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 Shen and Li (2017), 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 Shen and Li (2017))**.** *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

$$\begin{aligned}
\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\|_2 &\le \sqrt{\zeta}\left\|\boldsymbol{\beta}^t - \eta\widehat{\boldsymbol{G}} - \boldsymbol{\beta}^*\right\|_2 \\
&= \sqrt{\zeta}\left\|\boldsymbol{\beta}^t - \eta\boldsymbol{G} - \boldsymbol{\beta}^* + \eta(\boldsymbol{G} - \widehat{\boldsymbol{G}})\right\|_2 \\
&\le \sqrt{\zeta}\|\boldsymbol{\beta}^t - \eta\boldsymbol{G} - \boldsymbol{\beta}^*\|_2 + \sqrt{\zeta}\eta\left\|\boldsymbol{G} - \widehat{\boldsymbol{G}}\right\|_2
\end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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

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 Balakrishnan et al. (2017a)'s proof

A key part of the proof of the main theorem in Balakrishnan et al. (2017a) is to obtain an upper bound on the k-sparse operator norm. Specifically, their Lemmas A.2 and A.3 aim to show:

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

where $\widehat{\boldsymbol{G}}(w) = \mathsf{P}_{2\widetilde{k}}\left(\sum_{i=1}^{|\mathcal{S}|} w_i \boldsymbol{g}_i\right)$, $\widetilde{\Delta}(w) = \sum_{i=1}^{|\mathcal{S}|} w_i \boldsymbol{g}_i - \boldsymbol{G}$[3], 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 Balakrishnan et al. (2017a) 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

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

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), Balakrishnan et al. (2017a) solves the following convex program:

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

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 $\mathcal{N}\left(\mu, \boldsymbol{I}_d\right)$, where function $F\left(\cdot\right)$ 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:

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

It is clear that $\|\widehat{\boldsymbol{\Sigma}} - F\|_{\widetilde{\mathrm{k}},\mathrm{op}} = 1$. Solving the convex relaxation $\max_{\boldsymbol{H}} \operatorname{Tr}\left(\boldsymbol{E} \cdot \boldsymbol{H}\right)$ or $\max_{\boldsymbol{H}} \operatorname{Tr}(\widehat{\boldsymbol{\Sigma}} \cdot \boldsymbol{H})$ gives answer $\boldsymbol{H}^* = \begin{bmatrix} 1 \quad 0; 0 \quad 0 \end{bmatrix}$ and the corresponding $\lambda^* = 0.5625$, which is clearly not an upper bound of $\|\widehat{\boldsymbol{\Sigma}} - F\|_{\widetilde{\mathrm{k}},\mathrm{op}}$. Hence $\lambda^* \ge \|\widehat{\boldsymbol{\Sigma}} - F\|_{\widetilde{\mathrm{k}},\mathrm{op}}$ 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 $F\left(\boldsymbol{G}\right)$. 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 $F\left(\boldsymbol{G}\right)$. L_{cov} and L_{F} are two important properties of $F\left(\boldsymbol{G}\right)$, related to its smoothness. We first

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

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 $F\left(\boldsymbol{G}\right)$.

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*

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

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

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

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

Next, we apply the Stein-type Lemma Stein (1981) for $\boldsymbol{x} \sim \mathcal{N}\left(0, \boldsymbol{I}_d\right)$, and a function $f\left(x\right)$ whose second derivative exists:

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

By eq. (10), we have

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

□

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

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

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

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

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 $F\left(\boldsymbol{G}\right)$,

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

For the L_{F} term, we have

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

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:

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

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:

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

The balanced weighted distribution is defined to satisfy:

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

Notice that $\sum_{i \in \mathcal{S}_{\mathrm{bad}}} \widetilde{w}_i = O\left(\epsilon\right)$, and $\sum_{i \in \mathcal{S}_{\mathrm{bad}}} w_i = O\left(\frac{\epsilon}{1-2\epsilon}\right)$ 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 Balakrishnan et al. (2017a))**.** *Suppose G is k-sparse. Then we have the following result:*

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

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 Balakrishnan et al. (2017a).

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 $\delta = \widetilde{O}\left(\epsilon\right)$, such that with probability at least $1-\nu$, for any index subset $\mathcal{J} \subset [d]$, $|\mathcal{J}| \leq \widetilde{k}$ and for any $\mathcal{G}' \subset \mathcal{G}$, $|\mathcal{G}'| \geq (1-2\epsilon)|\mathcal{G}|$, the following inequalities hold:*

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

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

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 Balakrishnan et al. (2017a), 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*

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

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

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,

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

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

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:

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

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

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

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 $N(k, d, \epsilon, \nu)$ 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*

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

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 $\rho_{\text{sep}} = \Theta(\epsilon\big(\|\boldsymbol{G}^t\|_2^2 + \sigma^2\big))$ 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:

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

To prove the Main Theorem (Theorem 3.1) in Balakrishnan et al. (2017a), 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

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

with probability at least $1-\nu$. This exactly gives us a $\widetilde{O}\left(\epsilon^2\sigma^2\right)$-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 $\delta=\Omega\left(\epsilon\right)$ with $\delta=\Omega\left(\sqrt{\epsilon}\right)$, since the results for $\delta=\Omega\left(\epsilon\right)$ implies the results for $\delta=\Omega\left(\sqrt{\epsilon}\right)$.

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 $\delta=\widetilde{O}\left(\sqrt{\epsilon}\right)$ 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:*

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

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

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

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

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

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 $n=\Omega\left(\frac{k^2\log(d/\nu)}{\epsilon}\right)$ 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 $\{\boldsymbol{g}_i^t\}_{i=1}^n$. Algorithm 3 computes λ^* that satisfies*

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

If $\lambda^*\ge\rho_{\text{sep}}=C_\gamma\left(\|\boldsymbol{G}^t\|_2^2+\sigma^2\right)$, then with probability at least $1-\nu$, we have

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

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

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

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

By Theorem A.1 in Balakrishnan et al. (2017a), we have

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

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 $\mathcal{S}_{\text{good}}$, Balakrishnan et al. (2017a) 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

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

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

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

$$\begin{aligned}&\mathbb{E}_{i\in_u\mathcal{S}_{\text{good}}}\tau_i\\ &=\operatorname{Tr}\left(H^*\cdot\left(\mathbb{E}_{i\in_u\mathcal{S}_{\text{good}}}\left(\boldsymbol{g}_i-\widehat{\boldsymbol{G}}\right)^{\otimes 2}\right)\right)\end{aligned}$$

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

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. (2), H^* belongs to the Fantope $\mathcal{F}^1$ Vu et al. (2013), and thus for any matrix A, we have $\operatorname{Tr}\left(A \cdot H^*\right) \leq \|A\|_{\text{op}}$.

Thus, we have

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

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

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

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

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

Now, under the condition $\lambda^* \geq \rho_{\text{sep}} = \Theta\left(\|\boldsymbol{G}\|_2^2 + \sigma^2\right)$, 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

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

$$\geq \frac{\left\|\widehat{\Delta}\right\|_2^2}{\epsilon}.$$

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 $(\gamma-1)/\gamma$*:*

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

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

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

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

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

Finally,

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

□

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

We use the martingale technique in Xu et al. (2013), by defining T: $T=\min\{l:\mathcal{E}^l \text{ is true}\}$. Based on T, we define a random variable:

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

Lemma F.2 (Lemma 1 in Xu et al. (2013))**.** $\{Y^l, \mathcal{F}^l\}$ *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* $n=\Omega\big(\frac{k^2\log(d/\nu)}{\epsilon}\big)$ *ϵ-corrupted samples from Model 1.1 with* $\boldsymbol{\Sigma}=\boldsymbol{I}_d$*. Let* $\mathcal{S}_{\mathrm{in}}$ *be an ϵ-corrupted set of gradient samples* $\{\boldsymbol{g}_i^t\}_{i=1}^n$*. By setting* $\widetilde{k}=k'+k$*, if we run Algorithm 3 iteratively with initial set* $\mathcal{S}_{\mathrm{in}}$*, and subsequently on* $\mathcal{S}_{\mathrm{out}}$*, and use* $\rho_{\mathrm{sep}}=C_\gamma\left(\|\boldsymbol{G}^t\|_2^2+\sigma^2\right)$*, then this repeated use of Algorithm 3 will stop after at most* $\frac{1.1\gamma}{\gamma-1}\epsilon n$ *iterations, and output* $\widehat{\boldsymbol{G}}^t$*, such that*

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

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.*

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 Xu et al. (2013). 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 supermartingale. 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 $F(\boldsymbol{G})$ 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 $\mathcal{J}\subset[d]$, $|\mathcal{J}|\le\widetilde{k}$ and for any $\mathcal{G}'\subset\mathcal{G}$, $|\mathcal{G}'|\ge(1-2\epsilon)|\mathcal{G}|$, we have*

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

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

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

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

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 Diakonikolas et al. (2016), by replacing Theorem 5.50 with Theorem 5.44 in Vershynin (2010).

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 $N\left(k, d, \epsilon, \nu\right)$ 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. □

H Additional experiments

In the this section, we consider the actual running time to demonstrate the practical usefulness of the algorithm in high dimensions, by plotting against wall time (Section 6 has only plots against iteration number).

We show the scalability of our robust sparse regression algorithm under different setups. The setup for robust sparse regression is similar to Section 6 – 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

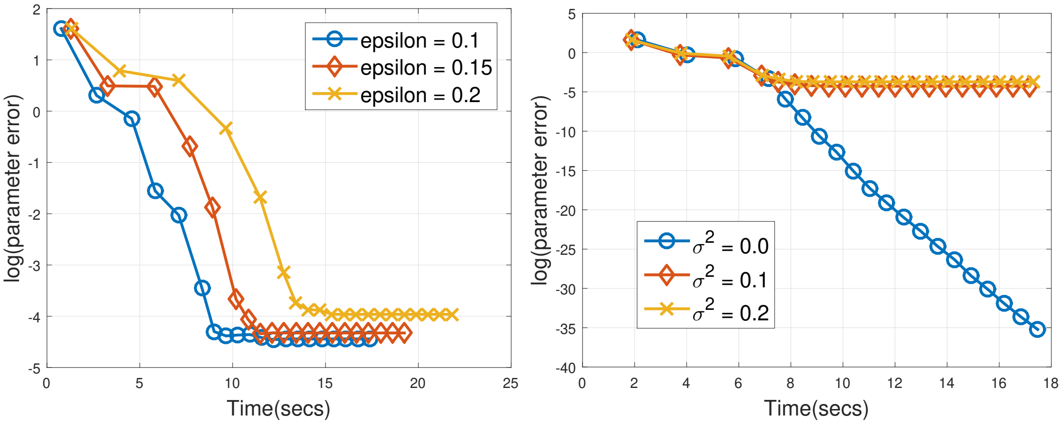

Figure 4: In this figure, we show the error accuracy vs. wall clock time for each iterate of Algorithm 1 via filtering (Algorithm 3). In both plots, we use the same setup in the paper and fix $d = 2000$, and clean data sample size $n = 1000$. In the left plot, the fraction of outliers epsilon is fixed as 0.1. In the right plot, the noise variance is fixed as 0.1, and we vary the epsilon. Since larger epsilon leads to more outliers, the computational time for $\epsilon = 0.2$ is slightly larger.

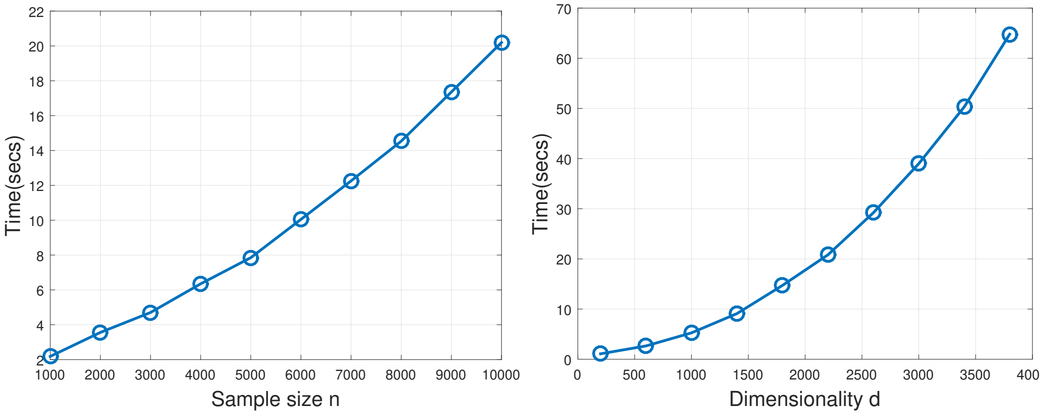

Figure 5: In this figure, we show the wall clock time vs. the sample size or the dimensionality. In both plots, we use $\epsilon = 0.1$ and fix the iteration number to be 20, which is sufficient to recover the parameter. In the left plot, we fix $d = 500$ and vary the sample size n. Since the number of outliers is linear in n when epsilon is fixed, the computational time has a linear dependence on n in theory and practice. In the right plot, we fix $n = 1000$ and vary the dimensionality d. Though the computational complexity depends on different Sparse PCA solvers (e.g., d'Aspremont et al. (2007)), we show that our algorithm can easily scale for high dimensions.

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 $-A\boldsymbol{\beta}^*$.

The error accuracy vs. wall clock time and scalability in high dimensions are summarized in Figure 4 and Figure 5. In particular, the convergence in Figure 4 with respect to clock time is similar to the convergence with respect to iterations in Figure 2. And Figure 5 shows the scalability of our robust sparse regression algorithm in very high dimensions.