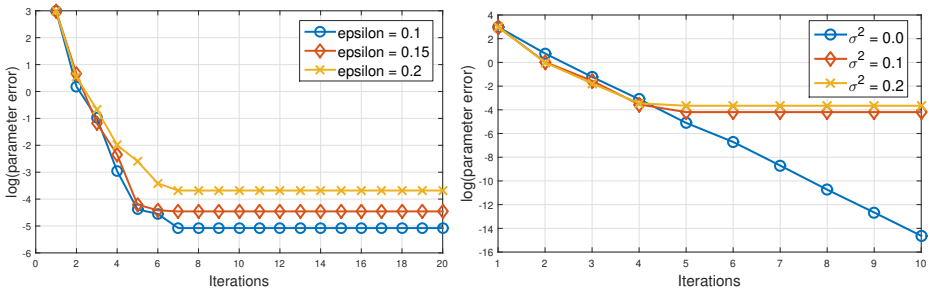

Figure 2: Empirical illustration of the linear convergence of $\log(\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2^2)$ vs. iteration counts in the Algorithm 1. In all cases, we fix $k = 5$, $d = 500$, and choose the sample complexity $n \propto 1/\epsilon$. The left plot considers different ϵ with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

we do not use the sample splitting technique in the simulations.

The entries of the true parameter $\boldsymbol{\beta}^*$ are set to be either $+1$ or -1, hence $\|\boldsymbol{\beta}^*\|_2^2 = k$ is fixed. The authentic $\boldsymbol{x}_i$s are generated from $\mathcal{N}(0, \boldsymbol{I}_d)$, and the authentic $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \xi_i$ as in Model 1.1. We set the covariates of the outliers as A, where A is a random ± 1 matrix of dimension $\epsilon n/(1-\epsilon) \times d$, and set the responses of outliers to $-A\boldsymbol{\beta}^*$.

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.

H.3 Robust sparse regression with unknown covariance matrix

Following Appendix H.2, we 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 Appendix H.2, 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.

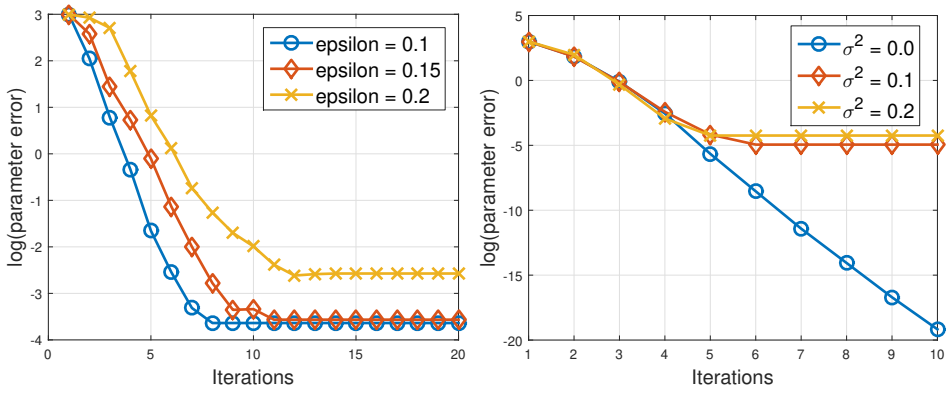

Figure 3: Empirical illustration of the linear convergence of $\log(\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2^2)$ vs. iteration counts in the Algorithm 1 with unknown covariance matrix which is a Toeplitz matrix with a decay $\boldsymbol{\Sigma}_{ij} = \exp^{-(i-j)^2}$. The other settings are the same as Figure 2. Even though the covariance matrix is unknown, we observe similar performance in linear convergence as Figure 2.