Robust Sparse Regression under Adversarial Corruption Supplementary Material

Yudong Chen Constantine Caramanis Shie Mannor

In this supplementary material, we prove the theoretical results in the main paper.

1 Proof of Theorem 1

Recall that $y = [y^{\mathcal{A}}; y^{\mathcal{O}}]$ and $X = [X^{\mathcal{A}}; X^{\mathcal{O}}]$ with $y^{\mathcal{A}} = X^{\mathcal{A}}\beta^* + e$, and Λ^* is the true support. The adversary fixes some set $\hat{\Lambda}$ disjoint from the true support Λ^* with $|\hat{\Lambda}| = |\Lambda^*|$. It then chooses $\hat{\beta}$ and $y^{\mathcal{O}}$ such that $\hat{\beta}_{\hat{\Lambda}} = \beta^*_{\Lambda^*}$ $\hat{\beta}_{\hat{\Lambda}^c} = 0$, and $y^{\mathcal{O}} = X^{\mathcal{O}}\hat{\beta}$ with $X^{\mathcal{O}}$ to be determined later. By assumption we have $h(\hat{\beta}) = h(\beta^*) \le R$, so $\hat{\beta}$ is feasible. Its objective value is $f(y - X\hat{\beta}) = f([y^{\mathcal{A}} - X^{\mathcal{A}}_{\hat{\Lambda}}\beta^*_{\Lambda^*}; 0]) \le C$ for some finite constant C. The adversary further chooses $X^{\mathcal{O}}$ such that $X^{\mathcal{O}}_{\Lambda^*} = 0$ and $X^{\mathcal{O}}_{\hat{\Lambda}}$ is large. Any $\tilde{\beta}$ supported on Λ^* has objective value

$$f(y - X\tilde{\beta}) = f([y^{\mathcal{A}} - X^{\mathcal{A}}\tilde{\beta}; X^{\mathcal{O}}(\hat{\beta} - \tilde{\beta})]) = f([y^{\mathcal{A}} - X^{\mathcal{A}}\tilde{\beta}; X^{\mathcal{O}}_{\Lambda}\beta^*_{\Lambda^*}]) \ge f([0; X^{\mathcal{O}}_{\Lambda}\beta^*_{\Lambda^*}]),$$

which can be made bigger than C under the SCO Condition. Therefore, any solution $\tilde{\beta}$ with the correct support Λ^* has a higher objective value than $\hat{\beta}$, and thus is not the optimal solution.

2 Proof of Theorem 2

For simplicity we assume $\Lambda^* = \{1, \dots, k\}$, $\mathcal{A} = \{1, \dots, n\}$, and $\mathcal{O} = \{n+1, \dots, n+n_1\}$. We will show that the adversary choose $y^{\mathcal{O}}$ and $X^{\mathcal{O}}$ in such a way that any "correct" solution of the form $(\theta, \mathcal{S}, \Lambda^*)$ (i.e., with the correct support Λ^*) is not optimal because an alternative solution $(\hat{\theta}, \hat{\mathcal{S}}, \hat{\Lambda})$ with $\hat{\theta} = [1, \dots, 1]^\top$, $\hat{\mathcal{S}} = \{n_1 + 1, \dots, n + n_1\}$, $\hat{\Lambda} = \{2, \dots, k, k+1\}$ has smaller objective value.

Now for the details. The adversary chooses $\left(y^{\mathcal{O}}\right)_i = \sqrt{k}$ for all i, $X^{\mathcal{O}}_{\Lambda^*} = 0$, and $X^{\mathcal{O}}_{k+1} = y^{\mathcal{O}}$, hence $y^{\mathcal{O}} - X^{\mathcal{O}}_{\hat{\Lambda}}\hat{\theta} = 0$. To compute the objective values of the "correct" solution and the alternative solution, we need a simple technical lemma, which follows from standard results for the norms of random Gaussian matrix. The proof is given in the appendix.

Lemma 1. *If $n \gtrsim k^3 \log p$, we have*

$$\begin{aligned} \left\|e + X^{\mathcal{A}}_{\Lambda^*}\delta\right\|_2^2 &\ge \left(1 - \frac{1}{k}\right)(\sigma_e^2 + \|\delta\|_2^2), \forall \delta \in \mathbb{R}^k \\ \left\|e^{\hat{\mathcal{S}}/\mathcal{O}} + X_1^{\hat{\mathcal{S}}/\mathcal{O}} - X_{k+1}^{\hat{\mathcal{S}}/\mathcal{O}}\right\|_2^2 &\le \left(1 + \frac{1}{k}\right)\left(1 - \frac{n_1}{n}\right)(\sigma_e^2 + 2) \end{aligned}$$

with high probability.

Using the above lemma, we can upper-bound the objective value of the alternative solution:

$$\begin{aligned}
\left\|y^{\hat{S}} - X^{\hat{S}}_{\hat{\Lambda}}\hat{\theta}\right\|_2^2 &= \left\|y^{\mathcal{O}} - X^{\mathcal{O}}_{\hat{\Lambda}}\hat{\theta}\right\|_2^2 + \left\|y^{\hat{S}/\mathcal{O}} - X^{\hat{S}/\mathcal{O}}_{\hat{\Lambda}}\hat{\theta}\right\|_2^2 \\
&= 0 + \left\|y^{\hat{S}/\mathcal{O}} - X^{\hat{S}/\mathcal{O}}_{\Lambda^*}\beta^*_{\Lambda^*} + X^{\hat{S}/\mathcal{O}}_1 - X^{\hat{S}/\mathcal{O}}_{k+1}\right\|_2^{2\,2} \\
&= \left\|e_{\hat{S}/\mathcal{O}} + X^{\hat{S}/\mathcal{O}}_1 - X^{\hat{S}/\mathcal{O}}_{k+1}\right\|_2^2 \\
&\leq \left(1+\frac{1}{k}\right)\left(1-\frac{n_1}{n}\right)(\sigma_e^2+2). \qquad (1)
\end{aligned}$$

To lower-bound the objective value of solutions of the form $(\theta, \mathcal{S}, \Lambda^*)$, we distinguish two cases. If $\mathcal{S}\cap\mathcal{O} \neq \phi$, then the objective value is

$$\begin{aligned}
\left\|y^{\mathcal{S}} - X^{\mathcal{S}}_{\Lambda^*}\theta\right\|_2^2 &\geq \left\|y^{\mathcal{S}\cap\mathcal{O}} - X^{\mathcal{S}\cap\mathcal{O}}_{\Lambda^*}\theta\right\|_2^2 \\
&= \left\|y^{\mathcal{S}\cap\mathcal{O}}\right\|_2^2 \\
&\geq k \qquad (2)
\end{aligned}$$

If $\mathcal{S}\cap\mathcal{O} = \phi$, we have $\mathcal{S} = \mathcal{A}$ and thus

$$\begin{aligned}
\left\|y^{\mathcal{S}} - X^{\mathcal{S}}_{\Lambda^*}\theta\right\|_2^2 &= \left\|y^{\mathcal{A}} - X^{\mathcal{A}}_{\Lambda^*}\theta\right\|_2^2 \\
&= \left\|X^{\mathcal{A}}_{\Lambda^*}\beta^*_{\Lambda^*} + e - X^{\mathcal{A}}_{\Lambda^*}\theta\right\|_2^2 \\
&= \left\|e + X^{\mathcal{A}}_{\Lambda^*}(\beta^* - \theta)\right\|_2^2 \\
&\geq \left(1-\frac{1}{k}\right)\sigma_e^2, \qquad (3)
\end{aligned}$$

where we use the lemma in the inequality. When $n_1 > \frac{3n}{k+1}$ and $\sigma_e^2 = k$, we have $\min\left\{k, \left(1-\frac{1}{k}\right)\sigma_e^2\right\} > \left(1+\frac{1}{k}\right)\left(1-\frac{n_1}{n}\right)(\sigma_e^2+2)$. Combining (1), (2) and (3) concludes the proof.

2.1 Proof of the Lemma 1

Let $\theta' = \left[\sigma_e|\delta^\top\right]^\top$. We can write $\left\|e + X^{\mathcal{A}}_{\Lambda^*}\delta\right\|_2^2 = \|Z_1\theta'\|_2^2$ with $Z_1 \triangleq \left[\frac{1}{\sigma_e}e|X^{\mathcal{A}}_{\Lambda^*}\right]$. Note that Z_1 is an $n\times(k+1)$ matrix with i.i.d. $\mathcal{N}(0, \frac{1}{n})$ entries, whose smallest singular value can be bounded using standard results. For example, using Lemma 5.1 in [1] with $\Phi(\omega) = Z_1$, $N = k+1$, $T = \{1, \ldots, N\}$, $\delta = \frac{1}{3k}$ and $c_0(\delta/2) = 1/288k^2$, we have

$$\|Z_1\theta'\|_2^2 \geq \left(1-\frac{1}{3k}\right)^2\|\theta'\|_2^2 \geq \left(1-\frac{1}{k}\right)(\sigma_e^2 + \|\delta\|_2^2), \forall\delta$$

with probability at least

$$1 - 2e^{\frac{1}{288k^2}n - (k+1)\ln(36k)} \geq 1 - 2p^{-3}$$

provided $n \geq 576(k+1)^3\ln(36p)$. This proves the first inequality. The second lemma can be proved similarly using Lemma 5.1 in [1].

3 Proof of Theorem 3

We prove Theorem 3 in this section. We need two technical lemmas. The first lemma bounds the maximum of independent sub-Gaussian random variables. The proof follows from the definition of sub-Gaussianity and Chernoff bound, and is given in the next subsection.

Lemma 2. *Suppose $Z_1, \ldots, Z_m$ are m independent sub-Gaussian random variables with parameter σ. Then we have $\max_{i=1,\ldots,m} |Z_i| \le 4\sigma\sqrt{\log m + \log p}$. with high probability.*

The second lemma is a standard concentration result for the sum of squares of independent sub-Gaussian random variables. It follows directly from Eq. (74) in [2].

Lemma 3. *Let $Y_1, \ldots, Y_n$ be n i.i.d. zero-mean sub-Gaussian random variables with parameter $\frac{1}{\sqrt{n}}$ and variance at most $\frac{1}{n}$. Then we have*

$$|\sum_{i=1}^{n} Y_i^2 - 1| \le c_1\sqrt{\frac{\log p}{n}}$$

with high probability for some absolute constant c_1. Moreover, if $Z_1, \ldots, Z_n$ are also i.i.d. zero-mean sub-Gaussian random variables with parameter $\frac{1}{\sqrt{n}}$ and variance at most $\frac{1}{n}$, and independent of $Y_1, \ldots, Y_n$, then

$$|\sum_{i=1}^{n} Y_i Z_i| \le c_2\sqrt{\frac{\log p}{n}}$$

with high probability for some absolute constant c_2.

Remark. *When the above inequality holds, we write $\sum_{i=1}^n Y_i^2 \approx 1 \pm \sqrt{\frac{\log p}{n}}$ and $\sum_{i=1}^n Y_i Z_i \approx \pm\sqrt{\frac{\log p}{n}}$ w.h.p.*

We now turn to the proof of Theorem 3. We first consider the distributed corruption model. For simplicity we assume that for each columns of X and y there are at most $\frac{n_1}{2}$ entries corrupted (instead of n_1 entries). This will affect the statement of the theorem by a constant of 2.

Consider the trimmed inner product $h(j)$ between the jth column of X and y. Let $\mathcal{A}_j$ is the set of index i such that X_{ij} and y_i are both not corrupted. By assumption $|\mathcal{A}_j| \ge n$. By putting $|\mathcal{A}_j| - n$ clean indices in $\mathcal{A}_j^c$, we may assume $|\mathcal{A}_j| = n$ without loss of generality. By prescription of Algorithm 2, we can write $h(j)$ as

$$h(j) = \sum_{i\in\mathcal{A}_j} X_{ij}y_i - \sum_{\substack{i\in\text{trimmed}\\ \text{inliers}}} X_{ij}y_i + \sum_{\substack{i\in\text{remaining}\\ \text{outliers}}} X_{ij}y_i.$$

We estimate each term in the above sum.

1. Observe that

$$\sum_{i\in\mathcal{A}_j} X_{ij}y_i = \sum_{i\in\mathcal{A}_j} X_{ij}\left(\sum_{k=1}^{p} X_{ik}\beta_k^* + e\right) = \sum_{i\in\mathcal{A}_j} X_{ij}^2\beta_j^* + \sum_{i\in\mathcal{A}_j} X_{ij}\left(\sum_{k\neq j} X_{ik}\beta_k^* + e\right).$$

   (a) Because the points in $\mathcal{A}_j$ obeys the Sub-Gaussian model, Lemma 3 gives w.h.p. $\sum_{i\in\mathcal{A}_j} X_{ij}^2\beta_j^* \approx \beta_j^*\left(1 \pm \sqrt{\frac{1}{n}\log p}\right)$.

   (b) On the other hand, because X_{ik} and X_{ij} are independent when $k \neq j$, and $Z_i \triangleq \sum_{k\neq j} X_{ik}\beta_k^* + e$ are i.i.d. sub-Gaussian with parameter and standard deviation at most $\sqrt{\left(\|\beta^*\|_2^2 + \sigma_e^2\right)/n}$, we apply Lemma 3 to obtain $\sum_{i\in\mathcal{A}_j} X_{ij}Z_i \approx \pm\frac{1}{\sqrt{n}}\sqrt{\left(\|\beta^*\|_2^2 + \sigma_e^2\right)\log p}$ w.h.p.

2. Again due to independence and sub-Gaussianity of points in $\mathcal{A}_j$, Lemma 2 gives $\max_{i\in\mathcal{A}_j}|X_{ij}| \lesssim \sqrt{(\log p)/n}$ w.h.p. and $\max_{i\in\mathcal{A}_j}|y_i| \lesssim \sqrt{(\log p/n)\left(\|\beta^*\|_2^2+\sigma_e^2\right)}$ w.h.p. It follows that w.h.p.

$$\left|\sum_{\substack{i\in\text{trimmed}\\ \text{inliers}}} X_{ij}y_i\right| \le n_1\left(\max_{i\in\mathcal{A}}|X_{ij}|\right)\left(\max_{i\in\mathcal{A}}|y_i|\right) \lesssim n_1\cdot\sqrt{\frac{\log p}{n}}\cdot\sqrt{\frac{\log p}{n}\left(\|\beta^*\|_2^2+\sigma_e^2\right)}.$$

3. By prescription of the trimming procedure, either all outliers are trimmed, or the remaining outliers are no larger than the trimmed inliers. It follows from the last equation that w.h.p.

$$\left|\sum_{\substack{i\in\text{remaining}\\ \text{outliers}}} X_{ij}y_i\right| \le \sum_{\substack{i\in\text{remaining}\\ \text{outliers}}} |X_{ij}y_i| \le \sum_{\substack{i\in\text{trimmed}\\ \text{inliers}}} |X_{ij}y_i| \lesssim n_1\frac{\log p}{n}\cdot\sqrt{\|\beta^*\|_2^2+\sigma_e^2}.$$

Combining pieces, we have for all $j=1,\ldots,p$,

$$\left|h(j)-\beta_j^*\right| \lesssim \left|\beta_j^*\right|\sqrt{\frac{2}{n}\log p}+\frac{1}{\sqrt{n}}\sqrt{\left(\|\beta^*\|_2^2+\sigma_e^2\right)\log p}+n_1\cdot\frac{\log p}{n}\sqrt{\left(\|\beta^*\|_2^2+\sigma_e^2\right)}. \tag{4}$$

If RoTR correctly picks an index j in the true support Λ^*, then the error in estimating $\hat{\beta}_j$ is bounded by the expression above. If RoTR picks some incorrect index j not in Λ^*, then the difference between the corresponding $\hat{\beta}_j$ and the true $\beta^*_{j'}$ that should have been picked is still bounded by the expression above (up to constant factors). Therefore, we have

$$\left\|\hat{\beta}-\beta^*\right\|_2^2 \lesssim \sum_{j\in\Lambda^*}\left[\left|\beta_j^*\right|\sqrt{\frac{2}{n}\log p}+\frac{1}{\sqrt{n}}\sqrt{\left(\|\beta^*\|_2^2+\sigma_e^2\right)\log p}\ +n_1\frac{\log p}{n}\sqrt{\left(\|\beta^*\|_2^2+\sigma_e^2\right)}\right]^2.$$

The first part of the theorem then follows after straightforward algebra manipulation. On the other hand, RoTR picks the correct support as long as $|h(j)|>|h(j')|$ for all $j\in\Lambda^*, j'\in(\Lambda^*)^c$. In view of Eq.(4), we require

$$\begin{aligned} n &\gtrsim \max_j\left(\frac{\|\beta^*\|_2^2}{\beta_j^2}\right)\cdot\log p\cdot\left(1+\sigma_e^2/\left\|\beta^*\right\|_2^2\right)\\ \frac{n_1}{n} &\lesssim \frac{1}{\sqrt{\max_j\left(\frac{\|\beta^*\|_2^2}{\beta_j^2}\right)\cdot\left(1+\sigma_e^2/\left\|\beta^*\right\|_2^2\right)\log p}}\end{aligned}$$

One verifies that the above inequalities are satisfied under the conditions in the second part of the theorem.

Now consider the row corruption model. A careful examination of the proof above shows that, when there are n_1 corrupted rows, the set $\mathcal{A}_j$ still has cardinality at least n, and the proof thus holds under the row corruption model.

3.1 Proof of Lemma 2

Let $\hat{Z} = \max_i Z_i$. By definition of sub-Gaussianity, we have

$$\begin{aligned}\mathbb{E}\left[e^{t\hat{Z}/\sigma}\right] &= \mathbb{E}\left[\max_i e^{tZ_i/\sigma}\right] \\ &\leq \sum_i \mathbb{E}\left[e^{tZ_i/\sigma}\right] \\ &\leq me^{t^2/2} \\ &= e^{t^2/2+\log m}\end{aligned}$$

It follows from Markov Inequality that

$$\begin{aligned}P(\hat{Z} \geq \sigma t) &= P(e^{t\hat{Z}/\sigma} \geq e^{t^2}) \\ &\leq e^{-t^2}\mathbb{E}\left[e^{t\hat{Z}/\sigma}\right] \\ &\leq e^{-t^2+t^2/2+\log m} \\ &= e^{-\frac{1}{2}t^2+\log m}.\end{aligned}$$

By symmetry we have

$$P(\min_i Z_i \leq -\sigma t) \leq e^{-\frac{1}{2}t^2+\log m},$$

so a union bound gives

$$\begin{aligned}P(\max_i |Z_i| \geq \sigma t) &\leq P(\max_i Z_i \geq \sigma t) + P(\min_i Z_i \leq -\sigma t) \\ &\leq 2e^{-\frac{1}{2}t^2+\log m}.\end{aligned}$$

Taking $t = 4\sqrt{\log m + \log p}$ yields the result.

4 Proof of Theorem 4

We now prove Theorem 4. We first prove the theorem for Robust Lasso, and then for Robust Dantzig selector. We will use Lemma 2 and Lemma 3 given in the last section.

4.1 Proof for Robust Lasso

For simplicity, we only prove the theorem for the row corruption model. It is straightforward to adapt the proof for the distributed corruption model (cf. the proof for RoTR in the last section).

Let $\Delta := \hat{\beta} - \beta^*$, $F := \hat{\Gamma} - X^{\mathcal{A}\top}X^{\mathcal{A}}$, $f := \hat{\gamma} - X^{\mathcal{A}\top}X^{\mathcal{A}}\beta^*$, and $S = \text{support}(\beta^*)$. For any vector $b \in \mathbb{R}^p$, b_S is the vector with $(b_S)_i = b_i$ for $i \in S$ and $(b_S)_i = 0$ for $i \notin S$.

Because $\hat{\beta}$ satisfies the constraint in the optimization problem in Robust Lasso, we have

$$\begin{aligned}\|\beta^*\|_1 &\geq \|\beta^* + \Delta\|_1 \\ &= \|\beta^* + \Delta_S\|_1 + \|\Delta_{S^c}\|_1 \\ &\geq \|\beta^*\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 .\end{aligned}$$

It follows that $\|\Delta_{S^c}\| \leq \|\Delta_S\|_1$. Because $|S| = k$, we obtain the following inequality

$$\begin{aligned} \|\Delta\|_1 &= \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \\ &\leq 2\|\Delta_S\|_1 \\ &\leq 2\sqrt{k}\|\Delta_S\|_2 \\ &\leq 2\sqrt{k}\|\Delta\|_2 . \end{aligned} \tag{5}$$

Under the assumption for n in the theorem, standard results (e.g. Lemma 1 in [2]) guarantees that the authentic $X^{\mathcal{A}}$ satisfies Restricted Strong Convexity (RSC) under the assumption of the theorem:

$$u^\top (X^{\mathcal{A}\top} X^{\mathcal{A}}) u \geq \frac{1}{4}\lambda_{\min}(\Sigma_x)\|u\|_2 , \quad \forall u : \|u\|_1 \leq 2\sqrt{k}\|u\|_2 . \tag{6}$$

Combining with (5), we obtain

$$\begin{aligned} \Delta^\top \hat{\Gamma}^\top \Delta &= \Delta^\top (X^{\mathcal{A}\top} X^{\mathcal{A}})\Delta + \Delta^\top F \Delta \\ &\geq \frac{1}{2}\lambda_{\min}(\Sigma_x)\|\Delta\|_2^2 - \|F\|_\infty \sum_{i,j} |\Delta_i|\,|\Delta_j| \\ &= \frac{1}{2}\lambda_{\min}(\Sigma_x)\|\Delta\|_2^2 - \|F\|_\infty \|\Delta\|_1^2 \\ &\geq \frac{1}{2}\lambda_{\min}(\Sigma_x)\|\Delta\|_2^2 - 4k\|F\|_\infty \|\Delta\|_2^2 . \end{aligned} \tag{7}$$

The magnitude of F_{ij} can be bounded similarly to the proof of RoTR. To see this, let $\mathcal{T}$ be the set of trimmed indices, and recall that $\mathcal{A}$ and $\mathcal{O}$ are the sets of inliers and outliers, respectively. We can write

$$\begin{aligned} F_{ij} &= \langle X_i, X_j \rangle_{n_1} - \langle X_i^{\mathcal{A}}, X_j^{\mathcal{A}} \rangle \\ &= -\sum_{k \in \mathcal{T}\cap\mathcal{A}} X_{ki}X_{kj} + \sum_{k \in \mathcal{T}^c \cap \mathcal{O}} X_{ki}X_{kj} \\ &\leq 2n_1 \left(\max_{k\in\mathcal{A}} |X_{ki}|\right)\left(\left|\max_{k\in\mathcal{A}} X_{kj}\right|\right). \end{aligned}$$

Because X_{ki}, $k \in \mathcal{A}$ are independent sub-Gaussian variable parameters $\frac{1}{n}\sigma_x^2$, Lemma 2 gives $\max_{k\in\mathcal{A}} |X_{ki}| \lesssim \sigma_x\sqrt{\frac{\log p}{n}}$ w.h.p. It follows from a union bound over (i,j) that

$$\|F\|_\infty \lesssim \frac{n_1 \log p}{n}\sigma_x^2. \tag{8}$$

Under our assumption, we have $\frac{n_1}{n} \lesssim \frac{\lambda_{\min}(\Sigma_x)}{\sigma_x^2 k \log p}$, so

$$\|F\|_\infty \leq \frac{\lambda_{\min}(\Sigma_x)}{16k}. \tag{9}$$

We thus obtain

$$\Delta^\top \hat{\Gamma}^\top \Delta \geq \frac{\lambda_{\min}(\Sigma_x)}{4}\|\Delta\|_2^2 . \tag{10}$$

By Holder's inequality and (5), we have

$$\begin{aligned} \left\langle \hat{\gamma} - \hat{\Gamma}\beta^*, \Delta \right\rangle &\leq \left\|\hat{\gamma} - \hat{\Gamma}\beta^*\right\|_\infty \|\Delta\|_1 \\ &\leq 4\sqrt{k}\left\|\hat{\gamma} - \hat{\Gamma}\beta^*\right\|_\infty \|\Delta\|_2 \end{aligned}$$

Now note that

$$\begin{aligned}\left\|\hat{\gamma}-\hat{\Gamma}\beta^*\right\|_\infty &\leq \left\|X^{\mathcal{A}\top}X^{\mathcal{A}}\beta^*-\hat{\Gamma}\beta^*\right\|_\infty+\left\|\hat{\gamma}-X^{\mathcal{A}\top}X^{\mathcal{A}}\beta^*\right\|_\infty\\ &= \|F\beta^*\|_\infty+\|f\|_\infty.\end{aligned}$$

Using (8) and the k-sparsity of β^*, we can bound the first term with $\frac{\sqrt{k}n_1\log p}{n}\sigma_x^2\|\beta\|_2$. For the second term, we decompose f_j as

$$\begin{aligned}f_j &= \langle X_j,y\rangle_R-\left\langle X_j^{\mathcal{A}},X^{\mathcal{A}}\beta^*\right\rangle\\ &= \sum_{i\in\mathcal{T}^c}X_{ij}y_i-\left\langle X_j^{\mathcal{A}},X^{\mathcal{A}}\beta^*\right\rangle\\ &= \left(\sum_{i\in\mathcal{A}}X_{ij}y_i-\left\langle X_j^{\mathcal{A}},X^{\mathcal{A}}\beta^*\right\rangle\right)-\sum_{i\in\mathcal{T}\cap\mathcal{A}}X_{ij}y_i+\sum_{i\in\mathcal{T}^c\cap\mathcal{O}}X_{ij}y_i\\ &= \left\langle X_j^{\mathcal{A}},e\right\rangle-\sum_{i\in\mathcal{T}\cap\mathcal{A}}X_{ij}y_i+\sum_{i\in\mathcal{T}^c\cap\mathcal{O}}X_{ij}y_i.\end{aligned}$$

We have $\left\langle X_j^{\mathcal{A}},e\right\rangle\approx\sqrt{\frac{\sigma_e^2\log p}{n}}$ w.h.p. by Lemma 3. Under the sub-Gaussian Design model, each y_i, $i\in\mathcal{A}$ is sub-Gaussian with parameter $\frac{\sigma_e^2+\sigma_x^2\|\beta^*\|_2^2}{n}$. Using Lemma 2 similarly as before, we obtain

$$\left|-\sum_{i\in\mathcal{T}\cap\mathcal{A}}X_{ij}y_i+\sum_{i\in\mathcal{T}^c\cap\mathcal{O}}X_{ij}y_i\right|\lesssim\frac{n_1\log p}{n}\sigma_x^2\sqrt{\sigma_e^2+\sigma_x^2\|\beta^*\|_2^2}.$$

It follows from a union bound that

$$\|f\|_\infty\lesssim\sqrt{\frac{\sigma_e^2\log p}{n}}+\frac{n_1\log p}{n}\sigma_x^2\sqrt{\sigma_e^2+\sigma_x^2\|\beta^*\|_2^2}. \tag{11}$$

Combining pieces, we obtain

$$\left\langle\hat{\gamma}-\hat{\Gamma}\beta^*,\Delta\right\rangle\lesssim\|\Delta\|_2\left(\frac{kn_1\log p}{n}\sigma_x^2\|\beta\|_2+\sqrt{\frac{k\sigma_e^2\log p}{n}}+\frac{n_1\log p\sqrt{k}}{n}\sigma_x\sqrt{\sigma_e^2+\sigma_x^2\|\beta^*\|_2^2}\right). \tag{12}$$

By optimality of $\hat{\beta}$, we have

$$\frac{1}{2}\hat{\beta}^\top\hat{\Gamma}\hat{\beta}-\hat{\gamma}^\top\hat{\beta}\leq\frac{1}{2}\beta^{*\top}\hat{\Gamma}\beta^*-\hat{\gamma}^\top\beta^*.$$

Rearranging terms, we get

$$\frac{1}{2}\Delta^\top\hat{\Gamma}^\top\Delta\leq\left\langle\hat{\gamma}-\hat{\Gamma}\beta^*,\Delta\right\rangle.$$

Combining the above inequality with (10), (12) and (5), we obtain

$$\frac{1}{2\sqrt{k}}\|\Delta\|_1\leq\|\Delta\|_2\lesssim\frac{1}{\lambda_{\min}(\Sigma_x)}\left(\frac{kn_1\log p}{n}\sigma_x^2\|\beta\|_2+\sqrt{\frac{k\sigma_e^2\log p}{n}}+\frac{n_1\log p\sqrt{k}}{n}\sigma_x\sqrt{\sigma_e^2+\sigma_x^2\|\beta^*\|_2^2}\right),$$

which concludes the proof of the theorem.

4.2 Proof for Robust Dantzig selector

Define F and f as before. Using (8) and (11), we have

$$\begin{aligned}\left\|\hat{\Gamma}\beta^* - \hat{\gamma}\right\|_\infty &= \left\|(X^{\mathcal{A}\top}X^{\mathcal{A}} + F)\beta^* - (f + X^{\mathcal{A}\top}X^{\mathcal{A}}\beta^*)\right\|_\infty \\ &\leq \|F\beta^*\|_\infty + \|f\|_\infty \\ &\leq \|F\|_\infty \|\beta^*\|_1 + \|f\|_\infty \\ &\lesssim \frac{n_1 \log p}{n}\sigma_x^2 \|\beta^*\|_1 + \sqrt{\frac{\sigma_e^2 \log p}{n}} + \frac{n_1 \log p}{n}\sigma_x^2\sqrt{\sigma_e^2 + \sigma_x^2 \|\beta^*\|_2^2}.\end{aligned}$$

Under the assumption of the theorem, this means that β^* is feasible to the optimization problem in Robust Dantzig selector. Applying Lemma 1 in [3], we obtain

$$\|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1 . \tag{13}$$

Now observe that

$$\begin{aligned}X^{\mathcal{A}\top}X^{\mathcal{A}}\Delta &= X^{\mathcal{A}\top}X^{\mathcal{A}}\hat{\beta} - X^{\mathcal{A}\top}X^{\mathcal{A}}\beta^* \\ &= \hat{\Gamma}\hat{\beta} - F\hat{\beta} - X^{\mathcal{A}\top}X^{\mathcal{A}}\beta^* \\ &= \hat{\Gamma}\hat{\beta} - \hat{\gamma} + \hat{\gamma} - F\hat{\beta} - X^{\mathcal{A}\top}X^{\mathcal{A}}\beta^* \\ &= \hat{\Gamma}\hat{\beta} - \hat{\gamma} + (f + X^{\mathcal{A}\top}X^{\mathcal{A}}\beta^*) - F\hat{\beta} - X^{\mathcal{A}\top}X^{\mathcal{A}}\beta^* \\ &= \hat{\Gamma}\hat{\beta} - \hat{\gamma} + f - F\hat{\beta}.\end{aligned}$$

It follows that

$$\begin{aligned}\left\|X^{\mathcal{A}\top}X^{\mathcal{A}}\Delta\right\|_\infty &\leq \left\|\hat{\Gamma}\hat{\beta} - \hat{\gamma}\right\|_\infty + \|f\|_\infty + \left\|F\hat{\beta}\right\|_\infty \\ &\leq \mu\left\|\hat{\beta}\right\|_1 + \tau + \|f\|_\infty + \left\|F\hat{\beta}\right\|_\infty \\ &\leq \mu\left\|\hat{\beta}\right\|_1 + \tau + \|f\|_\infty + \|F\|_\infty \left\|\hat{\beta}\right\|_1 \\ &\leq (\mu + \|F\|_\infty)\|\beta^*\|_1 + \tau + \|f\|_\infty , \end{aligned} \tag{14}$$

where we use the fact that $\hat{\beta}$ is feasible in the second inequality and the optimality of $\hat{\beta}$ in the last inequality. Using the definition of $\kappa_q(s)$, κ_{RE} and (13) in [4] (recall that Δ satisfies (13)), we have

$$\begin{aligned}\left\|X^{\mathcal{A}\top}X^{\mathcal{A}}\Delta\right\|_\infty &\geq \kappa_2(k)\|\Delta\|_2 \\ &\gtrsim \frac{1}{\sqrt{k}}\kappa_{RE}(2k)\|\Delta\|_2 ,\end{aligned}$$

where κ_{RE} is defined as

$$\kappa_{RE}(2k) \triangleq \min_{|J|=2k, \|u_{J^c}\|_1 \leq \|u_J\|_1} \frac{\left|u^\top X^{\mathcal{A}\top}X^{\mathcal{A}}u\right|}{\|u_J\|_2^2}.$$

Using Lemma 1 in [2], we know w.h.p. $\kappa_{RE}(2k) \geq \frac{1}{2}\lambda_{\min}(\Sigma_x)$ under the assumption of the theorem. Combining this with (14), we obtain

$$\|\Delta\|_2 \lesssim \frac{\sqrt{k}}{\lambda_{\min}(\Sigma_x)}\left((\mu + \|F\|_\infty)\|\beta^*\|_1 + \tau + \|f\|_\infty\right).$$

It then follows from (13), (8), (9) and our choice of μ and τ that

$$\|\Delta\|_2 \lesssim \frac{1}{\lambda_{\min}(\Sigma_x)} \left(\frac{kn_1 \log p}{n} \sigma_x^2 \|\beta^*\|_2 + \sigma_e \sqrt{\frac{k \log p}{n}} + \frac{n_1 \log p \sqrt{k}}{n} \sigma_x \sqrt{\sigma_e^2 + \sigma_x^2 \|\beta^*\|_2^2} \right).$$

This concludes the proof of the theorem.

5 Guarantees for the Projected Gradient Descent Method

Combining (6), (7) with (9), we know that $\hat{\Gamma}$ satisfies the lower-Restricted Eigenvalue condition in [2] with $\alpha_1 = \frac{1}{2}\lambda_{\min}(\Sigma_x)$ and $\tau(n,p) \leq \frac{1}{4}\alpha_1$ under the condition of Theorem 4. A similar argument shows that $\hat{\Gamma}$ satisfies the upper-Restricted Eigenvalue condition as well with $\alpha_2 = \frac{3}{2}\lambda_{\max}(\Sigma_x)$. We can then apply Theorem 2 in [2] to conclude the following: for the projected gradient descend method, if we choose the step size $\eta = 3\lambda_{\max}(\Sigma_x)$, then there exist absolute constant $c_1, c_2 > 0, \gamma < 1$ such that w.h.p., for all $t > 0$,

$$\begin{aligned} \left\|\beta^t - \hat{\beta}\right\|_2^2 &\leq \gamma^t \left\|\beta^0 - \hat{\beta}\right\|_2^2 + c_1 \frac{\log p}{n} \left\|\hat{\beta} - \beta^*\right\|_1^2 + \left\|\hat{\beta} - \beta^*\right\|_2^2, \\ \left\|\beta^t - \hat{\beta}\right\|_1 &\leq 2\sqrt{k} \left\|\beta^t - \hat{\beta}\right\|_2 + 2\sqrt{k} \left\|\beta^* - \hat{\beta}\right\|_2 + 2 \left\|\beta^* - \hat{\beta}\right\|_2. \end{aligned}$$

This means for t large enough, $\|\beta^t - \beta^*\|_q \lesssim \left\|\hat{\beta} - \beta^*\right\|_q$ for $q = 1, 2$, so the output of the projected gradient descend method also obeys the error bounds in Theorem 4.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[2] P.L. Loh and M.J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Arxiv preprint arXiv:1109.3714*, 2012.

[3] M. Rosenbaum and A.B. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

[4] M. Rosenbaum and A.B. Tsybakov. Improved matrix uncertainty selector. *arXiv preprint arXiv:1112.4413*, 2011.