A. Introduction

In this supplement, we first provide additional experimental results on the proposed estimator with MCP regularization, followed by the details of technical proof for the main results, including proofs of theorems and auxiliary lemmas.

B. Additional Experimental Results

Regarding matrix completion and matrix sensing, we present additional experimental results of the proposed estimator with MCP penalty. Due to the similar properties and parameter settings of these two nonconvex penalties, the MCP penalty and SCAD penalty, the numerical behaviour of the proposed estimator with MCP penalty resembles the one with SCAD penalty, as shown in Figure 2.

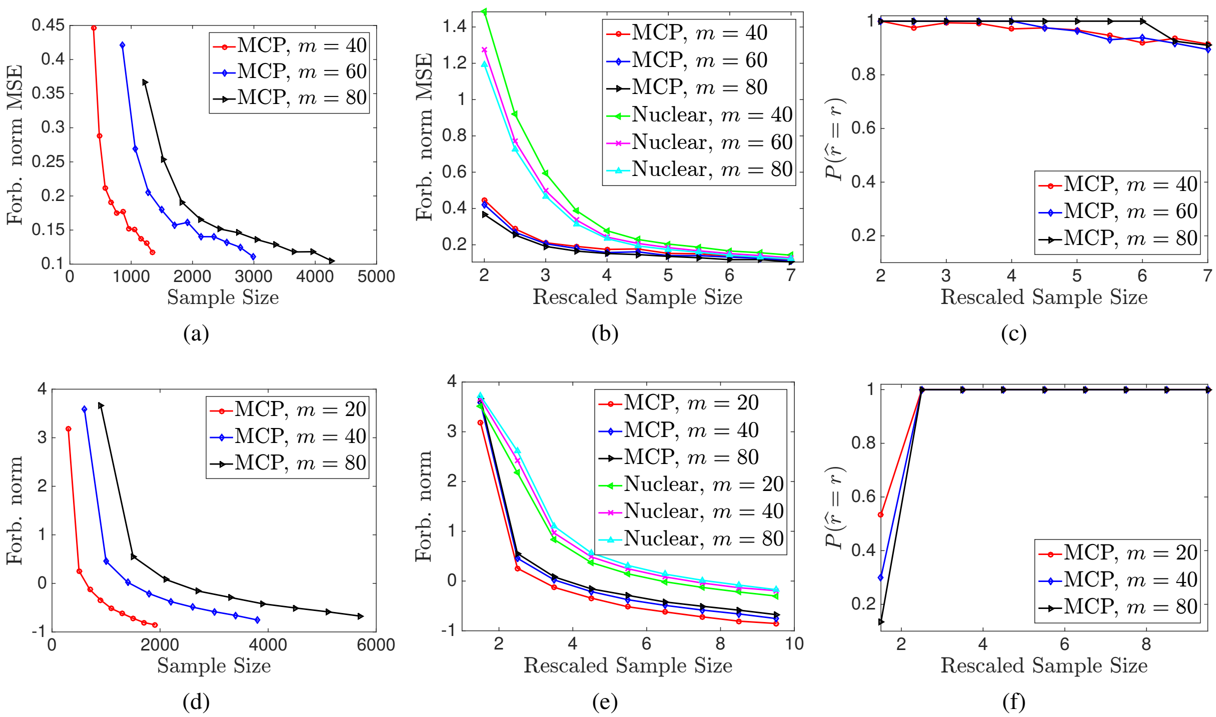

Figure 2. Simulation Results for Matrix Completion and Matrix Sensing with MCP penalty. Accordingly, the size of matrix and the rank are $m \times m$. The results of matrix completion, with rank $r = \lfloor \log^2 m \rfloor$, in Figure 2(a)-2(c) with the rescaled sample size $N = n/(rm \log m)$; while matrix sensing, for rank $r = 10$, is studied in Figure 2(d)-2(f) with rescaled sample size $N = n/(rm)$.

In detail, Figure 2(a)- 2(c) are the results for matrix completion. With the same settings as experiments shown in Figure 1, we have that the estimator with MCP penalty, a particular case of the proposed estimator with nonconvex penalty, behaviors in accordance with our theoretical analysis and outperforms the estimator with nuclear norm. For the other example, *i.e.*, matrix sensing, the results in Figure 2(d)- 2(f) manifest the superiority of the estimator with MCP penalty. Particularly, for both examples, we have with with high probability, the rank of the underlying matrix is recovered with high probability.

C. Background

For matrix $\mathbf{\Theta}^* \in \mathbb{R}^{m_1 \times m_2}$, which is exactly low-rank and has rank r, we have the singular value decomposition (SVD) form of $\mathbf{\Theta}^* = \mathbf{U}^* \mathbf{\Gamma}^* \mathbf{V}^{*\top}$, where $\mathbf{U}^* \in \mathbb{R}^{m_1 \times r}$, $\mathbf{V}^* \in \mathbb{R}^{m_2 \times r}$ are matrices consist of left and right singular vectors, and $\mathbf{\Gamma}^* = \mathrm{diag}(\gamma_1^*, \ldots, \gamma_r^*) \in \mathbb{R}^{r \times r}$. Based on $\mathbf{U}^*, \mathbf{V}^*$, we define the following two subspaces of $\mathbb{R}^{m_1 \times m_2}$:

$$\mathcal{F}(\mathbf{U}^*, \mathbf{V}^*) := \{\mathbf{\Delta} | \text{row}(\mathbf{\Delta}) \subseteq \mathbf{V}^* \text{ and } \text{col}(\mathbf{\Delta}) \subseteq \mathbf{U}^*\},$$

and

$$\mathcal{F}^{\perp}(\mathbf{U}^*, \mathbf{V}^*) := \{\mathbf{\Delta} | \text{row}(\mathbf{\Delta}) \perp \mathbf{V}^* \text{ and } \text{col}(\mathbf{\Delta}) \perp \mathbf{U}^*\},$$

where $\boldsymbol{\Delta} \in \mathbb{R}^{m_1 \times m_2}$ is an arbitrary matrix, and $\mathrm{row}(\boldsymbol{\Delta}) \subseteq \mathbb{R}^{m_2}$, $\mathrm{col}(\boldsymbol{\Delta}) \subseteq \mathbb{R}^{m_1}$ are the row space and column space of the matrix $\boldsymbol{\Delta}$. respectively. We will use the shorthand notation of $\mathcal{F}, \mathcal{F}^{\perp}$, when $(\mathbf{U}^*, \mathbf{V}^*)$ are clear from the context. Define $\boldsymbol{\Pi}_{\mathcal{F}}, \boldsymbol{\Pi}_{\mathcal{F}^{\perp}}$ as the projection operator onto the subspaces $\mathcal{F}$ and $\mathcal{F}^{\perp}$:

$$\begin{aligned} \boldsymbol{\Pi}_{\mathcal{F}}(\mathbf{A}) &= \mathbf{U}^*\mathbf{U}^{*\top}\mathbf{A}\mathbf{V}^*\mathbf{V}^{*\top}, \\ \boldsymbol{\Pi}_{\mathcal{F}^{\perp}}(\mathbf{A}) &= \big(\mathbf{I}_{m_1} - \mathbf{U}^*\mathbf{U}^{*\top}\big)\mathbf{A}\big(\mathbf{I}_{m_2} - \mathbf{V}^*\mathbf{V}^{*\top}\big). \end{aligned} \tag{C.1}$$

Thus, for all $\boldsymbol{\Delta} \in \mathbb{R}^{m_1 \times m_2}$, we have its orthogonal complement $\boldsymbol{\Delta}''$ with respect to the true low-rank matrix $\boldsymbol{\Theta}^*$ as follows:

$$\begin{aligned} \boldsymbol{\Delta}'' =& \big(\mathbf{I}_{m_1} - \mathbf{U}^*\mathbf{U}^{*\top}\big)\boldsymbol{\Delta}\big(\mathbf{I}_{m_2} - \mathbf{V}^*\mathbf{V}^{*\top}\big), \\ \boldsymbol{\Delta}' =& \boldsymbol{\Delta} - \boldsymbol{\Delta}'', \end{aligned} \tag{C.2}$$

where $\boldsymbol{\Delta}'$ is the component which has overlapped row and column space with $\boldsymbol{\Theta}^*$. (Negahban et al., 2012) gives detailed discussion about the concept of decomposibility and a large class of decomposable norms, among which the decomposability of the nuclear norm and Frobenius norm is relevant to our problem. For low-rank estimation, we have the equality that $\|\boldsymbol{\Theta}^* + \boldsymbol{\Delta}''\|_* = \|\boldsymbol{\Theta}^*\|_* + \|\boldsymbol{\Delta}''\|_*$ with $\boldsymbol{\Delta}''$ defined above.

D. Proof of the Main Results

D.1. Proof of Theorem 3.4

We first define $\widetilde{\mathcal{L}}_{n,\lambda}(\cdot)$ as follows,

$$\widetilde{\mathcal{L}}_{n,\lambda}(\boldsymbol{\Theta}) = \mathcal{L}_n(\boldsymbol{\Theta}) + \mathcal{Q}_\lambda(\boldsymbol{\Theta}).$$

Based on the the restrict strongly convexity of $\mathcal{L}_n$, and the curvature parameter of the non-convex penalty, if $\kappa(\mathfrak{X}) > \zeta_-$, we have the restrict strongly convexity of $\widetilde{\mathcal{L}}_{n,\lambda}(\cdot)$, as stated in the following lemma.

Lemma D.1. Under Assumption 3.1, if it is assumed that $\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2 \in \mathcal{C}$, we have

$$\widetilde{\mathcal{L}}_{n,\lambda}(\boldsymbol{\Theta}_2) \geq \widetilde{\mathcal{L}}_{n,\lambda}(\boldsymbol{\Theta}_1) + \langle \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\boldsymbol{\Theta}_1), \boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1 \rangle + \frac{\kappa(\mathfrak{X}) - \zeta_-}{2}\|\boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1\|_F^2.$$

Proof. Proof is provided in Section F.1. □

In the following, we prove that $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*$ lies in the cone $\mathcal{C}$, where

$$\mathcal{C} = \big\{\boldsymbol{\Delta} \in \mathbb{R}^{m_1 \times m_2} \big| \|\Pi_{\mathcal{F}^{\perp}}(\boldsymbol{\Delta})\|_* \leq 5\|\Pi_{\mathcal{F}}(\boldsymbol{\Delta})\|_*\big\}.$$

Lemma D.2. Under Assumption 3.1, the condition $\kappa(\mathfrak{X}) > \zeta_-$, and the regularization parameter $\lambda \geq 2\big\|\mathfrak{X}^*(\boldsymbol{\epsilon})\big\|_2/n$, we have

$$\big\|\Pi_{\mathcal{F}}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\big\|_* \leq 5\big\|\Pi_{\mathcal{F}^{\perp}}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\big\|_*.$$

Proof. Proof is provided in Section F.2. □

Now we are ready to prove Theorem 3.4.

Proof of Theorem 3.4. According to Lemma D.1, we have

$$\widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\boldsymbol{\Theta}}) \geq \widetilde{\mathcal{L}}_{n,\lambda}(\boldsymbol{\Theta}^*) + \langle \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\boldsymbol{\Theta}^*), \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^* \rangle + \frac{\kappa(\mathfrak{X}) - \zeta_-}{2}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2, \tag{D.1}$$

$$\widetilde{\mathcal{L}}_{n,\lambda}(\boldsymbol{\Theta}^*) \geq \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\boldsymbol{\Theta}}) + \langle \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\boldsymbol{\Theta}}), \boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}} \rangle + \frac{\kappa(\mathfrak{X}) - \zeta_-}{2}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}\|_F^2. \tag{D.2}$$

Meanwhile, since $\|\cdot\|_*$ is convex, we have

$$\lambda\|\widehat{\mathbf{\Theta}}\|_* \geq \lambda\|\mathbf{\Theta}^*\|_* + \lambda\langle\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*, \mathbf{W}^*\rangle, \tag{D.3}$$

$$\lambda\|\mathbf{\Theta}^*\|_* \geq \lambda\|\widehat{\mathbf{\Theta}}\|_* + \lambda\langle\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}, \mathbf{W}^*\rangle, \tag{D.4}$$

where $\mathbf{W}^* \in \|\mathbf{\Theta}^*\|_*$.

Adding (D.1) to (D.4), we have

$$0 \geq \big\langle \nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*) + \lambda\mathbf{W}^*, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\big\rangle + \big\langle \nabla\widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}) + \lambda\widehat{\mathbf{W}}, \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle + (\kappa(\mathfrak{X}) - \zeta_-)\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F^2.$$

Since $\widehat{\mathbf{\Theta}}$ is the solution to the SDP (2.2), $\widehat{\mathbf{\Theta}}$ satisfies the optimality condition (variational inequality), for any $\mathbf{\Theta}' \in \mathbb{R}^{m_1\times m_2}$, it holds that

$$\max_{\mathbf{\Theta}'} \big\langle \nabla\widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}) + \lambda\widehat{\mathbf{W}}, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}'\big\rangle \leq 0,$$

which implies

$$\big\langle \nabla\widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}) + \lambda\widehat{\mathbf{W}}, \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle \geq 0.$$

Hence,

$$\begin{aligned}
&(\kappa(\mathfrak{X}) - \zeta_-)\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F^2 \leq \big\langle \nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*) + \lambda\mathbf{W}^*, \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle \\
&\leq \Big\langle \mathbf{\Pi}_{\mathcal{F}^\perp}\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*) + \lambda\mathbf{W}^*\big), \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\Big\rangle + \Big\langle \mathbf{\Pi}_{\mathcal{F}}\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*) + \lambda\mathbf{W}^*\big), \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\Big\rangle.
\end{aligned} \tag{D.5}$$

Recall that $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}(\mathbf{\Theta}^*)$ is the vector of (ordered) singular values of $\mathbf{\Theta}^*$. In the following, we decompose (D.5) into three parts with regard to the magnitudes of the singular values of $\mathbf{\Theta}^*$.

(1) $i \in S^c$ that $(\boldsymbol{\gamma}^*)_i = 0$;

(2) $i \in S_1$ that $(\boldsymbol{\gamma}^*)_i \geq \nu$;

(3) $i \in S_2$ that $\nu > (\boldsymbol{\gamma}^*)_i > 0$.

Note that $S_1 \cup S_2 = S$.

(1) For $i \in S^c$, it correspond to the projector $\mathbf{\Pi}_{\mathcal{F}^\perp}(\cdot)$ since $\boldsymbol{\gamma}\big(\mathbf{\Pi}_{\mathcal{F}^\perp}(\mathbf{\Theta}^*)\big) = (\boldsymbol{\gamma}^*)_{S^c} = \mathbf{0}$.

Based on the regularity condition (iii) in Assumption 3.3 that $q_\lambda'(0) = 0$, we have that $\nabla\mathcal{Q}_\lambda(\mathbf{\Theta}^*) = \mathbf{U}^* q_\lambda'(\mathbf{\Gamma}^*)\mathbf{V}^{*\top}$ where $\mathbf{\Gamma}^* \in \mathbb{R}^{r\times r}$ is the diagonal matrix with $\operatorname{diag}(\mathbf{\Gamma}^*) = \boldsymbol{\gamma}^*$, we have

$$\begin{aligned}
\mathbf{\Pi}_{\mathcal{F}^\perp}(\nabla\mathcal{Q}_\lambda(\mathbf{\Theta}^*)) &= \big(\mathbf{I}_{m_1} - \mathbf{U}^*\mathbf{U}^{*\top}\big)\mathbf{U}^* q_\lambda'(\mathbf{\Gamma}^*)\mathbf{V}^{*\top}\big(\mathbf{I}_{m_2} - \mathbf{V}^*\mathbf{V}^{*\top}\big) \\
&= (\mathbf{U}^* - \mathbf{U}^*) q_\lambda'(\mathbf{\Gamma}^*)\big(\mathbf{V}^{*\top} - \mathbf{V}^{*\top}\big) \\
&= \mathbf{0}.
\end{aligned}$$

Therefore,

$$\mathbf{\Pi}_{\mathcal{F}^\perp}(\nabla\mathcal{Q}_\lambda(\mathbf{\Theta}^*)) = \mathbf{0}.$$

Meanwhile, we have

$$\big\|\mathbf{\Pi}_{\mathcal{F}^\perp}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2 \leq \big\|\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big\|_2 = \frac{\|\mathfrak{X}^*(\boldsymbol{\epsilon})\|_2}{n} \leq \lambda.$$

For $\mathbf{Z}^* = -\lambda^{-1}\mathbf{\Pi}_{\mathcal{F}^\perp}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)$, we have $\mathbf{W}^* = \mathbf{U}^*\mathbf{V}^{*\top} + \mathbf{Z}^* \in \partial\|\mathbf{\Theta}^*\|_*$ because $\|\mathbf{Z}^*\|_2 \leq 1$ and $\mathbf{Z}^* \in \mathcal{F}^\perp$, which satisfies the condition of $\mathbf{W}^*$ to be subgradient of $\|\mathbf{\Theta}^*\|_*$. With this particular choice of $\mathbf{W}^*$, we have

$$\mathbf{\Pi}_{\mathcal{F}^\perp}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*) + \lambda\mathbf{W}^*\big) = \mathbf{\Pi}_{\mathcal{F}^\perp}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big) + \lambda\mathbf{Z}^* = \mathbf{0},$$

which implies that

$$\big\langle \mathbf{\Pi}_{\mathcal{F}^\perp}\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*) + \lambda\mathbf{W}^*\big), \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle = \langle \mathbf{0}, \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\rangle = 0. \tag{D.6}$$

(2) Consider $i \in S_1$ that $(\boldsymbol{\gamma}^*)_i \geq \nu$. Let $|S_1| = r_1$. Define a subspace of $\mathcal{F}$ associated with S_1 as follows

$$\mathcal{F}_{S_1}(\mathbf{U}^*, \mathbf{V}^*) := \{\mathbf{\Delta} \in \mathbb{R}^{m_1\times m_2} | \text{row}(\mathbf{\Delta}) \subset \mathbf{V}^*_{S_1} \text{ and } \text{col}(\mathbf{\Delta}) \subset \mathbf{U}^*_{S_1}\},$$

where $\mathbf{U}^*_{S_1}$ and $\mathbf{V}^*_{S_1}$ is the matrix with the i^{th} row of $\mathbf{U}^*$ and $\mathbf{V}^*$ where $i \in S_1$.

Recall that $\mathcal{P}_\lambda(\mathbf{\Theta}^*) = \mathcal{Q}_\lambda(\mathbf{\Theta}^*) + \lambda\|\mathbf{\Theta}^*\|_*$. We have

$$\nabla\mathcal{P}_\lambda(\mathbf{\Theta}^*) = \nabla\mathcal{Q}_\lambda(\mathbf{\Theta}^*) + \lambda(\mathbf{U}^*\mathbf{V}^{*\top} + \mathbf{Z}^*).$$

Projecting $\nabla\mathcal{P}_\lambda(\mathbf{\Theta}^*)$ into the subspace $\mathcal{F}_{S_1}$, we have

$$\begin{aligned}
\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\mathcal{P}_\lambda(\mathbf{\Theta}^*)\big) &= \mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\mathcal{Q}_\lambda(\mathbf{\Theta}^*) + \lambda\mathbf{U}^*\mathbf{V}^{*\top} + \lambda\mathbf{Z}^*\big) \\
&= \mathbf{U}^*_{S_1} q'_\lambda(\mathbf{\Gamma}^*_{S_1})(\mathbf{V}^*_{S_1})^\top + \lambda\mathbf{U}^*_{S_1}(\mathbf{V}^*_{S_1})^\top \\
&= \mathbf{U}^*_{S_1}\big(q'_\lambda(\mathbf{\Gamma}^*_{S_1}) + \lambda\mathbf{I}_{S_1}\big)(\mathbf{V}^*_{S_1})^\top,
\end{aligned}$$

where $\mathbf{\Gamma}^*_{S_1} \in \mathbb{R}^{r_1\times r_1}$ and $\big(q'_\lambda(\mathbf{\Gamma}^*_{S_1}) + \lambda\mathbf{I}_{S_1}\big)$ is a diagonal matrix that $\big(q'_\lambda(\mathbf{\Gamma}^*_{S_1}) + \lambda\mathbf{I}_{S_1}\big)_{ii} = 0$ for $i \notin S_1$, and for all $i \in S_1$,

$$\big(q'_\lambda(\mathbf{\Gamma}^*_{S_1}) + \lambda\mathbf{I}_{S_1}\big)_{ii} = q'_\lambda\big((\boldsymbol{\gamma}^*)_i\big) + \lambda = p'_\lambda\big((\boldsymbol{\gamma}^*)_i\big) = 0,$$

where the last equality is because $p_\lambda(\cdot)$ satisfies the regularity condition (i) with $(\boldsymbol{\gamma}^*)_i \geq \nu$ for $i \in S_1$. Thus, we have $q'_\lambda(\mathbf{D}_{S_1}) + \lambda\mathbf{I}_{S_1} = \mathbf{0}$, which indicates that $\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\mathcal{P}_\lambda(\mathbf{\Theta}^*)\big) = \mathbf{0}$. Therefore, we have

$$\begin{aligned}
\big\langle \mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*) + \lambda\mathbf{W}^*\big), \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle &= \big\langle \mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*) + \nabla\mathcal{P}_\lambda(\mathbf{\Theta}^*)\big), \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle \\
&= \big\langle \mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big), \mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big)\big\rangle \\
&\leq \big\|\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2 \cdot \big\|\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big)\big\|_*,
\end{aligned}$$

where the last inequality is derived from the Hölder inequality. What remains is to bound $\big\|\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big)\big\|_*$. By the properties of projection on to the subspace $\mathcal{F}_{S_1}$, we have

$$\big\|\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big)\big\|_* \leq \sqrt{r_1}\big\|\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big)\big\|_F \leq \sqrt{r_1}\big\|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\|_F,$$

where the second inequality is due to the fact that $\text{rank}\big(\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big)\big) \leq r_1$. Therefore, we have

$$\big\langle \mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*) + \lambda\mathbf{W}^*\big), \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle \leq \sqrt{r_1}\big\|\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2 \cdot \big\|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\|_F. \tag{D.7}$$

(3) Finally, consider $i \in S_2$ that $(\boldsymbol{\gamma}^*)_i \leq \nu$. Let $|S_2| = r_2$. Define a subspace of $\mathcal{F}$ associated with S_2 as follows

$$\mathcal{F}_{S_2}(\mathbf{U}^*, \mathbf{V}^*) := \big\{\mathbf{\Delta} \in \mathbb{R}^{m_1\times m_2} | \text{row}(\mathbf{\Delta}) \subset \mathbf{V}^*_{S_2} \text{ and } \text{col}(\mathbf{\Delta}) \subset \mathbf{U}^*_{S_2}\big\},$$

where $\mathbf{U}^*_{S_2}$ and $\mathbf{V}^*_{S_2}$ is the matrix with the i^{th} row of $\mathbf{U}^*$ and $\mathbf{V}^*$ where $i \in S_2$. It is obvious that for all $\mathbf{\Delta} \in \mathbb{R}^{m_1\times m_2}$, the following decomposition holds

$$\mathbf{\Pi}_{\mathcal{F}}(\mathbf{\Delta}) = \mathbf{\Pi}_{\mathcal{F}_{S_1}}(\mathbf{\Delta}) + \mathbf{\Pi}_{\mathcal{F}_{S_2}}(\mathbf{\Delta}).$$

In addition, since $\mathbf{U}^*$, $\mathbf{V}^*$ are unitary matrices, we have

$$\mathcal{F}_{S_1} \subset \mathcal{F}^\perp_{S_2}, \text{ and } \mathcal{F}_{S_2} \subset \mathcal{F}^\perp_{S_1},$$

where $\mathcal{F}_{S_1}^{\perp}, \mathcal{F}_{S_2}^{\perp}$ denote the complementary subspace of $\mathcal{F}_{S_1}$ and $\mathcal{F}_{S_2}$, respectively. Similar to analysis in (2) on S_1, we have

$$\mathbf{\Pi}_{\mathcal{F}_{S_2}}\big(\nabla\mathcal{Q}_\lambda(\mathbf{\Theta}^*)\big) = \mathbf{U}_{S_2}^* q_\lambda'(\mathbf{\Gamma}_{S_2}^*)(\mathbf{V}_{S_2}^*)^\top,$$

where $q_\lambda'(\mathbf{\Gamma}_{S_2}^*)$ is a diagonal matrix that $\big(q_\lambda'(\mathbf{\Gamma}_{S_2}^*)\big)_{ii} = 0$ for $i \notin S_2$, and for all $i \in S_2$, $\big(q_\lambda'(\mathbf{\Gamma}_{S_2}^*)\big)_{ii} = q_\lambda'\big((\gamma^*)_i\big) \le \lambda$, since $(\gamma^*)_i \le \nu$ and $q_\lambda(\cdot)$ satisfies the regularity condition (iv). Therefore

$$\big\|\mathbf{\Pi}_{\mathcal{F}_{S_2}}\big(\nabla\mathcal{Q}_\lambda(\mathbf{\Theta}^*)\big)\big\|_2 = \max_{i\in S_2}\big(q_\lambda'(\mathbf{\Gamma}_{S_2}^*)\big)_{ii} \le \lambda. \tag{D.8}$$

Meanwhile, we have

$$\big\|\mathbf{\Pi}_{\mathcal{F}_{S_2}}(\lambda\mathbf{W}^*)\big\|_2 \le \big\|\mathbf{\Pi}_{\mathcal{F}}\big(\lambda\mathbf{U}^*\mathbf{V}^{*\top}\big)\big\|_2 = \lambda, \tag{D.9}$$

where the first inequality is due the fact that $\mathcal{F}_{S_2} \in \mathcal{F}$, and last equality comes from the fact that $\big\|\mathbf{U}^*\mathbf{V}^{*\top}\big\|_2 = 1$. Therefore, we have

$$\big\|\mathbf{\Pi}_{\mathcal{F}_{S_2}}(\lambda\mathbf{W}^*)\big\|_2 \le \lambda. \tag{D.10}$$

In addition, we have the fact that $\big\|\mathbf{\Pi}_{\mathcal{F}_{S_2}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2 \le \big\|\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big\|_2 \le \lambda.$, which indicates that

$$\begin{aligned}
&\big\langle\mathbf{\Pi}_{\mathcal{F}_{S_2}}\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*) + \lambda\mathbf{W}^*\big), \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle = \big\langle\mathbf{\Pi}_{\mathcal{F}_{S_2}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*) + \nabla\mathcal{Q}_\lambda(\mathbf{\Theta}^*) + \lambda\mathbf{W}^*\big), \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle \\
&= \big\langle\mathbf{\Pi}_{\mathcal{F}_{S_2}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big), \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle + \big\langle\mathbf{\Pi}_{\mathcal{F}_{S_2}}\big(\nabla\mathcal{Q}_\lambda(\mathbf{\Theta}^*)\big), \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle + \big\langle\mathbf{\Pi}_{\mathcal{F}_{S_2}}(\lambda\mathbf{W}^*), \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle \\
&\le \Big[\big\|\mathbf{\Pi}_{\mathcal{F}_{S_2}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2 + \big\|\mathbf{\Pi}_{\mathcal{F}_{S_2}}\big(\nabla\mathcal{Q}_\lambda(\mathbf{\Theta}^*)\big)\big\|_2 + \big\|\mathbf{\Pi}_{\mathcal{F}_{S_2}}(\lambda\mathbf{W}^*)\big\|_2\Big]\big\|\mathbf{\Pi}_{\mathcal{F}_{S_2}}(\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}})\big\|_*,
\end{aligned}$$

where the last inequality is due to Hölder's inequality. Since we have obtained the bound for each term, as in (D.8), (D.9), (D.10), we have

$$\begin{aligned}
\big\langle\mathbf{\Pi}_{\mathcal{F}_{S_2}}\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*) + \lambda\mathbf{W}^*\big), \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle &\le 3\lambda\|\mathbf{\Pi}_{\mathcal{F}_{S_2}}(\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}})\|_* \\
&\le 3\lambda\sqrt{r_2}\|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\|_F,
\end{aligned} \tag{D.11}$$

where the last inequality utilizes the fact that $\text{rank}(\mathbf{\Pi}_{\mathcal{F}_{S_2}}(\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}})) \le r_2$.

Adding (D.6), (D.7), and (D.11), we have

$$\begin{aligned}
\big(\kappa(\mathfrak{X}) - \zeta_-\big)\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F^2 &\le \big\langle\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*) + \lambda\mathbf{W}^*, \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\rangle \\
&\le \sqrt{r_1}\big\|\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2 \cdot \big\|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\big\|_F + 3\lambda\sqrt{r_2}\|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\|_F,
\end{aligned}$$

which indicate that

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F \le \frac{\sqrt{r_1}}{\kappa(\mathfrak{X}) - \zeta_-}\big\|\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2 + \frac{3\lambda\sqrt{r_2}}{\kappa(\mathfrak{X}) - \zeta_-}.$$

This completes the proof. □

D.2. Proof of Theorem 3.5

Before presenting the proof of Theorem 3.5, we need the following lemma.

Lemma D.3 (Deterministic Bound)**.** Suppose $\mathbf{\Theta}^* \in \mathbb{R}^{m_1\times m_2}$ has rank r, $\mathfrak{X}(\cdot)$ satisfies RSC with respect to $\mathcal{C}$. Then the error bound between the oracle estimator $\widehat{\mathbf{\Theta}}_O$ and true $\mathbf{\Theta}^*$ satisfies

$$\|\widehat{\mathbf{\Theta}}_O - \mathbf{\Theta}^*\|_F \le \frac{2\sqrt{r}\big\|\Pi_{\mathcal{F}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2}{\kappa(\mathfrak{X})}, \tag{D.12}$$

Proof. Proof is provided in Section F.3. □

Proof of Theorem 3.5. Suppose $\widehat{\mathbf{W}} \in \partial\|\widehat{\boldsymbol{\Theta}}\|_*$, since $\widehat{\boldsymbol{\Theta}}$ is the solution to the SDP (2.2), the variational inequality yields

$$\max_{\boldsymbol{\Theta}'} \big\langle \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}', \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\boldsymbol{\Theta}}) + \lambda \widehat{\mathbf{W}} \big\rangle \le 0. \tag{D.13}$$

In the following, we will show that there exists some $\widehat{\mathbf{W}}_O \in \partial\|\widehat{\boldsymbol{\Theta}}_O\|_*$ such that, for all $\boldsymbol{\Theta}' \in \mathbb{R}^{m_1\times m_2}$,

$$\max_{\boldsymbol{\Theta}'} \big\langle \widehat{\boldsymbol{\Theta}}_O - \boldsymbol{\Theta}', \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\boldsymbol{\Theta}}_O) + \lambda \widehat{\mathbf{W}}_O \big\rangle \le 0. \tag{D.14}$$

Recall that $\widetilde{\mathcal{L}}_{n,\lambda}(\boldsymbol{\Theta}) = \mathcal{L}_n(\boldsymbol{\Theta}) + \mathcal{Q}_\lambda(\boldsymbol{\Theta})$. By projecting the components of the inner product of the LHS in (D.14) into two complementary spaces $\mathcal{F}$ and $\mathcal{F}^\perp$, we have the following decomposition

$$\begin{aligned}
&\big\langle \widehat{\boldsymbol{\Theta}}_O - \boldsymbol{\Theta}', \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\boldsymbol{\Theta}}_O) + \lambda \widehat{\mathbf{W}}_O \big\rangle \\
&= \underbrace{\big\langle \boldsymbol{\Pi}_{\mathcal{F}}(\widehat{\boldsymbol{\Theta}}_O - \boldsymbol{\Theta}'), \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\boldsymbol{\Theta}}_O) + \lambda \widehat{\mathbf{W}}_O \big\rangle}_{I_1} + \underbrace{\big\langle \boldsymbol{\Pi}_{\mathcal{F}^\perp}(\widehat{\boldsymbol{\Theta}}_O - \boldsymbol{\Theta}'), \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\boldsymbol{\Theta}}_O) + \lambda \widehat{\mathbf{W}}_O \big\rangle}_{I_2}.
\end{aligned} \tag{D.15}$$

Analysis of Term I_1. Let $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}(\boldsymbol{\Theta}^*)$, $\widehat{\boldsymbol{\gamma}}_O = \boldsymbol{\gamma}(\widehat{\boldsymbol{\Theta}}_O)$ be the vector of (ordered) singular values of $\boldsymbol{\Theta}^*$ and $\widehat{\boldsymbol{\Theta}}_O$, respectively. By the perturbation bounds for singular values, the Weyl's inequality (Weyl, 1912), we have that

$$\max_i \big|(\boldsymbol{\gamma}^*)_i - (\widehat{\boldsymbol{\gamma}}_O)_i\big| \le \big\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_O\big\|_2 \le \big\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_O\big\|_F.$$

Since Lemma D.3 provides the Frobenius norm on the estimation error $\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_O$, we obtain that

$$\max_i \big|(\boldsymbol{\gamma}^*)_i - (\widehat{\boldsymbol{\gamma}}_O)_i\big| \le \frac{2\sqrt{r}}{n\kappa(\mathfrak{X})}\|\mathfrak{X}^*(\boldsymbol{\epsilon})\|_2.$$

If it is assumed that $S = \mathrm{supp}(\boldsymbol{\sigma}^*)$, we have $|S| = r$. The triangle inequality yields that

$$\begin{aligned}
\min_{i\in S} \big|(\widehat{\boldsymbol{\gamma}}_O)_i\big| &= \min_{i\in S}\big|(\widehat{\boldsymbol{\gamma}}_O)_i - (\boldsymbol{\gamma}^*)_i + (\boldsymbol{\gamma}^*)_i\big| \ge -\max_{i\in S}\big|(\widehat{\boldsymbol{\gamma}}_O - \boldsymbol{\gamma}^*)_i\big| + \min_{i\in S}\big|(\boldsymbol{\gamma}^*)_i\big| \\
&\ge -\frac{2\sqrt{r}}{n\kappa(\mathfrak{X})}\|\mathfrak{X}^*(\boldsymbol{\epsilon})\|_2 + \nu + \frac{2\sqrt{r}}{n\kappa(\mathfrak{X})}\|\mathfrak{X}^*(\boldsymbol{\epsilon})\|_2 \\
&= \nu,
\end{aligned}$$

where the inequality on the second line is derived based on the condition that $\min_{i\in S}\big|(\boldsymbol{\gamma}^*)_i\big| \ge \nu + 2n^{-1}\sqrt{r}\|\mathfrak{X}^*(\boldsymbol{\epsilon})\|_*/\kappa(\mathfrak{X})$. Based on the definition of oracle estimator (3.2), $\widehat{\boldsymbol{\Theta}}_O \in \mathcal{F}$, which implies $\mathrm{rank}(\widehat{\boldsymbol{\Theta}}_O) = r$. Therefore, we have

$$(\widehat{\boldsymbol{\gamma}}_O)_1 \ge (\widehat{\boldsymbol{\gamma}}_O)_2 \ge \ldots \ge (\widehat{\boldsymbol{\gamma}}_O)_r \ge \nu > 0 = (\widehat{\boldsymbol{\gamma}}_O)_{r+1} = (\widehat{\boldsymbol{\gamma}}_O)_m = 0. \tag{D.16}$$

By the definition of Oracle estimator, we have $\widehat{\boldsymbol{\Theta}}_O = \mathbf{U}^*\widehat{\boldsymbol{\Gamma}}_O\mathbf{V}^{*\top}$, where $\widehat{\boldsymbol{\Gamma}}_O$ is the diagonal matrix with $\mathrm{diag}(\widehat{\boldsymbol{\Gamma}}_O) = \widehat{\boldsymbol{\gamma}}_O$. Since $\mathcal{P}_\lambda(\boldsymbol{\Theta}) = \mathcal{Q}_\lambda(\boldsymbol{\Theta}) + \lambda\|\boldsymbol{\Theta}\|_*$, we have

$$\begin{aligned}
\boldsymbol{\Pi}_{\mathcal{F}}\big(\nabla\mathcal{P}_\lambda(\widehat{\boldsymbol{\Theta}}_O)\big) &= \boldsymbol{\Pi}_{\mathcal{F}}\big(\nabla\mathcal{Q}_\lambda(\widehat{\boldsymbol{\Theta}}_O) + \lambda\partial\|\widehat{\boldsymbol{\Theta}}_O\|_*\big) \\
&= \boldsymbol{\Pi}_{\mathcal{F}}\big(\mathbf{U}^* q_\lambda'(\widehat{\boldsymbol{\Gamma}}_O)\mathbf{V}^{*\top} + \lambda\mathbf{U}^*\mathbf{V}^{*\top} + \lambda\widehat{\mathbf{Z}}_O\big) \\
&= \mathbf{U}^*\Big(q_\lambda'\big((\widehat{\boldsymbol{\Gamma}}_O)_S\big) + \lambda\mathbf{I}_r\Big)\mathbf{V}^{*\top},
\end{aligned} \tag{D.17}$$

where $\widehat{\mathbf{Z}}_O \in \mathcal{F}^\perp$, $\|\widehat{\mathbf{Z}}_O\|_2 \le 1$, and $(\widehat{\boldsymbol{\Gamma}}_O)_S \in \mathbb{R}^{r\times r}$ is a diagonal matrix with $\mathrm{diag}\big((\widehat{\boldsymbol{\Gamma}}_O)_S\big) = (\widehat{\boldsymbol{\gamma}}_O)_S$. The first equality in (D.17) is based on the definition of $\nabla\mathcal{Q}_\lambda(\cdot)$ and $\partial\|\cdot\|_*$, while the second is to simply project each component into the subspace $\mathcal{F}$. Since $p_\lambda(t) = q_\lambda(t) + \lambda|t|$, we have $p_\lambda'(t) = q_\lambda'(t) + \lambda t$ for all $t > 0$. Consider the diagonal matrix $q_\lambda'\big((\widehat{\boldsymbol{\Gamma}}_O)_S\big) + \lambda\mathbf{I}_r$, we have the i^{th} ($i \in S$) element on the diagonal that

$$\Big(q_\lambda'\big((\widehat{\boldsymbol{\Gamma}}_O)_S\big) + \lambda\mathbf{I}_r\Big)_{ii} = q_\lambda'\big((\widehat{\boldsymbol{\gamma}}_O)_i\big) + \lambda = p_\lambda'\big((\widehat{\boldsymbol{\gamma}}_O)_i\big).$$

Since $p_\lambda(\cdot)$ satisfies the regularity condition (ii), that $p'_\lambda(t) = 0$ for all $t \geq \nu$, we have $p'_\lambda\big((\widehat{\gamma}_O)_i\big) = 0$ for $i \in S$, in light of the fact that $(\widehat{\gamma}_O)_i \geq \nu > 0$. Therefore, the diagonal matrix $q'_\lambda\big((\widehat{\boldsymbol{\Gamma}}_O)_S\big) + \lambda \mathbf{I}_r = \mathbf{0}$, substituting which into (D.17) yields

$$\boldsymbol{\Pi}_{\mathcal{F}}\big(\nabla \mathcal{P}_\lambda(\widehat{\boldsymbol{\Theta}}_O)\big) = \mathbf{0}. \tag{D.18}$$

Since $\widehat{\boldsymbol{\Theta}}_O$ is a minimizer of (3.2) over $\mathcal{F}$, we have the following optimality condition that for all $\boldsymbol{\Theta}' \in \mathbb{R}^{m_1 \times m_2}$,

$$\max_{\boldsymbol{\Theta}'} \big\langle \boldsymbol{\Pi}_{\mathcal{F}}(\widehat{\boldsymbol{\Theta}}_O - \boldsymbol{\Theta}'), \nabla \mathcal{L}_n(\widehat{\boldsymbol{\Theta}}_O)\big\rangle \leq 0. \tag{D.19}$$

Substitute (D.18) and (D.19) into item I_1, we have for all $\widehat{\mathbf{W}}_O \in \partial \|\widehat{\boldsymbol{\Theta}}_O\|_*$,

$$\begin{aligned}
&\max_{\boldsymbol{\Theta}'} \big\langle \boldsymbol{\Pi}_{\mathcal{F}}(\widehat{\boldsymbol{\Theta}}_O - \boldsymbol{\Theta}'), \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\boldsymbol{\Theta}}_O) + \lambda \widehat{\mathbf{W}}_O\big\rangle \\
&= \max_{\boldsymbol{\Theta}'} \big\langle \boldsymbol{\Pi}_{\mathcal{F}}(\widehat{\boldsymbol{\Theta}}_O - \boldsymbol{\Theta}'), \nabla \mathcal{L}_n(\widehat{\boldsymbol{\Theta}}_O)\big\rangle + \max_{\boldsymbol{\Theta}'} \big\langle \boldsymbol{\Pi}_{\mathcal{F}}(\widehat{\boldsymbol{\Theta}}_O - \boldsymbol{\Theta}'), \boldsymbol{\Pi}_{\mathcal{F}}\big(\nabla \mathcal{P}_\lambda(\widehat{\boldsymbol{\Theta}}_O)\big)\big\rangle \\
&\leq 0.
\end{aligned} \tag{D.20}$$

Analysis of Term I_2. By definition of $\nabla \mathcal{Q}_\lambda(\boldsymbol{\Theta})$, and the condition that $q'_\lambda(\cdot)$ satisfies the regularity condition (iii) in Assumption 3.3, we have the SVD of $\nabla \mathcal{Q}_\lambda(\boldsymbol{\Theta}_O)$ as $\nabla \mathcal{Q}_\lambda(\widehat{\boldsymbol{\Theta}}_O) = \mathbf{U}^* q'_\lambda(\widehat{\boldsymbol{\Gamma}}_O)\mathbf{V}^{*\top}$, where $\widehat{\boldsymbol{\Gamma}}_O \in \mathbb{R}^{r \times r}$ is a diagonal matrix. Projecting $\nabla \mathcal{Q}_\lambda(\widehat{\boldsymbol{\Theta}}_O)$ into $\mathcal{F}^\perp$ yields that

$$\begin{aligned}
\boldsymbol{\Pi}_{\mathcal{F}^\perp}\big(\nabla \mathcal{Q}_\lambda(\widehat{\boldsymbol{\Theta}}_O)\big) &= \big(\mathbf{I}_{m_1} - \mathbf{U}^*\mathbf{U}^{*\top}\big)\mathbf{U}^* q'_\lambda\big((\widehat{\boldsymbol{\Gamma}}_O)\big)\mathbf{V}^{*\top}\big(\mathbf{I}_{m_1} - \mathbf{V}^*\mathbf{V}^{*\top}\big) \\
&= \big(\mathbf{U}^* - \mathbf{U}^*\big) q'_\lambda\big((\widehat{\boldsymbol{\Gamma}}_O)_{S^c}\big)\big(\mathbf{V}^{*\top} - \mathbf{V}^{*\top}\big) \\
&= \mathbf{0}.
\end{aligned}$$

Thus,

$$\boldsymbol{\Pi}_{\mathcal{F}^\perp}\big(\nabla \mathcal{Q}_\lambda(\widehat{\boldsymbol{\Theta}}_O)\big) = \mathbf{0}. \tag{D.21}$$

Therefore,

$$I_2 = \big\langle \boldsymbol{\Pi}_{\mathcal{F}^\perp}(-\boldsymbol{\Theta}'), \boldsymbol{\Pi}_{\mathcal{F}^\perp}\big(\nabla \mathcal{L}_n(\widehat{\boldsymbol{\Theta}}_O) + \lambda \widehat{\mathbf{W}}_O\big)\big\rangle.$$

Moreover, the triangle inequality yields

$$\begin{aligned}
\|\nabla \mathcal{L}_n(\widehat{\boldsymbol{\Theta}}_O)\|_2 &\leq \|\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*)\|_2 + \|\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*) - \nabla \mathcal{L}_n(\widehat{\boldsymbol{\Theta}}_O)\|_2 \\
&\leq \|\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*)\|_2 + \|\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*) - \nabla \mathcal{L}_n(\widehat{\boldsymbol{\Theta}}_O)\|_F \\
&\leq \|\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*)\|_2 + \rho(\mathfrak{X})\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_O\|_F,
\end{aligned} \tag{D.22}$$

where the second inequality comes from the fact that $\|\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*) - \nabla \mathcal{L}_n(\widehat{\boldsymbol{\Theta}}_O)\|_2 \leq \|\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*) - \nabla \mathcal{L}_n(\widehat{\boldsymbol{\Theta}}_O)\|_F$, while the last inequality is obtained by the restricted strong smoothness (Assumption 3.2), which is equivalent to

$$\|\nabla \mathcal{L}_n(\boldsymbol{\Theta}) - \nabla \mathcal{L}_n(\boldsymbol{\Theta} + \widehat{\boldsymbol{\Delta}}_O)\|_F \leq \rho(\mathfrak{X})\|\widehat{\boldsymbol{\Delta}}_O\|_F,$$

over the restricted set $\mathcal{C}$; since $\Pi_{\mathcal{F}^\perp}(\widehat{\boldsymbol{\Delta}}_O) = \mathbf{0}$, it is evident that $\widehat{\boldsymbol{\Delta}}_O \in \mathcal{C}$.

Substitute (D.12) of Lemma D.3 into (D.22), we have

$$\Big\|\boldsymbol{\Pi}_{\mathcal{F}^\perp}\big(\nabla \mathcal{L}_n(\widehat{\boldsymbol{\Theta}}_O)\big)\Big\|_2 \leq \big\|\nabla \mathcal{L}_n(\widehat{\boldsymbol{\Theta}}_O)\big\|_2 \leq \big\|\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*)\big\|_2 + \frac{2\sqrt{r}\rho(\mathfrak{X})}{n\kappa(\mathfrak{X})}\big\|\mathfrak{X}^*(\boldsymbol{\epsilon})\big\|_2 \leq \lambda,$$

where the last inequality follows from the choice of λ.

By setting $\widehat{\mathbf{Z}}_O = -\lambda^{-1}\boldsymbol{\Pi}_{\mathcal{F}^\perp}\big(\nabla \mathcal{L}_n(\widehat{\boldsymbol{\Theta}}_O)\big)$, such that $\widehat{\mathbf{W}}_O = \mathbf{U}^*\mathbf{V}^{*\top} + \widehat{\mathbf{Z}}_O \in \partial\|\widehat{\boldsymbol{\Theta}}_O\|_*$ since $\widehat{\mathbf{Z}}_O$ satisfies the condition $\widehat{\mathbf{Z}}_O \in \mathcal{F}^\perp$, $\|\widehat{\mathbf{Z}}_O\|_2 \leq 1$, we have

$$\boldsymbol{\Pi}_{\mathcal{F}^\perp}\big(\nabla \mathcal{L}_n(\widehat{\boldsymbol{\Theta}}_O) + \lambda \widehat{\mathbf{W}}_O\big) = \mathbf{0},$$

which implies that

$$I_2 = \big\langle \mathbf{\Pi}_{\mathcal{F}^\perp}(-\mathbf{\Theta}'), \mathbf{0} \big\rangle = 0. \tag{D.23}$$

Substitute (D.20) and (D.23) into (D.15), we obtain (D.14) that

$$\max_{\mathbf{\Theta}'} \big\langle \widehat{\mathbf{\Theta}}_O - \mathbf{\Theta}', \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}_O) + \lambda \widehat{\mathbf{W}}_O \big\rangle \le 0.$$

Now we are going to prove that $\widehat{\mathbf{\Theta}}_O = \mathbf{\Theta}^*$.

Applying Lemma D.1, we have

$$\widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}) \ge \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}_O) + \big\langle \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}_O), \widehat{\mathbf{\Theta}} - \widehat{\mathbf{\Theta}}_O \big\rangle + \frac{\kappa(\mathfrak{X}) - \zeta_-}{2} \|\widehat{\mathbf{\Theta}}_O - \widehat{\mathbf{\Theta}}\|_F^2, \tag{D.24}$$

$$\widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}_O) \ge \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}) + \big\langle \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}), \widehat{\mathbf{\Theta}}_O - \widehat{\mathbf{\Theta}} \big\rangle + \frac{\kappa(\mathfrak{X}) - \zeta_-}{2} \|\widehat{\mathbf{\Theta}}_O - \widehat{\mathbf{\Theta}}\|_F^2. \tag{D.25}$$

On the other hand, because of the convexity of nuclear norm $\|\cdot\|_*$, we obtain

$$\lambda \|\widehat{\mathbf{\Theta}}\|_* \ge \lambda \|\widehat{\mathbf{\Theta}}_O\|_* + \lambda \langle \widehat{\mathbf{\Theta}} - \widehat{\mathbf{\Theta}}_O, \widehat{\mathbf{W}}_O \rangle, \tag{D.26}$$

$$\lambda \|\widehat{\mathbf{\Theta}}_O\|_* \ge \lambda \|\widehat{\mathbf{\Theta}}\|_* + \lambda \langle \widehat{\mathbf{\Theta}}_O - \widehat{\mathbf{\Theta}}, \widehat{\mathbf{W}} \rangle. \tag{D.27}$$

Add (D.24) to (D.27), we obtain

$$0 \ge \underbrace{\big\langle \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}) + \lambda \widehat{\mathbf{W}}, \widehat{\mathbf{\Theta}}_O - \widehat{\mathbf{\Theta}} \big\rangle}_{I_3} + \underbrace{\big\langle \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}_O) + \lambda \widehat{\mathbf{W}}_O, \widehat{\mathbf{\Theta}} - \widehat{\mathbf{\Theta}}_O \big\rangle}_{I_4} + \big(\kappa(\mathfrak{X}) - \zeta_-\big) \|\widehat{\mathbf{\Theta}}_O - \widehat{\mathbf{\Theta}}\|_F^2. \tag{D.28}$$

Analysis of Term I_3. By (D.13), we have

$$\big\langle \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}) + \lambda \widehat{\mathbf{W}}, \widehat{\mathbf{\Theta}} - \widehat{\mathbf{\Theta}}_O \big\rangle \le \max_{\mathbf{\Theta}'} \big\langle \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}) + \lambda \widehat{\mathbf{W}}, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}' \big\rangle \le 0. \tag{D.29}$$

Therefore $I_3 \ge 0$.

Analysis of Term I_4. By (D.14), we have

$$\big\langle \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}_O) + \lambda \widehat{\mathbf{W}}_O, \widehat{\mathbf{\Theta}}_O - \widehat{\mathbf{\Theta}} \big\rangle \le \max_{\mathbf{\Theta}'} \big\langle \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}_O) + \lambda \widehat{\mathbf{W}}_O, \widehat{\mathbf{\Theta}}_O - \mathbf{\Theta}' \big\rangle \le 0. \tag{D.30}$$

Therefore $I_4 \ge 0$. Substituting (D.29) and (D.30) into (D.28) yields that

$$\big(\kappa(\mathfrak{X}) - \zeta_-\big) \|\widehat{\mathbf{\Theta}}_O - \widehat{\mathbf{\Theta}}\|_F^2 \le 0,$$

which holds if and only if

$$\widehat{\mathbf{\Theta}}_O = \widehat{\mathbf{\Theta}}, \tag{D.31}$$

because $\kappa(\mathfrak{X}) > \zeta_-$.

By Lemma D.3, we obtain the error bound

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F = \|\widehat{\mathbf{\Theta}}_O - \mathbf{\Theta}^*\|_F \le \frac{2\sqrt{r} \big\| \Pi_{\mathcal{F}}\big(\nabla \mathcal{L}_n(\mathbf{\Theta}^*)\big) \big\|_2}{\kappa(\mathfrak{X})},$$

which completes the proof. □

E. Proof of the Results for Specific Examples

In this section, we provide the detailed proofs for corollaries of specific examples presented in Section 3.2. We will first establish the RSC condition for both examples, followed by proofs of the corollaries and more results on oracle property respecting two specific examples of matrix completion.

Particularly, the proofs include the following components: (i) establish the RSC condition, obtaining $\kappa(\mathfrak{X})$ by which Assumption 3.1 holds with high probability; (ii) estimate $\|\nabla \mathcal{L}_n(\mathbf{\Theta}^*)\|_2$ for the choice of the regularity parameter λ; (iii) establish the RSS condition, obtaining $\rho(\mathfrak{X})$ by which Assumption 3.2 holds with high probability.

E.1. Matrix Completion

As shown in (Candès & Recht, 2012) with various examples, it is insufficient to recover the low-rank matrix, since it is infeasible to recover overly "spiky" matrices which have very few large entries. Some existing work (Candès & Recht, 2012) imposes stringent matrix incoherence conditions to preclude such matrices; these assumptions are relaxed in more recent work (Negahban & Wainwright, 2012; Gunasekar et al., 2014) by restricting the spikiness ratio, which is defined as follows:

$$\alpha_{\text{sp}}(\mathbf{\Theta}) = \frac{\sqrt{m_1 m_2}\|\mathbf{\Theta}\|_\infty}{\|\mathbf{\Theta}\|_F}.$$

Assumption E.1. These exists a known α^*, such that

$$\|\mathbf{\Theta}^*\|_\infty = \frac{\alpha_{\text{sp}}(\mathbf{\Theta}^*)\|\mathbf{\Theta}^*\|_F}{\sqrt{m_1 m_2}} \le \alpha^*.$$

For the example of matrix completion, we have the following matrix concentration inequality, which follows from Proof of Corollary 1 in (Negahban & Wainwright, 2012).

Proposition E.2. Let $\mathbf{X}_i$ uniformly distributed on $\mathcal{X}$, and $\{\xi_k\}_{k=1}^n$ be a finite sequence of independent Gaussian variables with variance σ^2. There exist constants C_1, C_2 that with probability at least $1 - C_2/M$, we have

$$\Big\|\frac{1}{n}\sum_{i=1}^n \xi_i \mathbf{X}_i\Big\|_2 \le C_1\sigma\sqrt{\frac{M\log M}{m_1 m_2 n}}.$$

Furthermore, the following Lemma plays a key rule in obtaining faster rate for estimator with nonconvex penalties. Particularly, the following Lemma will provide an upper bound on $\big\|\Pi_{\mathcal{F}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2$.

Lemma E.3. If ξ_i is Gaussian noise with variance σ^2. $\mathcal{S}$ is a r-dimensional subspace. It holds with probability at least $1 - C_2/M$,

$$\Big\|\Pi_{\mathcal{S}}\Big(\frac{1}{n}\sum_{i=1}^n \xi_i \mathbf{X}_i\Big)\Big\|_2 \le C_1\sigma\sqrt{\frac{r\log M}{m_1 m_2 n}},$$

where C_1, C_2 are universal constants.

Proof. Proof is provided in Section F.4. □

In addition, we have the following Lemma (Theorem 1 in (Negahban & Wainwright, 2012)), which plays central role in establishing the RSC condition.

Lemma E.4. There are universal constants, $k_1, k_2, C_1, \ldots, C_5$, such that as long as $n > C_2 M\log M$, if the following condition is satisfied that

$$\sqrt{m_1 m_2}\frac{\|\mathbf{\Delta}\|_\infty}{\|\mathbf{\Delta}\|_F}\frac{\|\mathbf{\Delta}\|_*}{\|\mathbf{\Delta}\|_F} \le \frac{\sqrt{rn}}{k_1 r_1\sqrt{\log M} + k_2\sqrt{r_2 M\log M}}, \tag{E.1}$$

we have

$$\bigg|\frac{\big\|\mathfrak{X}_n(\mathbf{\Delta})\big\|_2}{\sqrt{n}} - \frac{\|\mathbf{\Delta}\|_F}{\sqrt{m_1 m_2}}\bigg| \le \frac{7}{8}\frac{\|\mathbf{\Delta}\|_F}{\sqrt{m_1 m_2}}\Big[1 + \frac{C_1\alpha_{\text{sp}}(\mathbf{\Delta})}{\sqrt{n}}\Big], \tag{E.2}$$

with probability greater than $1 - C_3\exp(-C_4 M\log M)$.

Proof of Corollary 3.6. With regard to the example of matrix completion, we consider a partially observed setting, *i.e.*, only the entries over the subset $\mathcal{X}$. A uniform sampling model is assumed that

$$\forall (i,j) \in \mathcal{X}, i \sim \text{uniform}([m_1]), j \sim \text{uniform}([m_2]).$$

Recall that $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$. In this proof, we consider two cases, depending on if the condition in (E.1) holds or not.

1. The condition in (E.1) does not hold.
2. The condition in (E.1) does hold.

CASE 1. If the condition in (E.1) is violated, it implies that

$$\begin{aligned}\|\widehat{\boldsymbol{\Delta}}\|_F^2 &\leq \sqrt{m_1 m_2}\|\widehat{\boldsymbol{\Delta}}\|_\infty \cdot \|\widehat{\boldsymbol{\Delta}}\|_* \frac{k_1 r_1\sqrt{\log M} + k_2\sqrt{r_2 M\log M}}{\sqrt{rn}} \\ &\leq \sqrt{m_1 m_2}(2\alpha^*)\big(\|\widehat{\boldsymbol{\Delta}}'\|_* + \|\widehat{\boldsymbol{\Delta}}''\|_*\big)\frac{k_1 r_1\sqrt{\log M} + k_2\sqrt{r_2 M\log M}}{\sqrt{rn}} \\ &\leq 12\alpha^*\sqrt{r m_1 m_2}\|\widehat{\boldsymbol{\Delta}}'\|_F\frac{k_1 r_1\sqrt{\log M} + k_2\sqrt{r_2 M\log M}}{\sqrt{rn}},\end{aligned}$$

where $\widehat{\boldsymbol{\Delta}}' = \Pi_{\mathcal{F}}(\widehat{\boldsymbol{\Delta}})$ and $\widehat{\boldsymbol{\Delta}}'' = \Pi_{\mathcal{F}^\perp}(\widehat{\boldsymbol{\Delta}})$, the second inequality follows from $\|\widehat{\boldsymbol{\Delta}}\|_\infty \leq \|\widehat{\boldsymbol{\Theta}}\|_\infty + \|\boldsymbol{\Theta}^*\|_\infty \leq 2\alpha^*$, and the decomposibility of nuclear norm that $\|\widehat{\boldsymbol{\Delta}}\|_* = \|\widehat{\boldsymbol{\Delta}}'\|_* + \|\widehat{\boldsymbol{\Delta}}''\|_*$; while the third inequality is based on the cone condition $\|\widehat{\boldsymbol{\Delta}}'\|_* \leq 5\|\widehat{\boldsymbol{\Delta}}''\|_*$ and $\|\widehat{\boldsymbol{\Delta}}'\|_* \leq \sqrt{r}\|\widehat{\boldsymbol{\Delta}}'\|_F$.
Moreover, since $\|\widehat{\boldsymbol{\Delta}}'\|_F \leq \|\widehat{\boldsymbol{\Delta}}\|_F$, we obtain that

$$\frac{1}{\sqrt{m_1 m_2}}\|\widehat{\boldsymbol{\Delta}}\|_F \leq 12\alpha^*\Big(k_1 r_1\sqrt{\frac{\log M}{n}} + k_1\sqrt{\frac{r_2 M\log M}{n}}\Big). \tag{E.3}$$

CASE 2. The condition in (E.1) is satisfied.

As implied by (E.2), we have

$$\frac{\big\|\mathfrak{X}_n(\boldsymbol{\Delta})\big\|_2}{\sqrt{n}} \geq \frac{1}{8}\frac{\|\boldsymbol{\Delta}\|_F}{\sqrt{m_1 m_2}}\Big[1 - \frac{C_1'\alpha_{\rm sp}(\boldsymbol{\Delta})}{\sqrt{n}}\Big],$$

If $C_1'\alpha_{\rm sp}(\widehat{\boldsymbol{\Delta}})/\sqrt{n} > 1/2$, we have

$$\|\widehat{\boldsymbol{\Delta}}\|_F \leq 2C_2\sqrt{m_1 m_2}\frac{\|\widehat{\boldsymbol{\Delta}}\|_\infty}{\sqrt{n}} \leq 4C_2\alpha^*\sqrt{\frac{m_1 m_2}{n}}. \tag{E.4}$$

If $C_1'\alpha_{\rm sp}(\widehat{\boldsymbol{\Delta}})/\sqrt{n} \leq 1/2$, we have

$$\frac{\big\|\mathfrak{X}_n(\widehat{\boldsymbol{\Delta}})\big\|_2^2}{n} \geq \frac{C_6^2}{m_1 m_2}\|\widehat{\boldsymbol{\Delta}}\|_F^2. \tag{E.5}$$

In order to establish the RSC condition, we need to show that (E.5) is equivalent to Assumption 3.1.

$$\begin{aligned}&\mathcal{L}_n(\boldsymbol{\Theta}^* + \widehat{\boldsymbol{\Delta}}) - \mathcal{L}_n(\boldsymbol{\Theta}^*) - \big\langle \nabla\mathcal{L}_n(\boldsymbol{\Theta}^*), \widehat{\boldsymbol{\Delta}}\big\rangle \\ &= \frac{1}{2n}\sum_{i=1}^n\big(\langle\boldsymbol{\Theta}^* + \widehat{\boldsymbol{\Delta}}, \mathbf{X}_i\rangle - y_i\big)^2 + \frac{1}{2n}\sum_{i=1}^n\big(\langle\boldsymbol{\Theta}^*, \mathbf{X}_i\rangle - y_i\big)^2 - \frac{1}{n}\sum_{i=1}^n\big(\langle\boldsymbol{\Theta}^*, \mathbf{X}_i\rangle - y_i\big)\langle\mathbf{X}_i, \widehat{\boldsymbol{\Delta}}\rangle \\ &= \frac{\big\|\mathfrak{X}_n(\widehat{\boldsymbol{\Delta}})\big\|_2^2}{n}.\end{aligned}$$

Thus, we have that (E.5) establishes the RSC condition, and $\kappa(\mathfrak{X}) = C_6^2/(m_1 m_2)$.

After establishing the RSC condition, what remains is to upper bound $n^{-1}\big\|\mathfrak{X}^*(\boldsymbol{\epsilon})\big\|_2$ and $n^{-1}\big\|\Pi_{\mathcal{F}_{S_1}}\big(\mathfrak{X}^*(\boldsymbol{\epsilon})\big)\big\|_2$. By Proposition E.2, we have that with high probability,

$$\frac{1}{n}\big\|\mathfrak{X}^*(\boldsymbol{\epsilon})\big\|_2 \leq C_6\sigma\sqrt{\frac{M\log M}{m_1 m_2 n}}; \tag{E.6}$$

By Lemma E.3, we have that with high probability,

$$\frac{1}{n}\big\|\Pi_{\mathcal{F}_{S_1}}\big(\mathfrak{X}^*(\boldsymbol{\epsilon})\big)\big\|_2 \le C_7\sigma\sqrt{\frac{r_1\log M}{m_1m_2n}}. \tag{E.7}$$

Substituting (E.6) and (E.7) into Theorem 3.4, we have that there exist positive constants C_1', C_2' such that

$$\frac{1}{\sqrt{m_1m_2}}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F \le C_1'\sigma r_1\sqrt{\frac{\log M}{n}} + C_2'\sigma\sqrt{\frac{r_2M\log M}{n}}. \tag{E.8}$$

Putting pieces (E.3), (E.4), and (E.8) together, we have

$$\frac{1}{\sqrt{m_1m_2}}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F \le \max\{\alpha^*, \sigma\}\left[C_3r_1\sqrt{\frac{\log M}{n}} + C_4\sqrt{\frac{r_2M\log M}{n}}\right],$$

which completes the proof. □

Corollary E.5. Under the conditions of Theorem 3.5, suppose $\mathbf{X}_i$ uniformly distributed on $\mathcal{X}$. These exists positive constants $C_1, \ldots, C_4$, for any $t > 0$, if $\kappa(\mathfrak{X}) = C_1/(m_1m_2) > \zeta_-$ and $\boldsymbol{\gamma}^*$ satisfies

$$\min_{i\in S}\big|(\boldsymbol{\gamma}^*)_i\big| \ge \nu + C_2\sigma\sqrt{rm_1m_2}\sqrt{\frac{M\log M}{n}},$$

where $S = \mathrm{supp}(\boldsymbol{\sigma}^*)$, for estimator in (2.2) with regularization parameter

$$\lambda \ge C_3(1+\sqrt{r})\sigma\sqrt{\frac{M\log M}{nm_1m_2}},$$

we have that with high probability, $\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Theta}}_O$, which yields that $\mathrm{rank}(\widehat{\boldsymbol{\Theta}}) = \mathrm{rank}(\widehat{\boldsymbol{\Theta}}_O) = \mathrm{rank}(\boldsymbol{\Theta}^*) = r$. In addition, we have

$$\frac{1}{\sqrt{m_1m_2}}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F \le C_4r\sigma\sqrt{\frac{\log M}{n}}. \tag{E.9}$$

Proof of Corollary E.5. As shown in the proof of Corollary 3.6, we have $\kappa(\mathfrak{X}) = C_1/(m_1m_2)$, together with (E.6) and (E.7), in order to prove Corollary E.5, according to Theorem 3.5, what remains is to obtain $\rho(\mathfrak{X})$ in Assumption 3.2. It can be shown that Assumption 3.2 is equivalent as

$$\frac{\rho(\mathfrak{X})}{2}\|\widehat{\boldsymbol{\Delta}}\|_F^2 \ge \frac{1}{n}\|\mathfrak{X}(\widehat{\boldsymbol{\Delta}})\|_2^2.$$

We consider the following cases depending on if (E.1) holds or not.

CASE 1. If the condition in (E.1) is violated,

$$\frac{1}{n}\|\mathfrak{X}(\widehat{\boldsymbol{\Delta}})\|_F^2 \le \|\widehat{\boldsymbol{\Delta}}\|_\infty^2 \le \|\widehat{\boldsymbol{\Delta}}\|_F^2,$$

which implies that $\rho(\mathfrak{X}) = 1$.

CASE 2. The condition in (E.1) is satisfied. As implied by Lemma E.4, when $n \ge C_5^2\alpha^* \ge C_5^2\alpha_{\mathrm{sp}}(\widehat{\boldsymbol{\Delta}})$, we have that with high probability, the following holds:

$$\frac{C_6}{m_1m_2}\|\widehat{\boldsymbol{\Delta}}\|_F^2 \ge \frac{1}{n}\|\mathfrak{X}(\widehat{\boldsymbol{\Delta}})\|_2^2.$$

Thus, $\rho(\mathfrak{X}) = C_6/(m_1m_2)$, which completes the proof. □

E.2. Matrix Sensing With Dependent Sampling

In this subsection, we provide the proof for the results on matrix sensing. In particular, we will first establish the RSC condition for the application of matrix sensing, followed by the proof on faster convergence rate and more results on the oracle property.

In order to establish the RSC condition, we need the following lemma (Proposition 1 in (Negahban & Wainwright, 2011)).

Lemma E.6. Consider the sampling operator of $\boldsymbol{\Sigma}$-ensemble, it holds with probability at least $1-2\exp(-n/32)$ that

$$\frac{\|\mathfrak{X}(\boldsymbol{\Delta})\|_2}{\sqrt{n}} \geq \frac{1}{4}\big\|\sqrt{\boldsymbol{\Sigma}}\mathrm{vec}(\boldsymbol{\Delta})\big\|_2 - 12\pi(\boldsymbol{\Sigma})\bigg(\sqrt{\frac{m_1}{n}}+\sqrt{\frac{m_2}{n}}\bigg)\|\boldsymbol{\Delta}\|_*.$$

In addition, we need the upper bound of $n^{-1}\big\|\mathfrak{X}^*(\boldsymbol{\epsilon})\big\|_2$, as stated in the following Proposition (Lemma 6, (Negahban & Wainwright, 2011)).

Proposition E.7. With high probability, there are universal constants C_1, C_2 and C_3 such that

$$\mathbb{P}\bigg[\frac{\big\|\mathfrak{X}^*(\boldsymbol{\epsilon})\big\|_2}{n} \geq C_1\sigma\pi(\boldsymbol{\Sigma})\bigg(\sqrt{\frac{m_1}{n}}+\sqrt{\frac{m_2}{n}}\bigg)\bigg] \leq C_2\exp\big(-C_3(m_1+m_2)\big),$$

where $\pi(\boldsymbol{\Sigma})^2 = \sup_{\|\mathbf{u}\|_2=1,\|\mathbf{v}\|_2=1}\mathrm{Var}(\mathbf{u}^\top\mathbf{X}\mathbf{v})$.

Proof of Corollary 3.8. To begin with, we need to establish the RSC condition as in Assumption 3.1. According to Lemma E.6, we have that

$$\frac{\|\mathfrak{X}(\widehat{\boldsymbol{\Delta}})\|_2}{\sqrt{n}} \geq \frac{\sqrt{\lambda_{\min}(\boldsymbol{\Sigma})}}{4}\|\widehat{\boldsymbol{\Delta}}\|_F - 12\pi(\boldsymbol{\Sigma})\bigg(\sqrt{\frac{m_1}{n}}+\sqrt{\frac{m_2}{n}}\bigg)\|\widehat{\boldsymbol{\Delta}}\|_*.$$

By the decomposibility of nuclear norm, we have that

$$\|\widehat{\boldsymbol{\Delta}}\|_* = \|\widehat{\boldsymbol{\Delta}}'\|_* + \|\widehat{\boldsymbol{\Delta}}''\|_* \leq 6\|\widehat{\boldsymbol{\Delta}}'\|_* = 6\sqrt{r}\|\widehat{\boldsymbol{\Delta}}'\|_F \leq 6\sqrt{r}\|\widehat{\boldsymbol{\Delta}}\|_F, \tag{E.10}$$

where $\widehat{\boldsymbol{\Delta}}' = \Pi_{\mathcal{F}}(\widehat{\boldsymbol{\Delta}})$ and $\widehat{\boldsymbol{\Delta}}'' = \Pi_{\mathcal{F}^\perp}(\widehat{\boldsymbol{\Delta}})$.
By substituting (E.10) into Proposition E.6, we have that

$$\begin{aligned}\frac{\|\mathfrak{X}(\widehat{\boldsymbol{\Delta}})\|_2}{\sqrt{n}} &\geq \frac{\sqrt{\lambda_{\min}(\boldsymbol{\Sigma})}}{4}\|\widehat{\boldsymbol{\Delta}}\|_F - 72\sqrt{r}\pi(\boldsymbol{\Sigma})\bigg(\sqrt{\frac{m_1}{n}}+\sqrt{\frac{m_2}{n}}\bigg)\|\widehat{\boldsymbol{\Delta}}\|_F\\ &= \bigg[\frac{\sqrt{\lambda_{\min}(\boldsymbol{\Sigma})}}{4} - 72\sqrt{r}\pi(\boldsymbol{\Sigma})\bigg(\sqrt{\frac{m_1}{n}}+\sqrt{\frac{m_2}{n}}\bigg)\bigg]\|\widehat{\boldsymbol{\Delta}}\|_F.\end{aligned}$$

Thus, for $n > C_1 r\pi^2(\boldsymbol{\Sigma})m_1m_2/\lambda_{\min}(\boldsymbol{\Sigma})$ where C_1 is sufficiently large such that

$$72\sqrt{r}\pi(\boldsymbol{\Sigma})\bigg(\sqrt{\frac{m_1}{n}}+\sqrt{\frac{m_2}{n}}\bigg) \leq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{8},$$

we have

$$\frac{\|\mathfrak{X}(\widehat{\boldsymbol{\Delta}})\|_2}{\sqrt{n}} \geq \frac{\sqrt{\lambda_{\min}(\boldsymbol{\Sigma})}}{8}\|\widehat{\boldsymbol{\Delta}}\|_F,$$

which implies that

$$\frac{\|\mathfrak{X}(\widehat{\boldsymbol{\Delta}})\|_2^2}{n} \geq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{64}\|\widehat{\boldsymbol{\Delta}}\|_F^2.$$

Therefore, $\kappa(\mathfrak{X}) = \lambda_{\min}(\boldsymbol{\Sigma})/32$ such that the following holds,

$$\frac{\|\mathfrak{X}(\widehat{\boldsymbol{\Delta}})\|_2^2}{n} \geq \frac{\kappa(\mathfrak{X})}{2}\|\widehat{\boldsymbol{\Delta}}\|_F^2,$$

which establishes the RSC condition for matrix sensing.

On the other hand, we have

$$\big\|\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2 = \big\|\mathbf{U}^*_{S_1}\mathbf{U}^{*\top}_{S_1}\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\mathbf{V}^*_{S_1}\mathbf{V}^{*\top}_{S_1}\big\|_2 = \big\|\mathbf{U}^{*\top}_{S_1}\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\mathbf{V}^*_{S_1}\big\|_2,$$

where the second inequality follows from the property of left and right singular vectors $\mathbf{U}^*_{S_1}, \mathbf{V}^*_{S_1}$.

It is worth noting that $\mathbf{U}^{*\top}_{S_1}\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\mathbf{V}^*_{S_1} \in \mathbb{R}^{r_1\times r_1}$. By Proposition E.7, we have that

$$\begin{aligned}\big\|\mathbf{U}^{*\top}\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\mathbf{V}^*\big\|_2 &\le 2C_0\sigma\pi(\mathbf{\Sigma})\sqrt{\frac{M}{n}},\\ \big\|\mathbf{U}^{*\top}_{S_1}\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\mathbf{V}^*_{S_1}\big\|_2 &\le 2C_0\sigma\pi(\mathbf{\Sigma})\sqrt{\frac{r_1}{n}},\end{aligned} \tag{E.11}$$

which hold with probability at lease $1 - C_1\exp(-C_2 r_1)$.

The upper bound is obtained directed from Theorem 3.4 and (E.11). Thus, we complete the proof. □

Corollary E.8. Under the condition of Theorem 3.5, for some universal constants $C_1,\ldots,C_6$ if $\kappa(\mathfrak{X}) = C_1\lambda_{\min}(\mathbf{\Sigma}) > \zeta_-$ and $\boldsymbol{\gamma}^*$ satisfies

$$\min_{i\in S}\big|(\boldsymbol{\gamma}^*)_i\big| \ge \nu + C_2\sigma\pi(\mathbf{\Sigma})\frac{\sqrt{r}(\sqrt{m_1}+\sqrt{m_2})}{\sqrt{n}\lambda_{\min}(\mathbf{\Sigma})},$$

where $S = \mathrm{supp}(\boldsymbol{\gamma}^*)$, for estimator in (2.2) with regularization parameter

$$\lambda \ge C_3\big(1 + \frac{\sqrt{r}\lambda_{\max}(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})}\big)\sigma\pi(\mathbf{\Sigma})\bigg(\sqrt{\frac{m_1}{n}} + \sqrt{\frac{m_2}{n}}\bigg),$$

we have that $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}}_O$, which yields that $\mathrm{rank}(\widehat{\mathbf{\Theta}}) = \mathrm{rank}(\widehat{\mathbf{\Theta}}_O) = \mathrm{rank}(\mathbf{\Theta}^*) = r$, with probability at least $1 - C_4\exp\big(-C_5(m_1+m_2)\big)$. In addition, we have

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F \le \frac{C_6 r\pi(\mathbf{\Sigma})}{\sqrt{n}\lambda_{\min}(\mathbf{\Sigma})}. \tag{E.12}$$

Proof of Corollary E.8. The proof follows from the proof of Corollary 3.8 and Theorem 3.5. As shown in the proof of Corollary 3.8, we have $\kappa(\mathfrak{X}) = C_1\lambda_{\min}(\mathbf{\Sigma})$, together with (E.11), in order to prove Corollary E.8, according to Theorem 3.5, what remains is to obtain $\rho(\mathfrak{X})$ in Assumption 3.2, respecting the example of matrix sensing.

According to Assumption 3.2, we have that $\rho(\mathfrak{X}) = \lambda_{\max}(\mathbf{H}_n)$, where $\mathbf{H}_n$ is the Hessian matrix of $\mathcal{L}_n(\cdot)$. Based on the definition of $\mathcal{L}_n(\cdot)$, we have

$$\mathbf{H}_n = n^{-1}\sum_{i=1}^{n}\mathrm{vec}(\mathbf{X}_i)\mathrm{vec}(\mathbf{X}_i)^\top.$$

Thus $\mathbb{E}[\mathbf{H}_n] = \mathbf{\Sigma}$. By concentration, we have that when n is sufficiently large, with high probability, $\lambda_{\max}(\mathbf{H}_n) \le 2\lambda_{\max}(\mathbf{\Sigma})$, which is equivalent to $\rho(\mathfrak{X}) \le 2\lambda_{\max}(\mathbf{\Sigma})$, holding with high probability, where n is sufficiently large. This completes the proof. □

F. Proof of Auxiliary Lemmas

F.1. Proof of Lemma D.1

Proof. By the restricted strong convexity assumption (Assumption 3.1), we have

$$\mathcal{L}_n(\mathbf{\Theta}_2) \ge \mathcal{L}_n(\mathbf{\Theta}_1) + \langle\nabla\mathcal{L}_n(\mathbf{\Theta}_1), \mathbf{\Theta}_2 - \mathbf{\Theta}_1\rangle + \frac{\kappa(\mathfrak{X})}{2}\|\mathbf{\Theta}_2 - \mathbf{\Theta}_1\|_F^2. \tag{F.1}$$

In the following, we will show the strong smoothness of $\mathcal{Q}_\lambda(\cdot)$, based on the regularity condition (ii), which imposes constraint on the level of nonconvexity of $q_\lambda(\cdot)$. Assume $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}(\mathbf{\Theta}_1), \boldsymbol{\gamma}_2 = \boldsymbol{\gamma}(\mathbf{\Theta}_2)$ are the vectors of singular values of $\mathbf{\Theta}_1, \mathbf{\Theta}_2$, respectively, and the singular values in γ_1, γ_2 are nonincreasing. For $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$, we have the following singular value decompositions:

$$\mathbf{\Theta}_1 = \mathbf{U}_1\mathbf{\Gamma}_1\mathbf{V}_1^\top,$$
$$\mathbf{\Theta}_2 = \mathbf{U}_2\mathbf{\Gamma}_2\mathbf{V}_2^\top,$$

where $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2 \in \mathbb{R}^{m\times m}$ are diagonal matrix with $\mathbf{\Gamma}_1 = \mathrm{diag}(\gamma_1), \mathbf{\Gamma}_2 = \mathrm{diag}(\gamma_2)$. For each pair of singular values of $\mathbf{\Theta}_1, \mathbf{\Theta}_2$: $\big((\gamma_1)_i, (\gamma_2)_i\big)$ where $i = 1, 2, \ldots, m$, we have

$$-\zeta_-\big((\gamma_1)_i - (\gamma_2)_i\big)^2 \le \big[q_\lambda'\big((\gamma_1)_i\big) - q_\lambda'\big((\gamma_2)_i\big)\big]\big((\gamma_1)_i - (\gamma_2)_i\big),$$

which is equivalent to

$$\big\langle\big(-q_\lambda'(\mathbf{\Gamma}_1)\big) - \big(-q_\lambda'(\mathbf{\Gamma}_2)\big), \mathbf{\Gamma}_1 - \mathbf{\Gamma}_2\big\rangle \le \zeta_-\|\mathbf{\Gamma}_1 - \mathbf{\Gamma}_2\|_F^2,$$

which yields

$$\big\langle\big(-\nabla\mathcal{Q}_\lambda(\mathbf{\Theta}_1)\big) - \big(-\nabla\mathcal{Q}_\lambda(\mathbf{\Theta}_2)\big), \mathbf{\Theta}_1 - \mathbf{\Theta}_2\big\rangle \le \zeta_-\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_F^2. \tag{F.2}$$

Since (F.2) is the definition of strongly smoothness of $-\mathcal{Q}(\cdot)$, it can be show to be equivalent to the following inequality that

$$\mathcal{Q}_\lambda(\mathbf{\Theta}_2) \ge \mathcal{Q}_\lambda(\mathbf{\Theta}_1) + \langle\nabla\mathcal{Q}(\mathbf{\Theta}_1), \mathbf{\Theta}_2 - \mathbf{\Theta}_1\rangle - \frac{\zeta_-}{2}\|\mathbf{\Theta}_2 - \mathbf{\Theta}_1\|_F^2. \tag{F.3}$$

Adding up (F.1) and (F.3), we complete the proof. □

F.2. Proof of Lemma D.2

Proof. By Lemma D.1, we have that

$$\widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}) + \lambda\|\widehat{\mathbf{\Theta}}\|_* - \widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*) - \lambda\|\mathbf{\Theta}^*\|_* \ge \langle\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\rangle + \lambda\|\widehat{\mathbf{\Theta}}\|_* - \lambda\|\mathbf{\Theta}^*\|_*. \tag{F.4}$$

For the first term on the RHS in (F.4), we have the following lower bound

$$\begin{aligned}\langle\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\rangle &= \big\langle\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*), \Pi_{\mathcal{F}}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\big\rangle + \big\langle\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*), \Pi_{\mathcal{F}^\perp}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\big\rangle \\ &\ge -\underbrace{\big\|\Pi_{\mathcal{F}}\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*)\big)\big\|_2}_{I_1}\big\|\Pi_{\mathcal{F}}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\big\|_* \\ &\quad - \underbrace{\big\|\Pi_{\mathcal{F}^\perp}\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*)\big)\big\|_2}_{I_2}\big\|\Pi_{\mathcal{F}^\perp}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\big\|_*,\end{aligned} \tag{F.5}$$

where the last inequality follows from Hölder's inequality.

Analysis of term I_1. It can be shown that $\nabla\mathcal{L}_n(\mathbf{\Theta}^*) = -\mathfrak{X}^*(\boldsymbol{\epsilon})/n$. Based on the condition that $\lambda > 2n^{-1}\|\mathfrak{X}^*(\boldsymbol{\epsilon})\|_2$, we have that

$$\|\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\|_2 \le \lambda/2. \tag{F.6}$$

Moreover, by condition (iv) in Assumption 3.3 and (F.6), we obtain that

$$\big\|\Pi_{\mathcal{F}}\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*)\big)\big\|_2 = \big\|\Pi_{\mathcal{F}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*) + \mathcal{Q}_\lambda(\mathbf{\Theta}^*)\big)\big\|_2 \le 3\lambda/2.$$

Analysis of term I_2. Since $\Pi_{\mathcal{F}^\perp}(\mathbf{\Theta}^*) = \mathbf{0}$, we have that

$$\big\|\Pi_{\mathcal{F}^\perp}\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*)\big)\big\|_2 = \big\|\Pi_{\mathcal{F}^\perp}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2 \le \lambda/2. \tag{F.7}$$

Putting pieces (F.6) and (F.7) into (F.5), we obtain

$$\langle\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\rangle \ge -3\lambda/2\big\|\Pi_{\mathcal{F}}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\big\|_* - \lambda/2\big\|\Pi_{\mathcal{F}^\perp}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\big\|_*. \tag{F.8}$$

Meanwhile, we have the lower bound on $\lambda\|\widehat{\mathbf{\Theta}}\|_* - \lambda\|\mathbf{\Theta}\|_*$ that

$$\begin{aligned}\lambda\|\widehat{\mathbf{\Theta}}\|_* - \lambda\|\mathbf{\Theta}\|_* &= \lambda\big\|\Pi_{\mathcal{F}}(\widehat{\mathbf{\Theta}})\big\|_* + \lambda\big\|\Pi_{\mathcal{F}^\perp}(\widehat{\mathbf{\Theta}})\big\|_* - \lambda\|\mathbf{\Theta}\|_* \\ &\geq -\lambda\big\|\Pi_{\mathcal{F}}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\big\|_* + \lambda\big\|\Pi_{\mathcal{F}^\perp}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\big\|_* \end{aligned} \tag{F.9}$$

Adding (F.8) and (F.9) yields that

$$\langle \nabla \widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}^*), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\rangle + \lambda\|\widehat{\mathbf{\Theta}}\|_* - \lambda\|\mathbf{\Theta}\|_* = -5\lambda/2\big\|\Pi_{\mathcal{F}}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\big\|_* + \lambda/2\big\|\Pi_{\mathcal{F}^\perp}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\big\|_*. \tag{F.10}$$

Due to the fact that $\widehat{\mathbf{\Theta}}$ is the global minimizer of (2.2), provided the condition that $\kappa(\mathfrak{X}) > \zeta_-$, we have

$$\widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}) + \lambda\|\widehat{\mathbf{\Theta}}\|_* - \widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}) - \lambda\|\mathbf{\Theta}^*\|_* \leq 0. \tag{F.11}$$

Substituting (F.10) and (F.11) into (F.4), since $\lambda > 0$, we have that

$$\big\|\Pi_{\mathcal{F}^\perp}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\big\|_* \leq 5\big\|\Pi_{\mathcal{F}}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\big\|_*,$$

which completes the proof. □

F.3. Proof of Lemma D.3

Proof. $\widehat{\mathbf{\Delta}}_O = \widehat{\mathbf{\Theta}}_O - \mathbf{\Theta}^*$. According to observation model (2.1) and definition of $\mathfrak{X}(\cdot)$, we have

$$\begin{aligned}\mathcal{L}_n(\widehat{\mathbf{\Theta}}_O) - \mathcal{L}_n(\mathbf{\Theta}^*) &= \frac{1}{2n}\sum_{i=1}^n \big(y_i - \mathfrak{X}_i(\mathbf{\Theta}^* + \widehat{\mathbf{\Delta}}_O)\big)^2 - \frac{1}{2n}\sum_{i=1}^n \big(y_i - \mathfrak{X}_i(\mathbf{\Theta}^*)\big)^2 \\ &= \frac{1}{2n}\sum_{i=1}^n \big(\epsilon_i - \mathfrak{X}_i(\widehat{\mathbf{\Delta}}_O)\big)^2 - \frac{1}{2n}\sum_{i=1}^n \epsilon_i^2 \\ &= \frac{1}{2n}\|\mathfrak{X}(\widehat{\mathbf{\Delta}}_O)\|_2^2 - \frac{1}{n}\langle \mathfrak{X}^*(\boldsymbol{\epsilon}), \widehat{\mathbf{\Delta}}_O\rangle,\end{aligned}$$

where $\mathfrak{X}^*(\boldsymbol{\epsilon}) = \sum_{i=1}^n \epsilon_i \mathbf{X}_i$ is the adjoint of the operator $\mathfrak{X}$. Because the oracle estimator $\widehat{\mathbf{\Theta}}_O$ minimizes $\mathcal{L}_n(\cdot)$ over the subspace $\mathcal{F}$, while $\mathbf{\Theta}^* \in \mathcal{F}$, we have $\mathcal{L}_n(\widehat{\mathbf{\Theta}}_O) - \mathcal{L}_n(\mathbf{\Theta}^*) \leq 0$, which yields

$$\frac{1}{2n}\|\mathfrak{X}(\widehat{\mathbf{\Delta}}_O)\|_2^2 \leq \frac{1}{n}\langle \mathfrak{X}^*(\boldsymbol{\epsilon}), \widehat{\mathbf{\Delta}}_O\rangle. \tag{F.12}$$

On the other hand, recall that by the RSC condition (Assumption 3.1), we have

$$\mathcal{L}_n(\mathbf{\Theta} + \mathbf{\Delta}) \geq \mathcal{L}_n(\mathbf{\Theta}) + \langle \nabla \mathcal{L}_n(\mathbf{\Theta}), \mathbf{\Delta}\rangle + \kappa(\mathfrak{X})/2\|\mathbf{\Delta}\|_F^2,$$

which implies that

$$\frac{1}{2n}\|\mathfrak{X}(\widehat{\mathbf{\Delta}}_O)\|_2^2 - \frac{1}{n}\langle \mathfrak{X}^*(\boldsymbol{\epsilon}), \widehat{\mathbf{\Delta}}_O\rangle - \langle \nabla \mathcal{L}_n(\mathbf{\Theta}^*), \mathbf{\Delta}\rangle = \frac{1}{2n}\|\mathfrak{X}(\widehat{\mathbf{\Delta}}_O)\|_2^2 \geq \frac{\kappa(\mathfrak{X})}{2}\|\widehat{\mathbf{\Delta}}_O\|_F^2. \tag{F.13}$$

Substituting (F.13) into (F.12), we have

$$\frac{\kappa(\mathfrak{X})}{2}\|\widehat{\mathbf{\Delta}}_O\|_F^2 \leq \frac{1}{2n}\|\mathfrak{X}(\widehat{\mathbf{\Delta}}_O)\|_2^2 \leq \frac{1}{n}\langle \mathfrak{X}^*(\boldsymbol{\epsilon}), \widehat{\mathbf{\Delta}}_O\rangle. \tag{F.14}$$

Therefore,

$$\|\widehat{\mathbf{\Delta}}_O\|_F^2 \leq \frac{2\big\langle \Pi_{\mathcal{F}}\big(\mathfrak{X}^*(\boldsymbol{\epsilon})\big), \widehat{\mathbf{\Delta}}_O\big\rangle}{n\kappa(\mathfrak{X})} \leq \frac{2\big\|\Pi_{\mathcal{F}}\big(\mathfrak{X}^*(\boldsymbol{\epsilon})\big)\big\|_2 \|\widehat{\mathbf{\Delta}}_O\|_*}{n\kappa(\mathfrak{X})},$$

where the last inequality is due to Hölder inequality. Moreover, since the rank $\mathbf{\Delta}_O$ is r, we have the fact that $\|\widehat{\mathbf{\Delta}}_O\|_* \leq \sqrt{r}\|\widehat{\mathbf{\Delta}}_O\|_F$, which indicates that

$$\|\widehat{\mathbf{\Delta}}_O\|_F^2 \leq \frac{2\sqrt{r}\big\|\Pi_{\mathcal{F}}\big(\mathfrak{X}^*(\boldsymbol{\epsilon})\big)\big\|_2 \cdot \|\widehat{\mathbf{\Delta}}_O\|_F}{n\kappa(\mathfrak{X})}.$$

Therefore, we have the following deterministic error bound

$$\|\widehat{\boldsymbol{\Delta}}_O\|_F \le \frac{2\sqrt{r}\big\|\Pi_{\mathcal{F}}\big(\mathfrak{X}^*(\boldsymbol{\epsilon})\big)\big\|_2}{n\kappa(\mathfrak{X})} = \frac{2\sqrt{r}\big\|\Pi_{\mathcal{F}}\big(\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\big)\big\|_2}{\kappa(\mathfrak{X})},$$

where the last equality results from the fact that $\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*) = -\mathfrak{X}^*(\boldsymbol{\epsilon})/n$.

Thus, we complete the proof. □

F.4. Proof of Lemma E.3

In order to prove Lemma E.3, we need the Ahlswede-Winter Matrix Bound. To begin with, we introduce the definition of $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$, followed by some established results on $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$.

The sub-Gaussian norm of X, denoted by $\|X\|_{\psi_2}$, is defined as follows

$$\|X\|_{\psi_2} = \sup_{p\ge 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}.$$

It is known that if $\mathbb{E}[X]=0$, then $\mathbb{E}[\exp(tX)] \le \exp(Ct^2\|X\|_{\psi_2}^2)$ for all $t\in\mathbb{R}$.

The sub-Exponential norm of X, denoted by $\|X\|_{\psi_1}$, is defined as follows

$$\|X\|_{\psi_1} = \sup_{p\ge 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}.$$

By (Vershynin, 2010), we have the following Lemma.

Lemma F.1. For Z_1 and Z_2 being two sub-Gaussian random variables, Z_1Z_2 is a sub-exponential random variable with

$$\|Z_1Z_2\|_{\psi_1} \le C\max\big\{\|Z_1\|_{\psi_2}^2,\ \|Z_2\|_{\psi_2}^2\big\},$$

where $C>0$ is an absolute constant.

Theorem F.2 (Ahlswede-Winter Matrix Bound)**.** (Negahban & Wainwright, 2012) Let $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$ be random matrices of size $m_1\times m_2$. Let $\|\mathbf{Z}_i\|_{\psi_1}\le K$ for all i such that $\|\mathbf{Z}_i\|_{\psi_1}$ is upper bounded by K. Furthermore, we have $\delta_i^2 = \max\big\{\big\|\mathbb{E}[\mathbf{Z}_i^\top\mathbf{Z}_i]\big\|_2, \big\|\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i^\top]\big\|_2\big\}$, and $\delta^2=\sum_{i=1}^n\delta_i^2$. Then we have

$$\mathbb{P}\bigg(\bigg\|\sum_{i=1}^n\mathbf{Z}_i\bigg\|_2\ge t\bigg) \le m_1m_2\max\Big\{\exp\Big(-\frac{t^2}{4\delta^2}\Big), \exp\Big(-\frac{t}{2K}\Big)\Big\}.$$

Now we are ready to prove Lemma E.3.

Proof of Lemma E.3. Since $\mathbf{U}^*$ and $\mathbf{V}^*$ are singular vectors, for $\mathcal{S}=\mathcal{F}(\mathbf{U}^*,\mathbf{V}^*)$, we have

$$\begin{aligned}\frac{1}{n}\bigg\|\Pi_{\mathcal{S}}\Big(\sum_{i=1}^n\xi_i\mathbf{X}_i\Big)\bigg\|_2 &= \frac{1}{n}\bigg\|\mathbf{U}^*\mathbf{U}^{*\top}\Big(\sum_{i=1}^n\xi_i\mathbf{X}_i\Big)\mathbf{V}^*\mathbf{V}^{*\top}\bigg\|_2\\ &= \frac{1}{n}\bigg\|\mathbf{U}^{*\top}\Big(\sum_{i=1}^n\xi_i\mathbf{X}_i\Big)\mathbf{V}^*\bigg\|_2.\end{aligned}$$

Recall that $\mathbf{X}_i=\mathbf{e}_{j(i)}\mathbf{e}_{k(i)}^\top$. Let $\mathbf{Y}_i=\epsilon_i\mathbf{X}_i=\epsilon_i\mathbf{e}_{j(i)}\mathbf{e}_{k(i)}^\top$. We have $\|\mathbf{Y}_i\|_{\psi_1}\le C\sigma^2$. Let $\mathbf{Z}_i=\mathbf{U}^{*\top}\mathbf{Y}_i\mathbf{V}^*\in\mathbb{R}^{r\times r}$. We have

$$\|\mathbf{Z}_i\|_{\psi_1} = \big\|\mathbf{U}^{*\top}\mathbf{Y}_i\mathbf{V}^*\big\|_{\psi_1}.$$

Based on the definition of $\mathbf{Y}_i$, we have that $\|\mathbf{Z}_i\|_{\psi_1}<C\sigma$. By applying Theorem F.1, we have

$$\|\mathbf{Z}_i\|_{\psi_1}\le C'\sigma^2.$$

Thus, $K = C'\sigma^2$.

Furthermore, we have

$$\begin{aligned}\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i^\top] &= \mathbb{E}[\mathbf{U}^{*\top}\mathbf{Y}_i\mathbf{V}^*\mathbf{V}^{*\top}\mathbf{Y}_i^\top\mathbf{U}^*] = \mathbb{E}[\epsilon_i^2\mathbf{U}^{*\top}\mathbf{e}_{j(i)}\mathbf{e}_{k(i)}^\top\mathbf{V}^*\mathbf{V}^{*\top}\mathbf{e}_{k(i)}\mathbf{e}_{j(i)}^\top\mathbf{U}^*]\\ &= \sigma^2\mathbb{E}[\mathbf{U}^{*\top}\mathbf{e}_{j(i)}\mathbf{e}_{k(i)}^\top\mathbf{V}^*\mathbf{V}^{*\top}\mathbf{e}_{k(i)}\mathbf{e}_{j(i)}^\top\mathbf{U}^*]\end{aligned}$$

Based on the definition of spectral norm, we have

$$\begin{aligned}\big\|\mathbf{U}^{*\top}\mathbf{e}_{j(i)}\mathbf{e}_{k(i)}^\top\mathbf{V}^*\mathbf{V}^{*\top}\mathbf{e}_{k(i)}\mathbf{e}_{j(i)}^\top\mathbf{U}^*\big\|_2 &= \max_{\|\mathbf{a}\|_2=1}\mathbf{a}^\top\mathbf{U}^{*\top}\mathbf{e}_{j(i)}\mathbf{e}_{k(i)}^\top\mathbf{V}^*\mathbf{V}^{*\top}\mathbf{e}_{k(i)}\mathbf{e}_{j(i)}^\top\mathbf{U}^*\mathbf{a}\\ &= \max_{\|\mathbf{b}\|_2=1}\mathbf{b}^\top\mathbf{e}_{j(i)}\mathbf{e}_{k(i)}^\top\mathbf{V}^*\mathbf{V}^{*\top}\mathbf{e}_{k(i)}\mathbf{e}_{j(i)}^\top\mathbf{b},\end{aligned}$$

where the second equality follows by setting $\mathbf{b} = \mathbf{U}^*\mathbf{a} \in \mathbb{R}^{m_1}$. In addition, we have

$$\mathbf{b}^\top\mathbf{e}_{j(i)}\mathbf{e}_{k(i)}^\top\mathbf{V}^*\mathbf{V}^{*\top}\mathbf{e}_{k(i)}\mathbf{e}_{j(i)}^\top\mathbf{b} = \mathbf{b}_{j(i)}\mathbf{v}_k^*\mathbf{v}_k^{*\top}\mathbf{b}_{j(i)} = \mathbf{b}_{j(i)}^2\|\mathbf{v}_k^*\|_2^2,$$

where $\mathbf{v}_k^*$ is the k-th row of $\mathbf{V}^*$. Thus

$$\begin{aligned}\big\|\mathbb{E}[\mathbf{U}^{*\top}\mathbf{e}_{j(i)}\mathbf{e}_{k(i)}^\top\mathbf{V}^*\mathbf{V}^{*\top}\mathbf{e}_{k(i)}\mathbf{e}_{j(i)}^\top\mathbf{U}^*]\big\|_2 &= \bigg\|\frac{1}{m_1m_2}\sum_{j=1}^{m_1}\sum_{k=2}^{m_2}\mathbf{U}^{*\top}\mathbf{e}_j\mathbf{e}_k^\top\mathbf{V}^*\mathbf{V}^{*\top}\mathbf{e}_k\mathbf{e}_j^\top\mathbf{U}^*\bigg\|_2\\ &= \frac{1}{m_1m_2}\max_{\|\mathbf{a}\|_2=1}\mathbf{a}^\top\sum_{j=1}^{m_1}\sum_{k=2}^{m_2}\mathbf{U}^{*\top}\mathbf{e}_j\mathbf{e}_k^\top\mathbf{V}^*\mathbf{V}^{*\top}\mathbf{e}_k\mathbf{e}_j^\top\mathbf{U}^*\mathbf{a}\\ &= \frac{1}{m_1m_2}\max_{\|\mathbf{b}\|_2=1}\sum_{j=1}^{m_1}\sum_{k=2}^{m_2}b_j^2\|\mathbf{v}_k^*\|_2^2.\end{aligned}$$

Since $\sum_{j=1}^{m_1}b_j^2 = 1$ and $\sum_{k=1}^{m_2}\|\mathbf{v}_k^*\|_2^2 = \|\mathbf{V}^*\|_F^2 = r$, we obtain that

$$\big\|\mathbb{E}[\mathbf{U}^{*\top}\mathbf{e}_{j(i)}\mathbf{e}_{k(i)}^\top\mathbf{V}^*\mathbf{V}^{*\top}\mathbf{e}_{k(i)}\mathbf{e}_{j(i)}^\top\mathbf{U}^*]\big\|_2 = \frac{r}{m_1m_2}.$$

Therefore, we have

$$\big\|\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i^\top]\big\|_2 = \frac{\sigma^2 r}{m_1m_2},$$

and the same result also applies to $\big\|\mathbb{E}[\mathbf{Z}_i^\top\mathbf{Z}_i]\big\|_2$.
By applying Theorem F.2, we obtain that

$$\mathbb{P}\bigg(\bigg\|\sum_{i=1}^n\xi_i\mathbf{Z}_i\bigg\|_2 \geq t\bigg) \leq m_1m_2\max\bigg\{\exp\Big(-\frac{m_1m_2t^2}{4n\sigma^2 r}\Big), \exp(-\frac{t}{2\sigma^2})\bigg\}.$$

Thus, with probability at least $1 - C_2M^{-1}$, we have

$$\bigg\|\sum_{i=1}^n\xi_i\mathbf{Z}_i\bigg\|_2 \leq C_1\sigma\sqrt{\frac{nr\log M}{m_1m_2}}$$

where $M = \max(m_1, m_2)$. It immediately implies that

$$\bigg\|\frac{1}{n}\sum_{i=1}^n\xi_i\mathbf{Z}_i\bigg\|_2 \leq C_1\sigma\sqrt{\frac{r\log M}{m_1m_2n}}, \tag{F.15}$$

which completes the proof. □