Supplementary Material to "Simultaneously sparse and low-rank matrix estimation via l_1-norm and nonconvex regularization"

Ling Peng, Shanshan Tang, Fei Huang, Weiying Ping

In this supplementary material, we provide the proofs for the main results discussed in Section A, along with some auxiliary lemmas. Computational algorithms are included in Section B. Throughout this appendix, we use C to represent a generic positive constant, the value of which may vary, even within the same line.

A Proof of main results

A.1 Proof of Theorem 1

Denote $\tilde{\Delta} := \tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*$, where $\tilde{\boldsymbol{\Theta}}$ is a stationary point in the program (1). Let $\tilde{\Delta} := \tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*$, we first prove that $\tilde{\Delta}$ lies in the cone $\mathcal{C}$ defined as

$$\mathcal{C} := \left\{ \Delta \in \mathbb{R}^{d_1 \times d_2} : \|S^{\perp}_{\boldsymbol{\Theta}^*} \circ \Delta\|_1 \leq 3\|S_{\boldsymbol{\Theta}^*} \circ \Delta\|_1 + 5\|\mathcal{P}_{\boldsymbol{\Theta}^*}(\Delta)\|_* + \|\Pi_{U_r^{\perp}} \Delta \Pi_{V_r^{\perp}}\|_* \right\}$$

Lemma 1 *Under Assumption A, for the choice of regularization parameters $\lambda \geq 2(1-\gamma)\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_\infty$ and $\eta \geq 2\gamma\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_{op}$ with constant $\gamma \in [0,1]$, we have*

$$\tilde{\Delta} \in \mathcal{C}.$$

Proof of Lemma 1. Since $\tilde{\boldsymbol{\Theta}}$ is a local minimum of (1), the following first-order condition is satisfied. Note that $\boldsymbol{\Theta}^*$ is feasible, take $\boldsymbol{\Theta} = \boldsymbol{\Theta}^*$ in (5), we have

$$\langle \nabla\mathcal{L}_n(\tilde{\boldsymbol{\Theta}}) + \nabla(\lambda\|\tilde{\boldsymbol{\Theta}}\|_1) + \nabla\mathcal{R}_\eta(\tilde{\boldsymbol{\Theta}}), \boldsymbol{\Theta}^* - \tilde{\boldsymbol{\Theta}} \rangle \geq 0.$$

On the other hand, since $\mathcal{L}_n$ is convex, by the monotonicity of the subdifferential, we have the following.

$$\langle \nabla\mathcal{L}_n(\tilde{\boldsymbol{\Theta}}) - \nabla\mathcal{L}_n(\boldsymbol{\Theta}^*), \tilde{\Delta} \rangle \geq 0,$$

Combining the above two inequalities, we have

$$\langle -V_{\lambda,\eta}(\tilde{\boldsymbol{\Theta}}) - \nabla\mathcal{Q}_\eta(\tilde{\boldsymbol{\Theta}}) - \nabla\mathcal{L}_n(\boldsymbol{\Theta}^*), \tilde{\Delta} \rangle \geq 0.$$

By the monotonicity of the subdifferential of the convex function, we have $\langle V_{\lambda,\eta}(\tilde{\boldsymbol{\Theta}}) - V_{\lambda,\eta}(\boldsymbol{\Theta}^*), \tilde{\Delta} \rangle \geq 0$, and consequently

$$\langle -V_{\lambda,\eta}(\boldsymbol{\Theta}^*) - \nabla\mathcal{Q}_\eta(\tilde{\boldsymbol{\Theta}}) - \nabla\mathcal{L}_n(\boldsymbol{\Theta}^*), \tilde{\Delta} \rangle \geq 0. \tag{A.1}$$

For any sparse matrix $\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}$ with singular value decomposition $\boldsymbol{\Theta} = U'\Gamma V'$ with Γ being the diagonal matrix of nonincreasing singular values, then $V_{\lambda,\eta}(\boldsymbol{\Theta})$ can be decomposed as

$$V_{\lambda,\eta}(\boldsymbol{\Theta}) = \lambda\left(U'V'^{\top} + \Pi_{U'^{\perp}}\boldsymbol{W}_*\Pi_{V'^{\perp}}\right) + \eta\left(S_{\boldsymbol{\Theta}} + \boldsymbol{W}_1 \circ S^{\perp}_{\boldsymbol{\Theta}}\right), \tag{A.2}$$

for some $\|\boldsymbol{W}_1\|_\infty \leq 1, \|\boldsymbol{W}_*\|_{op} \leq 1$, see Watson (1992). Here, $S_{\boldsymbol{\Theta}} = sign(\boldsymbol{\Theta}) \in \{0, \pm 1\}^{d_1 \times d_2}$ be the signed sparsity pattern of $\boldsymbol{\Theta}$, and $S^{\perp}_{\boldsymbol{\Theta}} \in \{0,1\}^{d_1 \times d_2}$ the complementary sparsity pattern. The symbol "$\circ$" denotes the Hadamard product. Moreover, since the function $\|\cdot\|_1$ and $\|\cdot\|_*$ are both convex, we can choose $\boldsymbol{W}_1$ and $\boldsymbol{W}_*$ (e.g. $W_* = \sum_{i=1}^r u_i v_i^{\top}$, where

u_i, v_i are the i-th left and right eigenvectors associated with the nonzero singular values of $\Pi_{U_r^\perp}\tilde{\Delta}\Pi_{V_r^\perp}$, respectively.) such that

$$\begin{aligned}\langle -V_{\lambda,\eta}(\boldsymbol{\Theta}^*), \tilde{\Delta}\rangle &= -\lambda\langle \partial\|\boldsymbol{\Theta}^*\|_1, \tilde{\Delta}\rangle - \eta\langle \partial\|\boldsymbol{\Theta}^*\|_*, \tilde{\Delta}\rangle \\ &= -\lambda\langle S_{\boldsymbol{\Theta}^*}, \tilde{\Delta}\rangle - \lambda\|S^\perp_{\boldsymbol{\Theta}^*}\circ\tilde{\Delta}\|_1 - \eta\langle U'V'^\top, \tilde{\Delta}\rangle - \eta\|\Pi_{U_r^\perp}\tilde{\Delta}\Pi_{V_r^\perp}\|_* \\ &\le \lambda\|S_{\boldsymbol{\Theta}^*}\circ\tilde{\Delta}\|_1 - \lambda\|S^\perp_{\boldsymbol{\Theta}^*}\circ\tilde{\Delta}\|_1 + \eta\|\mathcal{P}_{\boldsymbol{\Theta}^*}(\tilde{\Delta})\|_* - \eta\|\Pi_{U_r^\perp}\tilde{\Delta}\Pi_{V_r^\perp}\|_*. \end{aligned} \tag{A.3}$$

Assume that $\tilde{\boldsymbol{\Theta}}$ has SVD as $\tilde{\boldsymbol{\Theta}} = \tilde{U}\tilde{\Gamma}\tilde{V}^\top$, then $\nabla\mathcal{Q}_\eta(\tilde{\boldsymbol{\Theta}}) = \tilde{U}q'_\eta(\tilde{\Gamma})\tilde{V}^\top$, where $q'_\eta(\Gamma)$ denotes the diagonal matrix with $q'_\eta(\cdot)$ apply on every diagonal element of Γ. Recall that a_η satisfies the regularity condition (iii) in Assumption A. In particular, all subgradients and derivatives of q_η are bounded in magnitude by η. Applying the duality of matrix norms, we have

$$\langle -\nabla\mathcal{Q}_\eta(\tilde{\boldsymbol{\Theta}}), \tilde{\Delta}\rangle \le \|\nabla\mathcal{Q}_\eta(\tilde{\boldsymbol{\Theta}})\|_{op}\|\tilde{\Delta}\|_* = \|\tilde{U}q'_\eta(\tilde{\Gamma})\tilde{V}^\top\|_{op}\|\tilde{\Delta}\|_* \le \eta\|\tilde{\Delta}\|_*, \tag{A.4}$$

for any parameter $\gamma \in [0,1]$, by the choice of $\lambda \ge 2(1-\gamma)\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_\infty$ and $\eta \ge 2\gamma\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_{op}$, we have

$$\begin{aligned}&\langle -\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*), \tilde{\Delta}\rangle \\ \le&(1-\gamma)|\langle \nabla\mathcal{L}_n(\boldsymbol{\Theta}^*), \tilde{\Delta}\rangle| + \gamma|\langle \nabla\mathcal{L}_n(\boldsymbol{\Theta}^*), \tilde{\Delta}\rangle| \\ \le&(1-\gamma)\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_\infty\|\tilde{\Delta}\|_1 + \gamma\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_{op}\|\tilde{\Delta}\|_* \\ \le&\frac{\lambda}{2}(\|S_{\boldsymbol{\Theta}^*}\circ\tilde{\Delta}\|_1 + \|S^\perp_{\boldsymbol{\Theta}^*}\circ\tilde{\Delta}\|_1) + \frac{\eta}{2}(\|\mathcal{P}_{\boldsymbol{\Theta}^*}(\tilde{\Delta})\|_* + \|\Pi_{U_r^\perp}\tilde{\Delta}\Pi_{V_r^\perp}\|_*). \end{aligned} \tag{A.5}$$

Combining (A.1)), (A.3)-(A.5), it follows that

$$\begin{aligned}0 \le&\frac{3\lambda}{2}\|S_{\boldsymbol{\Theta}^*}\circ\tilde{\Delta}\|_1 - \frac{\lambda}{2}\|S^\perp_{\boldsymbol{\Theta}^*}\circ\tilde{\Delta}\|_1 + \frac{3\eta}{2}\|\mathcal{P}_{\boldsymbol{\Theta}^*}(\tilde{\Delta})\|_* - \frac{\eta}{2}\|\Pi_{U_r^\perp}\tilde{\Delta}\Pi_{V_r^\perp}\|_* + \eta\|\tilde{\Delta}\|_* \\ \le&\frac{3\lambda}{2}\|S_{\boldsymbol{\Theta}^*}\circ\tilde{\Delta}\|_1 - \frac{\lambda}{2}\|S^\perp_{\boldsymbol{\Theta}^*}\circ\tilde{\Delta}\|_1 + \frac{5\eta}{2}\|\mathcal{P}_{\boldsymbol{\Theta}^*}(\tilde{\Delta})\|_* + \frac{\eta}{2}\|\Pi_{U_r^\perp}\tilde{\Delta}\Pi_{V_r^\perp}\|_* \\ \le&\frac{\max\{\lambda,\eta\}}{2}\left(3\|S_{\boldsymbol{\Theta}^*}\circ\tilde{\Delta}\|_1 - \|S^\perp_{\boldsymbol{\Theta}^*}\circ\tilde{\Delta}\|_1 + 5\|\mathcal{P}_{\boldsymbol{\Theta}^*}(\tilde{\Delta})\|_* + \|\Pi_{U_r^\perp}\tilde{\Delta}\Pi_{V_r^\perp}\|_*\right)\end{aligned}$$

which implies that $\tilde{\Delta} \in \mathcal{C}$. □

Before proving Theorem 1, we need to prove the following lemmas.

Lemma 2 *Under the Assumption A, for any $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \mathbb{R}^{d_1\times d_2}$, we have*

$$\mathcal{Q}_\eta(\boldsymbol{\Theta}_1) \ge \mathcal{Q}_\eta(\boldsymbol{\Theta}_2) - \langle \nabla\mathcal{Q}_\eta(\boldsymbol{\Theta}_2), \boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\rangle - \frac{\mu}{2}\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_F^2. \tag{A.6}$$

Proof of Lemma 2. For $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \mathbb{R}^{d_1\times d_2}$, we have the following singular value decompositions:

$$\boldsymbol{\Theta}_1 = U_1\Gamma_1V_1^\top \quad \text{and} \quad \boldsymbol{\Theta}_2 = U_2\Gamma_2V_2^\top,$$

where $\Gamma_1 = \mathrm{diag}(\sigma_1(\boldsymbol{\Theta}_1), \cdots, \sigma_d(\boldsymbol{\Theta}_1))$ and $\Gamma_2 = \mathrm{diag}(\sigma_1(\boldsymbol{\Theta}_2), \cdots, \sigma_d(\boldsymbol{\Theta}_2))$ are the diagonal matrices with descending singular values, respectively. According to (vi) in Assumption A, since $q_\eta(t) + \frac{\mu}{2}t^2$ is convex, for any $j \in \{1, \cdots, d\}$ we have

$$\begin{aligned}&\left(q'_\eta(\sigma_j(\boldsymbol{\Theta}_2)) + \mu\sigma_j(\boldsymbol{\Theta}_2)\right)(\sigma_j(\boldsymbol{\Theta}_1) - \sigma_j(\boldsymbol{\Theta}_2)) \\ \le&q_\eta(\sigma_j(\boldsymbol{\Theta}_1)) + \frac{\mu}{2}\sigma_j^2(\boldsymbol{\Theta}_1) - \left(q_\eta(\sigma_j(\boldsymbol{\Theta}_2)) + \frac{\mu}{2}\sigma_j^2(\boldsymbol{\Theta}_2)\right),\end{aligned}$$

then rearranging the terms gives

$$q'_\eta(\sigma_j(\boldsymbol{\Theta}_2))(\sigma_j(\boldsymbol{\Theta}_1) - \sigma_j(\boldsymbol{\Theta}_2)) \le q_\eta(\sigma_j(\boldsymbol{\Theta}_1)) - q_\eta(\sigma_j(\boldsymbol{\Theta}_2)) + \frac{\mu}{2}(\sigma_j(\boldsymbol{\Theta}_1) - \sigma_j(\boldsymbol{\Theta}_2))^2.$$

Adding all the j on both sides, and by the Weyl's inequality, we have

$$\langle q'_\eta(\Gamma_2), \Gamma_1 - \Gamma_2 \rangle \le Q_\eta(\boldsymbol{\Theta}_1) - Q_\eta(\boldsymbol{\Theta}_2) + \frac{\mu}{2}\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_F^2,$$

where $q'_\eta(\Gamma_2) = \mathrm{diag}(q'_\eta(\sigma_1(\boldsymbol{\Theta}_2)), \cdots, q'_\eta(\sigma_d(\boldsymbol{\Theta}_2)))$ denotes applying q'_η to every singular value along the diagonal of Γ_2.

Recall that $\nabla Q_\eta(\boldsymbol{\Theta}_2) = U_2 q'_\eta(\Gamma_2) V_2^\top$, let the combined orthogonal matrix $\bar{U} = [U_1, U_2]$ and $\bar{V} = [V_1, V_2]$, it follows that

$$\begin{aligned}
\langle \nabla \mathcal{Q}_\eta(\boldsymbol{\Theta}_2), \boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2 \rangle &= \left\langle \bar{U} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & q'_\eta(\Gamma_2) \end{bmatrix} \bar{V}, \bar{U} \begin{bmatrix} \Gamma_1 & \mathbf{0} \\ \mathbf{0} & -\Gamma_2 \end{bmatrix} \bar{V} \right\rangle \\
&= \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & q'_\eta(\Gamma_2) \end{bmatrix}, \begin{bmatrix} \Gamma_1 & \mathbf{0} \\ \mathbf{0} & -\Gamma_2 \end{bmatrix} \right\rangle \\
&= \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & q'_\eta(\Gamma_2) \end{bmatrix}, \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \Gamma_1 & -\Gamma_2 \end{bmatrix} \right\rangle \\
&= \left\langle \bar{U} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & q'_\eta(\Gamma_2) \end{bmatrix} \bar{V}, \bar{U} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \Gamma_1 & -\Gamma_2 \end{bmatrix} \bar{V} \right\rangle \\
&= \left\langle U_2 q'_\eta(\Gamma_2) V_2^\top, U_2(\Gamma_1 - \Gamma_2) V_2^\top \right\rangle \\
&= \langle q'_\eta(\Gamma_2), \Gamma_1 - \Gamma_2 \rangle \\
&\le Q_\eta(\boldsymbol{\Theta}_1) - Q_\eta(\boldsymbol{\Theta}_2) + \frac{\mu}{2}\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_F^2.
\end{aligned}$$

This completes the proof. □

Lemma 3 *Suppose that $\mathcal{R}_\eta$ satisfies all the conditions in Assumption A. Then we have*

(i) $\mathcal{R}_\eta$ is decomposable with respect to the pair $(\mathbb{M}, \overline{\mathbb{M}}^\perp)$.
(ii) For any matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$, we have

$$\mathcal{R}_\eta(A \pm B) \le \mathcal{R}_\eta(A) + \mathcal{R}_\eta(B);$$
$$\mathcal{R}_\eta(A \pm B) \ge \mathcal{R}_\eta(A) - \mathcal{R}_\eta(B).$$

Proof of Lemma 3. (i). By construction, the spaces $\mathbb{M}$ and $\overline{\mathbb{M}}^\perp$ are such that any matrix $A \in \mathbb{M}$ and $B \in \overline{\mathbb{M}}^\perp$ have orthogonal row and column spaces. For a matrix $A \in M$, its SVD can be written as $A = U_A \Gamma_A V_A^\top$, where U_A and V_A are orthogonal matrices, and Γ_A is a diagonal matrix containing the singular values of A in non-increasing order. Similarly, for a matrix $B \in \overline{\mathbb{M}}^\perp$: $B = U_B \Gamma_B V_B^\top$, where U_B and V_B are orthogonal matrices, and Γ_B is a diagonal matrix containing the singular values of B.

Given the orthogonality of the row and column spaces, the matrix $A+B$ can be viewed as having a block-diagonal form on the appropriate basis. Specifically, in the basis formed by the combined orthogonal matrices $U = [U_A, U_B]$ and $V = [V_A, V_B]$, the matrix $A+B$ can be expressed as:

$$A + B = U \begin{pmatrix} \Gamma_A & 0 \\ 0 & \Gamma_B \end{pmatrix} V^\top.$$

Recall that $\mathcal{R}_\eta(A+B) = \sum_{j=1}^d \rho_\eta(\sigma_j(A+B))$ with ρ_η applying to each diagonal element in the block-diagonal form. Therefore, it is straightforward to get

$$\begin{aligned}
\mathcal{R}_\eta(A+B) = \sum_{j=1}^d \rho_\eta(\sigma_j(A+B)) &= \sum_{\sigma_j \in \mathrm{diag}(\Gamma_A)} \rho_\eta(\sigma_j) + \sum_{\sigma_{j'} \in \mathrm{diag}(\Gamma_B)} \rho_\eta(\sigma_{j'}) \\
&= \sum_j \rho_\eta(\sigma_j(A)) + \sum_{j'} \rho_\eta(\sigma_{j'}(B)) = \mathcal{R}_\eta(A) + \mathcal{R}_\eta(B).
\end{aligned}$$

This establishes the decomposability of $\mathcal{R}_\eta$ with respect to the pair $(\mathbb{M}, \overline{\mathbb{M}}^\perp)$.

(ii). Recall that ρ_η being a concave function, according to Rotfel'd (1967), for any matrices $A, B \in \mathbb{R}^{d_1\times d_2}$, we have

$$\sum_{j=1}^{d}\rho_\eta(\sigma_j(A+B)) \le \sum_{j=1}^{d}\rho_\eta(\sigma_j(A)) + \sum_{j=1}^{d}\rho_\eta(\sigma_j(B)),$$

which implies the subadditivity of $\mathcal{R}_\eta$, i.e., $\mathcal{R}_\eta(A+B) \le \mathcal{R}_\eta(A)+\mathcal{R}_\eta(B)$. On the other hand, according to Yue and So (2016), we have

$$\sum_{j=1}^{d}\rho_\eta(\sigma_j(A)) - \sum_{j=1}^{d}\rho_\eta(\sigma_j(B)) \le \sum_{j=1}^{d}\rho_\eta(\sigma_j(A-B)),$$

which implies $\mathcal{R}_\eta(A-B) \ge \mathcal{R}_\eta(A)-\mathcal{R}_\eta(B)$. Notably, since $\mathcal{R}_\eta(B)=\mathcal{R}_\eta(-B)$, we also have $\mathcal{R}_\eta(A-B) \le \mathcal{R}_\eta(A)+\mathcal{R}_\eta(B)$ and $\mathcal{R}_\eta(A+B) \ge \mathcal{R}_\eta(A)-\mathcal{R}_\eta(B)$. □

Lemma 4 *Suppose that $\mathcal{R}_\eta$ satisfies all the conditions in Assumption A. Let the matrices $\boldsymbol{\Theta}^*, \boldsymbol{\Theta} \in \mathbb{B}_q(R_q)\cap\mathbb{R}^{d_1\times d_2}$ and denote $\Delta := \boldsymbol{\Theta}-\boldsymbol{\Theta}^*$. Let $\Delta = U\Gamma_\Delta V^\top$ be the SVD of Δ with $\Gamma_\Delta = diag(\sigma_1(\Delta),\cdots,\sigma_d(\Delta))$ be the diagonal matrix of singular values in descending order. Let $\Delta' = U\Gamma_{\Delta'}V^\top$ with $\Gamma_{\Delta'} = diag(\sigma_1(\Delta),\cdots,\sigma_{2r}(\Delta),0,\cdots,0)$ and $\Delta'' = \Delta-\Delta'$, then we have*

$$\mathcal{R}_\eta(\boldsymbol{\Theta}^*) - \mathcal{R}_\eta(\boldsymbol{\Theta}) \le \eta(\|\Delta'\|_* - \|\Delta''\|_*). \tag{A.7}$$

Proof of Lemma 4. From (i) in Assumption A, we have $\mathcal{R}_\eta(\Pi_{U_r^\perp}\boldsymbol{\Theta}^*\Pi_{V_r^\perp}) = \sum_j \rho_\eta(0) = 0$. Note that the decompositions $\boldsymbol{\Theta}^* = \mathcal{P}_{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}^*) + \Pi_{U_r^\perp}\boldsymbol{\Theta}^*\Pi_{V_r^\perp}$ and $\Delta = \mathcal{P}_\Delta(\boldsymbol{\Theta}^*) + \Pi_{U_r^\perp}\Delta\Pi_{V_r^\perp}$ hold. According to Lemma 3, it follows that

$$\begin{aligned}
\mathcal{R}_\eta(\boldsymbol{\Theta}^*) - \mathcal{R}_\eta(\boldsymbol{\Theta}) =& \mathcal{R}_\eta(\boldsymbol{\Theta}^*) - \mathcal{R}_\eta(\boldsymbol{\Theta}^*+\Delta)\\
=& \mathcal{R}_\eta(\mathcal{P}_{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}^*)) - \mathcal{R}_\eta\left(\mathcal{P}_{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}^*) + \mathcal{P}_{\boldsymbol{\Theta}^*}(\Delta) + \Pi_{U_r^\perp}\Delta\Pi_{V_r^\perp}\right)\\
\le& \mathcal{R}_\eta(\mathcal{P}_{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}^*)) - \mathcal{R}_\eta\left(\mathcal{P}_{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}^*) + \Pi_{U_r^\perp}\Delta\Pi_{V_r^\perp}\right) + \mathcal{R}_\eta\left(\mathcal{P}_{\boldsymbol{\Theta}^*}(\Delta)\right)\\
=& \mathcal{R}_\eta(\mathcal{P}_{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}^*)) - \mathcal{R}_\eta\left(\mathcal{P}_{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}^*)\right) - \mathcal{R}_\eta\left(\Pi_{U_r^\perp}\Delta\Pi_{V_r^\perp}\right) + \mathcal{R}_\eta\left(\mathcal{P}_{\boldsymbol{\Theta}^*}(\Delta)\right)\\
=& \mathcal{R}_\eta\left(\mathcal{P}_{\boldsymbol{\Theta}^*}(\Delta)\right) - \mathcal{R}_\eta\left(\Pi_{U_r^\perp}\Delta\Pi_{V_r^\perp}\right).
\end{aligned}$$

Note that $\mathrm{rank}(\mathcal{P}_{\boldsymbol{\Theta}^*}(\Delta)) = 2r$. Since when a matrix is projected onto a subspace, its singular values will typically decrease or remain the same, and from (ii) in Assumption A that ρ_η is nondecreasing, we have

$$\begin{aligned}
\mathcal{R}_\eta\left(\mathcal{P}_{\boldsymbol{\Theta}^*}(\Delta)\right) - \mathcal{R}_\eta\left(\Pi_{U_r^\perp}\Delta\Pi_{V_r^\perp}\right) &\le \mathcal{R}_\eta(\Delta') - \mathcal{R}_\eta(\Delta'')\\
&= \sum_{j\in\{1,\cdots,2r\}}\rho_\eta(\sigma_j(\Delta)) - \sum_{j\in\{2r+1,\cdots,d\}}\rho_\eta(\sigma_j(\Delta)),
\end{aligned}$$

Finally, according to Lemma 5 in Loh and Wainwright (2015), we have

$$\begin{aligned}
\sum_{j\in\{1,\cdots,2r\}}\rho_\eta(\sigma_j(\Delta)) - \sum_{j\in\{2r+1,\cdots,d\}}\rho_\eta(\sigma_j(\Delta)) &\le \eta\left(\sum_{j\in\{1,\cdots,2r\}}\sigma_j(\Delta) + \sum_{j\in\{2r+1,\cdots,d\}}\sigma_j(\Delta)\right)\\
&= \eta(\|\Delta'\|_* - \|\Delta''\|_*).
\end{aligned}$$

This completes the proof of the Lemma. □

Proof of Theorem 1. According to Lemma 1, we have $\tilde{\Delta} \in \mathcal{C}$, and according to the RSC condition (2) we have

$$\langle \nabla \mathcal{L}_n(\tilde{\Theta}) - \nabla \mathcal{L}_n(\Theta^*), \tilde{\Delta} \rangle \geq \alpha \|\tilde{\Delta}\|_F^2 - \tau \frac{d_1 + d_2}{n} \|\tilde{\Delta}\|_*^2.$$

Then, combining with (4) gives

$$\alpha \|\tilde{\Delta}\|_F^2 - \tau \frac{d_1 + d_2}{n} \|\tilde{\Delta}\|_*^2 \leq -\langle \nabla \mathcal{L}_n(\Theta^*), \tilde{\Delta} \rangle + \langle \nabla(\lambda \|\tilde{\Theta}\|_1) + \nabla \mathcal{R}_\eta(\tilde{\Theta}), \Theta^* - \tilde{\Theta} \rangle. \quad \text{(A.8)}$$

On the other hand, recall that $\mathcal{R}_\eta(\Theta) = \eta \|\Theta\|_* + \mathcal{Q}_\eta(\Theta)$. According to (A.6) in Lemma 2, we have

$$\begin{aligned}
\langle \nabla \mathcal{R}_\eta(\tilde{\Theta}), \Theta^* - \tilde{\Theta} \rangle &= \langle \nabla(\eta \|\tilde{\Theta}\|_*) + \nabla \mathcal{Q}_\eta(\tilde{\Theta}), \Theta^* - \tilde{\Theta} \rangle \\
&\leq \langle \nabla(\eta \|\tilde{\Theta}\|_*), \Theta^* - \tilde{\Theta} \rangle + \mathcal{Q}_\eta(\Theta^*) - \mathcal{Q}_\eta(\tilde{\Theta}) + \frac{\mu}{2} \|\Theta^* - \tilde{\Theta}\|_F^2 \\
&\leq \eta \|\Theta^*\|_* - \eta \|\tilde{\Theta}\|_* + \mathcal{Q}_\eta(\Theta^*) - \mathcal{Q}_\eta(\tilde{\Theta}) + \frac{\mu}{2} \|\Theta^* - \tilde{\Theta}\|_F^2 \\
&= \mathcal{R}_\eta(\Theta^*) - \mathcal{R}_\eta(\tilde{\Theta}) + \frac{\mu}{2} \|\Theta^* - \tilde{\Theta}\|_F^2, \qquad \text{(A.9)}
\end{aligned}$$

where the third inequality is due to the convexity of $\|\cdot\|_*$. It follows from (A.7) in Lemma 4 that

$$\mathcal{R}_\eta(\Theta^*) - \mathcal{R}_\eta(\tilde{\Theta}) \leq \eta(\|\tilde{\Delta}'\|_* - \|\tilde{\Delta}''\|_*), \qquad \text{(A.10)}$$

with $\tilde{\Delta}', \tilde{\Delta}''$ defined as in Lemma 4. Recall that $\nabla(\lambda \|\tilde{\Theta}\|_1) \in \lambda \partial \|\Theta\|_1$ is the subgradient of $\lambda \|\Theta\|_1$, then by the monotonicity of the subdiffrential we have $\langle \nabla(\lambda \|\Theta^*\|_1) - \nabla(\lambda \|\tilde{\Theta}\|_1), \tilde{\Delta} \rangle \geq 0$. Thus,

$$\begin{aligned}
\langle \nabla(\lambda \|\tilde{\Theta}\|_1), \Theta^* - \tilde{\Theta} \rangle &\leq \langle \nabla(\lambda \|\Theta^*\|_1), -\tilde{\Delta} \rangle \\
&= -\lambda \langle S_{\Theta^*}, \tilde{\Delta} \rangle - \lambda \|S_{\Theta^*}^\perp \circ \tilde{\Delta}\|_1 \\
&\leq \lambda \|S_{\Theta^*} \circ \tilde{\Delta}\|_1 - \lambda \|S_{\Theta^*}^\perp \circ \tilde{\Delta}\|_1. \qquad \text{(A.11)}
\end{aligned}$$

Then combining (A.8)-(A.11), we have

$$\begin{aligned}
&\alpha \|\tilde{\Delta}\|_F^2 - \tau \frac{d_1 + d_2}{n} \|\tilde{\Delta}\|_*^2 \\
\leq &- \langle \nabla \mathcal{L}_n(\Theta^*), \tilde{\Delta} \rangle + \lambda \|S_{\Theta^*} \circ \tilde{\Delta}\|_1 - \lambda \|S_{\Theta^*}^\perp \circ \tilde{\Delta}\|_1 \\
&\eta(\|\tilde{\Delta}'\|_* - \|\tilde{\Delta}''\|_*) + \frac{\mu}{2} \|\tilde{\Delta}\|_F^2 \\
\leq &\frac{3}{2} \lambda \|S_{\Theta^*} \circ \tilde{\Delta}\|_1 - \frac{1}{2} \lambda \|S_{\Theta^*}^\perp \circ \tilde{\Delta}\|_1 + \frac{3}{2} \eta \|\tilde{\Delta}'\|_* - \frac{1}{2} \eta \|\tilde{\Delta}''\|_* \\
&+ \frac{\mu}{2} \|\tilde{\Delta}\|_F^2,
\end{aligned}$$

where the last inequality is since $\tilde{\Delta} = \tilde{\Delta}' + \tilde{\Delta}''$, by the choice of $\lambda \geq 2(1 - \gamma) \|\nabla \mathcal{L}_n(\Theta^*)\|_\infty$ and $\eta \geq 2\gamma \|\nabla \mathcal{L}_n(\Theta^*)\|_{op}$, we have

$$\begin{aligned}
\langle -\nabla \mathcal{L}_n(\Theta^*), \tilde{\Delta} \rangle &\leq (1 - \gamma) |\langle \nabla \mathcal{L}_n(\Theta^*), \tilde{\Delta} \rangle| + \gamma |\langle \nabla \mathcal{L}_n(\Theta^*), \tilde{\Delta} \rangle| \\
&\leq (1 - \gamma) \|\nabla \mathcal{L}_n(\Theta^*)\|_\infty \|\tilde{\Delta}\|_1 + \gamma \|\nabla \mathcal{L}_n(\Theta^*)\|_{op} \|\tilde{\Delta}\|_* \\
&\leq \frac{\lambda}{2} (\|S_{\Theta^*} \circ \tilde{\Delta}\|_1 + \|S_{\Theta^*}^\perp \circ \tilde{\Delta}\|_1) + \frac{\eta}{2} (\|\tilde{\Delta}'\|_* + \|\tilde{\Delta}''\|_*).
\end{aligned}$$

Provided $\alpha > \frac{\mu}{2}$, recall that $\|\tilde{\Delta}\|_* \leq 2R/\omega$ and $n \geq \frac{16R^2\tau^2}{\alpha^2 w^2}(d_1 + d_2)$, by the choice of $\eta \geq 2\alpha\sqrt{\frac{d_1+d_2}{n}}$, we have

$$\begin{aligned}
2\left(\alpha - \frac{\mu}{2}\right)\|\tilde{\Delta}\|_F^2 &\leq 3\lambda\|S_{\boldsymbol{\Theta}^*} \circ \tilde{\Delta}\|_1 - \lambda\|S_{\boldsymbol{\Theta}^*}^{\perp} \circ \tilde{\Delta}\|_1 + 3\eta\|\tilde{\Delta}'\|_* - \eta\|\tilde{\Delta}''\|_* \\
&\quad + 4\tau\frac{R}{\omega}\frac{d_1 + d_2}{n}\|\tilde{\Delta}\|_* \\
&\leq 3\lambda\|S_{\boldsymbol{\Theta}^*} \circ \tilde{\Delta}\|_1 - \lambda\|S_{\boldsymbol{\Theta}^*}^{\perp} \circ \tilde{\Delta}\|_1 + 3\eta\|\tilde{\Delta}'\|_* - \eta\|\tilde{\Delta}''\|_* \\
&\quad + \alpha\sqrt{\frac{d_1 + d_2}{n}}\|\tilde{\Delta}\|_* \\
&\leq 3\lambda\|S_{\boldsymbol{\Theta}^*} \circ \tilde{\Delta}\|_1 - \lambda\|S_{\boldsymbol{\Theta}^*}^{\perp} \circ \tilde{\Delta}\|_1 + \frac{7}{2}\eta\|\tilde{\Delta}'\|_* - \frac{1}{2}\eta\|\tilde{\Delta}''\|_*. \qquad (A.12)
\end{aligned}$$

which implies that

$$\begin{aligned}
2\left(\alpha - \frac{\mu}{2}\right)\|\tilde{\Delta}\|_F^2 &\leq 3\lambda\|S_{\boldsymbol{\Theta}^*} \circ \tilde{\Delta}\|_1 + \frac{7}{2}\eta\|\tilde{\Delta}'\|_* \\
&\leq \left(3\lambda\sqrt{s} + \frac{7}{2}\eta\sqrt{2r}\right)\|\tilde{\Delta}\|_F,
\end{aligned}$$

which implies

$$\|\tilde{\Delta}\|_F^2 \leq \frac{\left(3\lambda\sqrt{s} + \frac{7}{2}\eta\sqrt{2r}\right)^2}{4\left(\alpha - \frac{\mu}{2}\right)^2} \leq \frac{9\lambda^2 s}{2\left(\alpha - \frac{\mu}{2}\right)^2} + \frac{49\eta^2 r}{4\left(\alpha - \frac{\mu}{2}\right)^2}. \qquad (A.13)$$

Moreover, (A.12) implies that

$$\frac{1}{2}\eta\|\tilde{\Delta}''\|_* \leq 3\lambda\|S_{\boldsymbol{\Theta}^*} \circ \tilde{\Delta}\|_1 + \frac{7}{2}\eta\|\tilde{\Delta}'\|_*,$$

which combines with (A.13) further implies

$$\begin{aligned}
\|\tilde{\Delta}\|_* &\leq \frac{1}{\eta}\left(6\lambda\|S_{\boldsymbol{\Theta}^*} \circ \tilde{\Delta}\|_1 + 8\eta\|\tilde{\Delta}'\|_*\right) \\
&\leq \frac{2}{\eta}\left(3\lambda\sqrt{s} + 4\eta\sqrt{2r}\right)\|\tilde{\Delta}\|_F \\
&\leq \frac{\left(3\lambda\sqrt{s} + 4\eta\sqrt{2r}\right)^2}{\eta(\alpha - \frac{\mu}{2})}.
\end{aligned}$$

Thereby completing the proof of Theorem 1.

□

A.2 Proof of Proposition 1, Theorem 2 and Corollary 1

Lemma 5 *Under the conditions of (C.1) and (C.2), we have*

$$P\left(\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_{op} \geq c\sqrt{\frac{d_1 + d_2}{n}}\right) \leq c_1 \exp\left(-c_2(d_1 + d_2)\right) \qquad (A.14)$$

and

$$P\left(\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_\infty \geq c'\sqrt{\frac{\log(d_1 d_2)}{n}}\right) \leq c_1' \exp\left(-c_2' \log(d_1 d_2)\right), \qquad (A.15)$$

where $c, c', c_1, c_1', c_2, c_2'$ are some positive constants.

Proof of Lemma 5. Note that $\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*) = \frac{1}{n}\sum_{i=1}^n (b'(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle) - Y_i)\,\boldsymbol{X}_i$, it follows that

$$\begin{aligned}\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_{op} &= \|\frac{1}{n}\sum_{i=1}^n \left(b'(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle) - Y_i\right)\boldsymbol{X}_i\|_{op} \\ &= \|\frac{1}{n}\sum_{i=1}^n \left(b'(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle) - Y_i\right)\boldsymbol{X}_i - \mathbb{E}\left[\left(b'(\langle \boldsymbol{X}, \boldsymbol{\Theta}^*\rangle) - Y\right)\boldsymbol{X}\right]\|_{op},\end{aligned}$$

where the last equality is true since $\mathbb{E}[Y|\boldsymbol{X}] = b'(\langle \boldsymbol{X}, \boldsymbol{\Theta}^*\rangle)$.

To bound the above operator norm, we consider the 1/4-cover of $\mathcal{S}^{d_1-1}$ and $\mathcal{S}^{d_2-1}$, where $S^{d-1} := \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ denote the the unit ball in $\mathbb{R}^d$. Let $\boldsymbol{B}$ be any matrix of size $d_1 \times d_2$. Denote $\mathcal{N}^{d_1}$ and $\mathcal{N}^{d_2}$ as the 1/4-nets for $\mathcal{S}^{d_1-1}$ and $\mathcal{S}^{d_2-1}$, respectively. Note that by the definition of 1/4-net, for any $u \in \mathcal{S}^{d_1-1}$ and $v \in \mathcal{S}^{d_2-1}$, there exist $\tilde{u} \in \mathcal{N}^{d_1}$ and $\tilde{v} \in \mathcal{N}^{d_2}$ such that $\|u - \tilde{u}\|_2 \le 1/4$ and $\|v - \tilde{v}\|_2 \le 1/4$. It follows that

$$\begin{aligned}u^\top \boldsymbol{B} v &= u^\top \boldsymbol{B}(v - \tilde{v}) + (u - \tilde{u})^\top B\tilde{v} + \tilde{u}^\top \boldsymbol{B}\tilde{v} \\ &\le \frac{1}{4}\|\boldsymbol{B}\|_{op} + \frac{1}{4}\|\boldsymbol{B}\|_{op} + \sup_{\tilde{u}\in\mathcal{N}^{d_1}, \tilde{v}\in\mathcal{N}^{d_2}} \tilde{u}^\top \boldsymbol{B}\tilde{v} \\ &= \sup_{\tilde{u}\in\mathcal{N}^{d_1}, \tilde{v}\in\mathcal{N}^{d_2}} \tilde{u}^\top \boldsymbol{B}\tilde{v} + \frac{1}{2}\|\boldsymbol{B}\|_{op}.\end{aligned}$$

Taking the supremum over $u \in \mathcal{S}^{d_1-1}$ and $v \in \mathcal{S}^{d_2-1}$ yields that

$$\|\boldsymbol{B}\|_{op} \le 2 \sup_{\tilde{u}\in\mathcal{N}^{d_1}, \tilde{v}\in\mathcal{N}^{d_2}} \tilde{u}^\top \boldsymbol{B}\tilde{v}. \tag{A.16}$$

In the remaining proof, for fixed $u \in \mathcal{N}^{d_1}$ and $v \in \mathcal{N}^{d_2}$, for convenience, we denote the random variables $Z_i = u^\top \boldsymbol{X}_i v$ and $Z = u^\top \boldsymbol{X} v$. According to Lemma 2.7.7 in Vershynin (2018) of the relationship between sub-Gaussian norm and sub-exponential norm, combined with the sub-Gaussian norm for bounded variables, we have

$$\|\left(b'(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle) - Y_i\right) Z_i\|_{\psi_1} \le \|b'(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle) - Y_i\|_{\psi_2}\|Z_i\|_{\psi_2} \le M\kappa_{\boldsymbol{X}},$$

where the last inequality is because the variance of the response Y is bounded according to (C.2). Then by the Bernstein inequality (see Theorem 2.8.1 in Vershynin (2018)), for sufficiently small t, it follows that

$$\begin{aligned}&P\left(\left|\frac{1}{n}\sum_{i=1}^n \left(b'(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle) - Y_i\right)\boldsymbol{Z}_i - \mathbb{E}\left[\left(b'(\langle \boldsymbol{X}, \boldsymbol{\Theta}^*\rangle) - Y\right)\boldsymbol{Z}\right]\right| \ge t\right) \\ &\qquad \le 2\exp\left(-\frac{c_1 n t^2}{M^2\kappa_{\boldsymbol{X}}^2}\right),\end{aligned}$$

where the constant $c_1 > 0$. Thus, taking the union bound of all points in $\mathcal{N}^{d_1} \times \mathcal{N}^{d_2}$ with $|\mathcal{N}^{d_1} \times \mathcal{N}^{d_2}| = (1 + \frac{2}{\epsilon})^{d_1+d_2} = 9^{d_1+d_2}$ gives

$$\begin{aligned}&P\left(\left\|\frac{1}{n}\sum_{i=1}^n \left(b'(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle) - Y_i\right)\boldsymbol{X}_i - \mathbb{E}\left[\left(b'(\langle \boldsymbol{X}, \boldsymbol{\Theta}^*\rangle) - Y\right)\boldsymbol{X}\right]\right\|_{op} \ge 2t\right) \\ &\qquad \le 2\cdot 9^{d_1+d_2}\exp\left\{-\frac{c_1 n t^2}{M^2\kappa_{\boldsymbol{X}}^2}\right\}.\end{aligned}$$

Recall that $\mathbb{E}\left[\left(b'(\langle \boldsymbol{X}, \boldsymbol{\Theta}^*\rangle) - Y\right)\boldsymbol{Z}\right] = 0$. If we choose $t = c\sqrt{\frac{d_1+d_2}{n}}$, for any $\eta > 0$, there exists a constant $\gamma > 0$ such that $\max\{d_1, d_2\}/n < \gamma$, then it holds that

$$\begin{aligned}&P\left(\left\|\frac{1}{n}\sum_{i=1}^n \left(b'(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle) - Y_i\right)\boldsymbol{X}_i\right\|_{op} \ge c\sqrt{\frac{d_1+d_2}{n}}\right) \\ &\qquad \le c_1\exp\left(-c_2(d_1+d_2)\right),\end{aligned}$$

which implies (A.14).

Next, for each $1 \leq j \leq d_1$ and $1 \leq k \leq d_2$, we denote $p = jk \in \{1, \cdots, d_1 d_2\}$. For each $1 \leq i \leq n$, define the random vector $\boldsymbol{x}_i = \mathrm{vec}(\boldsymbol{X}_i)$, and define the random variable

$$U_{ip} := \left(b'(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle) - Y_i\right) \boldsymbol{x}_{ip},$$

where $\boldsymbol{x}_{ip}$ denotes the p-th element of random vector $\boldsymbol{x}_i$. Since $\boldsymbol{x}$ is sub-Gaussain, then for any $v \in \mathbb{R}^{d_1 d_2}$, the variable $v^\top \boldsymbol{x}$ is sub-Gaussian with parameter at most $\sigma_{\boldsymbol{x}}^2 \|v\|_2^2$.

Note that

$$\begin{aligned} P\left(\|\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*)\|_\infty \geq c' \sqrt{\frac{\log(d_1 d_2)}{n}}\right) &= P\left(\max_{1 \leq p \leq d_1 d_2} \left| \frac{1}{n} \sum_{i=1}^n U_{ip} \right| \geq t\right) \\ &\leq P(\mathcal{S}^c) + P\left(\max_{1 \leq p \leq d_1 d_2} \left| \frac{1}{n} \sum_{i=1}^n U_{ip} \right| \geq t | \mathcal{S}\right), \end{aligned} \tag{A.17}$$

where we define $\mathcal{S} := \left\{ \max_{p=1,\cdots,d_1 d_2} \left\{ \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_{ip}^2 \right\} \leq 2\sigma_{\boldsymbol{x}}^2 \right\}$. The standard bounds for the norms of sub-Gaussian vectors yields

$$P\left(\left| \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_{ip}^2 \right| \geq 2\sigma_{\boldsymbol{x}}^2\right) \leq 2\exp(-cn). \tag{A.18}$$

Thus, taking the union bound implies that

$$P(\mathcal{S}^c) \leq 2\exp(-cn + \log(d_1 d_2)) \leq c_1 \exp(-c_2 n),$$

where the last inequality follows since $n \gg \log(d_1 d_2)$. Here, c, c_1, c_2 are some universal constants. Next, it remains to bound the second term in (A.17). For any $t \in \mathbb{R}$, we have

$$\begin{aligned} \log \mathbb{E}[\exp(tU_{ip}) | \boldsymbol{x}_i] &= \log\left(\exp(tb'(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle)\boldsymbol{x}_{ip}) \cdot \mathbb{E}[-tY_i \boldsymbol{x}_{ip}]\right) \\ &= t\boldsymbol{x}_{ip} b'(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle) + b\left(-t\boldsymbol{x}_{ip} + \langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle\right) - b\left(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle\right), \end{aligned}$$

where the last equality is because $b(\cdot)$ is the cumulant generating function for the exponential family. Then, by Taylor's expansion, there exist some constants $c_i \in [0, 1]$ such that

$$\log \mathbb{E}[\exp(tU_{ip}) | \boldsymbol{x}_i] = \frac{t^2 \boldsymbol{x}_{ip}^2}{2} b''\left(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle - c_i t \boldsymbol{x}_{ip}\right) \leq \frac{M t^2 \boldsymbol{x}_{ip}^2}{2}, \tag{A.19}$$

since condition (C.2) ensures the soundness of b''. Consequently, (A.19) implies that conditioned on $\mathcal{S}$, $\frac{1}{n} \sum_{i=1}^n U_{ip}$ is sub-Gaussian with parameter at most $\sigma_U = M \max_{p=1,\cdots,d_1 d_2} \mathbb{E}[\boldsymbol{x}_{ip}^2]$. Then by the tail bound of the sub-Gaussian variable and the union bound, we have

$$P\left(\max_{1 \leq p \leq d_1 d_2} \left| \frac{1}{n} \sum_{i=1}^n U_{ip} \right| \geq t | \mathcal{S}\right) \leq d_1 d_2 \exp\left(-\frac{nt^2}{2\sigma_U^2}\right). \tag{A.20}$$

Taking $t = c' \sqrt{\frac{\log(d_1 d_2)}{n}}$, (A.17), (A.18) and (A.20) combined finish the proof of (A.15). □

Lemma 6 *For any random matrix $\boldsymbol{Z}_i \in \mathbb{R}^{d_1 \times d_2}$ that satisfies Condition 1, we have*

$$\mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^n \sigma_i \boldsymbol{Z}_i \right\|_{op}\right] \leq C \sqrt{\frac{d_1 + d_2}{n}}, \tag{A.21}$$

where $\{\sigma_i\}_{i=1}^n$ are i.i.d. Rademacher random variables satisfying $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$.

Proof of Lemma 6. Following the derivation of (A.16). Consider the matrix $\boldsymbol{B} = \sum_{i=1}^{n} \sigma_i \boldsymbol{X}_i$, then for any $s > 0$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left\{s\left\|\sum_{i=1}^{n}\sigma_i \boldsymbol{Z}_i\right\|_{op}\right\}\right] \\
\leq &\mathbb{E}\left[\exp\left\{2s\sup_{u\in\mathcal{N}^p, v\in\mathcal{N}^q} u^\top\left(\sum_{i=1}^{n}\sigma_i \boldsymbol{Z}_i\right)v\right\}\right] \\
\leq &\mathbb{E}\left[\exp\left\{2s\sum_{u\in\mathcal{N}^p, v\in\mathcal{N}^q} u^\top\left(\sum_{i=1}^{n}\sigma_i \boldsymbol{Z}_i\right)v\right\}\right] \\
\leq &9^{d_1+d_2}\sup_{u\in\mathcal{N}^p, v\in\mathcal{N}^q}\mathbb{E}\left[\exp\left\{2su^\top\left(\sum_{i=1}^{n}\sigma_i \boldsymbol{Z}_i\right)v\right\}\right] \\
\leq &9^{d_1+d_2}\sup_{u\in\mathcal{N}^p, v\in\mathcal{N}^q}\prod_{i=1}^{n}\mathbb{E}\left[\exp\left\{2su^\top\left(\sigma_i \boldsymbol{Z}_i\right)v\right\}\right] \\
\leq &9^{d_1+d_2}\exp\{Cs^2\} \leq \exp\{C(s^2+d_1+d_2)\},
\end{aligned} \tag{A.22}
$$

where the final inequality is derived from the property that, due to the sub-Gaussianity of the vectorized data $\boldsymbol{z}_i = \mathrm{vec}(\boldsymbol{Z}_i)$ and the boundedness of σ_i, the product $\sigma_i \boldsymbol{z}_i$ is also sub-Gaussian, with its moment generating function satisfying $\mathbb{E}\left[e^{sa^\top \frac{1}{n}\sum_i \sigma_i \boldsymbol{z}_i}\right] \leq e^{Cs^2\|a\|^2/n}$. Adhering to the approach employed in the proof of Lemma 3 by Lu et al. (2024), we proceed with the decomposition

$$
\begin{aligned}
\left\|\sum_{i=1}^{n}\sigma_i \boldsymbol{Z}_i\right\|_{op} &= \left\|\sum_{i=1}^{n}\sigma_i \boldsymbol{Z}_i\right\|_{op}\mathbb{I}\left\{\left\|\sum_{i=1}^{n}\sigma_i \boldsymbol{Z}_i\right\|_{op} \geq 4C\sqrt{d_1+d_2}\right\} \\
&\quad + \left\|\sum_{i=1}^{n}\sigma_i \boldsymbol{Z}_i\right\|_{op}\mathbb{I}\left\{\left\|\sum_{i=1}^{n}\sigma_i \boldsymbol{Z}_i\right\|_{op} < 4C\sqrt{d_1+d_2}\right\} \\
&=: Q_1 + Q_2.
\end{aligned}
$$

According to the definition of Q_1 and Equation (A.22), for a constant C_1 (at least 4 times C), it follows from the Markov inequality that

$$
\begin{aligned}
P\left(Q_1 \geq C_1\sqrt{d_1+d_2}\right) &= P\left(Q_1 \geq C_1\sqrt{d_1+d_2}\right) \\
&\leq \exp\{-sC_1\sqrt{d_1+d_2}\}\exp\{C(s^2+d_1+d_2)\} \\
&\leq \exp\left\{-\frac{C_1^2(d_1+d_2)}{4C} + C(d_1+d_2)\right\} \\
&\leq \exp\left\{\left(-\frac{C_1^2}{4C} + \frac{C_1^2}{8C}\right)(d_1+d_2)\right\} \\
&= \exp\left\{-\frac{C_1^2(d_1+d_2)}{8C}\right\},
\end{aligned} \tag{A.23}
$$

where the third inequality is obtained by setting $s = \frac{C_1\sqrt{d_1+d_2}}{2C}$, and the penultimate inequality results from the condition $C_1 \geq 4C$. Conversely, when $C_1 < 4C$, based on the definition of Q_1 and the derivation of equation (A.23), we have

$$
\begin{aligned}
P\left(Q_1 \geq C_1\sqrt{d_1+d_2}\right) = P\left(Q_1 \geq 4C\sqrt{d_1+d_2}\right) &\leq \exp\left\{-2C(d_1+d_2)\right\} \\
&\leq \exp\left\{-\frac{C_1^2(d_1+d_2)}{8C}\right\}.
\end{aligned} \tag{A.24}
$$

Combining equations (A.23) and (A.24) yields $P(Q_1 \geq t) \leq e^{-\frac{1}{8C}t^2}$. According to Lemma 2.2.1 in Van Der Vaart et al. (1996), we have that the ψ_2-norm of Q_1, denoted $\|Q_1\|_{\psi_2}$, is upper-bounded by $4\sqrt{C}/C_1$. This follows directly from the definition of the ψ_2-norm.

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n}\sigma_i \boldsymbol{Z}_i\right\|_{op}\right] \leq \left\|\left\|\sum_{i=1}^{n}\sigma_i \boldsymbol{Z}_i\right\|_{op}\right\|_{\psi_2} \leq \|Q_1\|_{\psi_2} + \|Q_2\|_{\psi_2} \leq C\sqrt{d_1+d_2},$$

where the final inequality is due to the fact that $\|Q_2\|_{\psi_2} \leq 4C\sqrt{d_1+d_2}$, as per the definition of Q_2. Consequently, we have $\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\sigma_i \boldsymbol{Z}_i\right\|_{op}\right] \leq C\sqrt{\frac{d_1+d_2}{n}}$. This concludes the proof of the lemma. □

Proof of Proposition 1. Applying a second-order Taylor expansion to the negative log-likelihood (11) between $\boldsymbol{\Theta}^* + \Delta$ and $\boldsymbol{\Theta}^*$, we conclude that for some $c_i \in [0,1]$,

$$\begin{aligned}\mathcal{E}_n(\Delta) &= \frac{1}{n}\sum_{i=1}^{n}\left(b'(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* + \Delta\rangle) - b'(\langle\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle)\right)\langle \boldsymbol{X}_i, \Delta\rangle \\ &= \frac{1}{n}\sum_{i=1}^{n} b''(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle + c_i\langle \boldsymbol{X}_i, \Delta\rangle)(\langle \boldsymbol{X}_i, \Delta\rangle)^2.\end{aligned}$$

The framework of the proof follows the lines in the proof of Proposition 2 in Negahban et al. (2012), we write this here to ensure the completeness of the proof. For a truncation level $r > 0$ to be chosen later, define the truncation functions:

$$\phi_r(t) = \begin{cases} t^2, & \text{if } |t| \leq \frac{r}{2}, \\ (r-t)^2, & \text{if } \frac{r}{2} \leq |t| \leq r, \\ 0, & \text{otherwise.}\end{cases} \qquad \text{and} \qquad \psi_r(t) = \begin{cases} t, & \text{if } |t| \leq r, \\ 0, & \text{otherwise.}\end{cases} \tag{A.25}$$

Introducing the truncation level $R \geq r$, consider the event $|\langle \boldsymbol{\Theta}^*, \boldsymbol{X}_i\rangle| > R$ or $|\langle \Delta, \boldsymbol{X}_i\rangle| > r$. Simple calculation shows

$$\begin{aligned}&b''(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle + c_i\langle \boldsymbol{X}_i, \Delta\rangle)(\langle \boldsymbol{X}_i, \Delta\rangle)^2 \\ \geq& b''(\psi_R(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle) + c_i\psi_r(\langle \boldsymbol{X}_i, \Delta\rangle))\phi_r\left(\langle \boldsymbol{X}_i, \Delta\rangle \mathbb{I}\{|\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle| \leq R\}\right).\end{aligned}$$

Denoting $L_b(R) = \min_{|t|\leq 2R} b''(t)$, adding the n samples, we have $b''(\psi_R(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle) + c_i\psi_r(\langle \boldsymbol{X}_i, \Delta\rangle)) \geq L_b(R)$. Consequently, we have

$$\mathcal{E}_n(\Delta) \geq L_b(R)\mathbb{E}_n\left[\phi_r\left(\langle \boldsymbol{X}, \Delta\rangle \mathbb{I}\{|\langle \boldsymbol{X}, \boldsymbol{\Theta}^*\rangle| \leq R\}\right)\right],$$

where $\mathbb{E}_n$ denotes the empirical expectation over n samples. For a given positive constant $\delta \in (0,1]$, we choose the following transaction parameters

$$R^2 := C_3 \qquad \text{and} \qquad r^2(\delta) := C_3\delta^2,$$

where $C_3 := \max\{1, 8\kappa_{\boldsymbol{X}}^2 \log\left(\frac{24\kappa_{\boldsymbol{X}}^2}{\kappa_l^2}\right)\}$. Next, it is sufficient to show

$$\mathbb{E}_n\left[\phi_r\left(\langle \boldsymbol{X}, \Delta\rangle \mathbb{I}\{|\langle \boldsymbol{X}, \boldsymbol{\Theta}^*\rangle| \leq R\}\right)\right] \geq \tilde{\alpha}\|\Delta\|_F^2 - \tilde{\tau}\sqrt{\frac{d_1+d_2}{n}}\|\Delta\|_*\|\Delta\|_F, \qquad \forall \|\Delta\|_F \leq \delta,$$

According to the homogeneity property $\phi_c(cz) = c^2\phi_1(z)$ addressed by Negahban et al. (2012), we can restrict our attention to $\delta = 1$. For convenience, denote

$$g_\Delta(\boldsymbol{X}) = \phi_r\left(\langle \boldsymbol{X}, \Delta\rangle \mathbb{I}\{|\langle \boldsymbol{X}, \boldsymbol{\Theta}^*\rangle| \leq R\}\right).$$

For each radius t, define the event

$$\mathcal{B}(t) := \left\{\mathbb{E}_n\left[g_\Delta(\boldsymbol{X})\right] < \kappa_1\left(1 - \kappa_2\sqrt{\frac{d_1+d_2}{n}}t\right), \text{for } \Delta \in \mathbb{S}_*(t) \cap \mathbb{S}_F(1)\right\},$$

where $\mathbb{S}_*(t) = \{\Delta \in \mathbb{R}^{d_1\times d_2} : \|\Delta\|_* = t\}$ and $\mathbb{S}_F(1) = \{\Delta \in \mathbb{R}^{d_1\times d_2} : \|\Delta\|_F = 1\}$ denote the spherical shell of F-norm radius one and l_1 norm radius t, respectively. To show that $P(\mathcal{B}(t))$ is sufficiently small, we define the random variable

$$Z(t) := \sup_{\Delta\in\mathbb{S}_*(t)\cap\mathbb{S}_F(1)} \left|(\mathbb{E}_n - \mathbb{E})\left[g_\Delta(\boldsymbol{X})\right]\right|.$$

To show that the probability of this event $\mathcal{B}(t)$ is very small, we take the following two steps:
(i) First, we show that for any $\Delta \in \mathbb{S}_F(1)$, we have

$$\mathbb{E}[g_\Delta(\boldsymbol{X})] \geq \frac{\kappa_l}{2}. \tag{A.26}$$

(ii) Second, we show the tail bound of $Z(t)$ by

$$P\left(Z(t) \geq \frac{\kappa_l}{4} + CK_3\kappa_X t\sqrt{\frac{d_1+d_2}{n}}\right) \leq c_1\exp(-c_2 n - c_2' t(d_1+d_2)). \tag{A.27}$$

According to Condition 1, for $\Delta \in \mathbb{S}_F(1)$, we have $\mathbb{E}[\langle \boldsymbol{X},\Delta\rangle^2] \geq \kappa_l\|\Delta\|_F^2 = \kappa_l$. Recall the definition of $g_\Delta(\boldsymbol{X})$, we have

$$\mathbb{E}_{\boldsymbol{X}}[\langle \boldsymbol{X},\Delta\rangle^2 - g_\Delta(\boldsymbol{X})] \leq \mathbb{E}_{\boldsymbol{X}}[\langle \boldsymbol{X},\Delta\rangle^2\mathbb{I}\{|\langle \boldsymbol{X},\boldsymbol{\Theta}^*\rangle| \geq R\}] + \mathbb{E}_{\boldsymbol{X}}[\langle \boldsymbol{X},\Delta\rangle^2\mathbb{I}\{\langle \boldsymbol{X},\Delta\rangle \geq \frac{r}{2}\}].$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}}[\langle \boldsymbol{X},\Delta\rangle^2\mathbb{I}\{|\langle \boldsymbol{X},\boldsymbol{\Theta}^*\rangle| \geq R\}] &\leq \sqrt{\mathbb{E}_{\boldsymbol{X}}[\langle \boldsymbol{X},\Delta\rangle^4]}\sqrt{P(|\langle \boldsymbol{X},\boldsymbol{\Theta}^*\rangle| \geq R)} \\
&\leq 3\kappa_{\boldsymbol{X}}^4\cdot 2\exp\left(-\frac{R^2}{2\kappa_{\boldsymbol{X}}^2}\right) \\
&= 6\kappa_{\boldsymbol{X}}^2\exp\left(-\frac{R^2}{2\kappa_{\boldsymbol{X}}^2}\right),
\end{aligned}$$

where the penultimate inequality is because $\langle \boldsymbol{X},\Delta\rangle$ and $\langle \boldsymbol{X},\boldsymbol{\Theta}^*\rangle$ are both sub-Gaussian with parameter $\kappa_{\boldsymbol{X}}$. Similar arguments yields

$$\mathbb{E}_{\boldsymbol{X}}[\langle \boldsymbol{X},\Delta\rangle^2\mathbb{I}\{|\langle \boldsymbol{X},\Delta\rangle| \geq \frac{r}{2}\}] \leq 6\kappa_{\boldsymbol{X}}^2\exp\left(-\frac{r^2}{8\kappa_{\boldsymbol{X}}^2}\right).$$

Therefore, for specific choice of $r^2 = R^2 = C_3 := \max\{1, 8\kappa_{\boldsymbol{X}}^2\log\left(\frac{24\kappa_{\boldsymbol{X}}^2}{\kappa_l^2}\right)\}$, each term above is at most $\kappa_l/4$, which complete the proof of (A.26).

Next, note that $\forall\Delta \in \mathbb{S}_F(1)$, we have $\|g_\Delta\|_\infty \leq r^2 = C_3$. Then by the bounded difference inequality, for any $u > 0$, we have

$$P\left(Z(t) - \mathbb{E}[Z(t)] \geq u\right) \leq 2\exp\left(-\frac{nu^2}{4C_3^2}\right).$$

Taking $u = \frac{\kappa_l}{4} + 2C_3\kappa_{\boldsymbol{X}}\sqrt{\frac{d_1+d_2}{n}}t$, we obtain

$$P\left(Z(t) \geq \mathbb{E}[Z(t)] + \frac{\kappa_l}{4} + 2C_3\kappa_{\boldsymbol{X}}\sqrt{\frac{d_1+d_2}{n}}t\right) \leq 2\exp\left(-\frac{n\kappa_l^2}{64C_3^2} - \kappa_{\boldsymbol{X}}^2 t(d_1+d_2)\right) \tag{A.28}$$

It remains to bound $\mathbb{E}[Z(t)]$. Note that by the definition in (A.25), the function ϕ_r is Lipschitz with parameter at most $L = 2r = 2\sqrt{C_3} \leq 2C_3$. By the standard symmetrization argument (see Lemma 2.3.2 in Van Der Vaart et al. (1996)), we have

$$\begin{aligned}
\mathbb{E}[Z(t)] &\leq 2\mathbb{E}_{\boldsymbol{X},\sigma_i}\left[\sup_{\Delta\in\mathbb{S}_*(t)\cap\mathbb{S}_F(1)}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g_\Delta(\boldsymbol{X}_i)\right|\right] \\
&\leq 2\mathbb{E}_{\boldsymbol{X},\sigma_i}\left[\sup_{\Delta\in\mathbb{S}_*(t)\cap\mathbb{S}_F(1)}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i \phi_r\left(\langle \boldsymbol{X}_i, \Delta\rangle \mathbb{I}\{|\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle| \leq R\}\right)\right|\right] \\
&\leq 8C_3\mathbb{E}_{\boldsymbol{X},\sigma_i}\left[\sup_{\Delta\in\mathbb{S}_*(t)\cap\mathbb{S}_F(1)}\left|\left\langle \frac{1}{n}\sum_{i=1}^n \sigma_i \boldsymbol{X}_i \mathbb{I}\{|\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle| \leq R\}, \Delta\right\rangle\right|\right] \\
&\leq 8C_3 t E_{\boldsymbol{X},\sigma_i}\left[\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i \boldsymbol{X}_i \mathbb{I}\{|\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle| \leq R\}\right\|_{op}\right],
\end{aligned}$$

where $\sigma_i, i = 1, \cdots, n$ are i.i.d. Rademacher variables and the penultimate inequality follows from the contraction inequality for the Rademacher complexity since ϕ_r is Lipschitz continuous (see, for example, Theorem 2.2 in Koltchinskii (2011a)).

The sub-Gaussian property is preserved under conditioning and scaling. If $u^\top \boldsymbol{X}_i v$ is sub-Gaussian, then $u^\top \left(\boldsymbol{X}_i \mathbb{I}\{|\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle| \leq R\}\right) v$, being a scaled (possibly zeroed) version of $u^\top \boldsymbol{X}_i v$, remains sub-Gaussian. Therefore, by (A.21) in Lemma 6, we conclude

$$\mathbb{E}[Z(t)] \leq C' C_3 \kappa_{\boldsymbol{X}} t \sqrt{\frac{d_1 + d_2}{n}}, \tag{A.29}$$

where C' is some positive constant. Finally, combining (A.29) with (A.28) yields (A.27). Together, equations (A.26) and (A.27) imply that $P(\mathcal{B}(t)) \leq c_1 \exp(-c_2 n - c_2' t(d_1 + d_2))$. This result implies the claim of Proposition 1 for $n \gg d_1 + d_2$. □

Proof of Theorem 2. From (12) in Proposition 1, we have

$$\mathcal{E}_n(\Delta) \geq \tilde{\alpha}\|\Delta\|_F^2 - \tilde{\tau}\sqrt{\frac{d_1 + d_2}{n}}\|\Delta\|_*\|\Delta\|_F, \qquad \forall \|\Delta\|_F \leq 1,$$

with probability at least $1 - c_1 \exp(-c_2 n)$. Note that

$$\tilde{\tau}\sqrt{\frac{d_1 + d_2}{n}}\|\Delta\|_*\|\Delta\|_F \leq \frac{\tilde{\alpha}}{2}\|\Delta\|_F^2 + \frac{\tilde{\tau}^2}{2\tilde{\alpha}}\frac{d_1 + d_2}{n}\|\Delta\|_*^2,$$

it follows that

$$\mathcal{E}_n(\Delta) \geq \frac{\tilde{\alpha}}{2}\|\Delta\|_F^2 - \frac{\tilde{\tau}^2}{2\tilde{\alpha}}\frac{d_1 + d_2}{n}\|\Delta\|_*^2,$$

which verified the RSC condition (2). Therefore, Lemma 5 combined with the result of Theorem 1 directly gives the result. □

Proof of Corollary 1. Recall the definition of γ, λ, η and that (13) holds for all such α, λ and η. Minimizing the right-hand side of (14), it follows that

$$\|\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F \leq C(\tilde{\alpha} - \mu)^{-1} \inf_{\gamma\in[0,1]}\left\{\gamma c\sqrt{\frac{s\log(d_1 d_2)}{n}} + (1-\gamma)c'\sqrt{\frac{r(d_1 + d_2)}{n}}\right\}, \tag{A.30}$$

with probability approaching one. The minimum is attained when γ is chosen so that the two terms inside the braces are equal. Moreover, the minimum value is clearly bounded above by

$$C(\tilde{\alpha} - \mu)^{-1} \min\left\{\sqrt{\frac{s\log(d_1 d_2)}{n}}, \sqrt{\frac{r(d_1 + d_2)}{n}}\right\}.$$

Denote the minimizer in (A.30) by $\tilde{\gamma}$. Then, we can set $\lambda = \tilde{\gamma} c\sqrt{\frac{s\log(d_1 d_2)}{n}}$ and $\eta = (1-\tilde{\gamma})c'\sqrt{\frac{r(d_1+d_2)}{n}}$. This completes the proof. □

A.3 Proof of Theorem 3

Before proving Theorem 3, we first define

$$\Phi(\boldsymbol{\Theta}) := \mathcal{L}_n(\boldsymbol{\Theta}) + \lambda\|\boldsymbol{\Theta}\|_1 + \mathcal{R}_\eta(\boldsymbol{\Theta}).$$

In addition, we need to prove the following technical lemmas.

Lemma 7 *Under the conditions of Theorem 3, suppose there exists a pair (ω, T) such that*

$$\Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}}) \le \Lambda, \qquad \forall t \ge T, \tag{A.31}$$

Then for any interaction $t \ge T$, it holds that

$$\|\boldsymbol{\Theta}^t - \hat{\boldsymbol{\Theta}}\|_* \le \left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)\left(2\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F + \|\boldsymbol{\Theta}^t - \hat{\boldsymbol{\Theta}}\|_F\right) + 2\min\left\{\frac{\Lambda}{\eta}, R\right\}.$$

Proof of Lemma 7. We first show that if $\lambda \ge 2(1-\gamma)\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_\infty$ and $\eta \ge 4\gamma\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_{op}$, then for any $\boldsymbol{\Theta} \in \Omega$ satisfying

$$\Phi(\boldsymbol{\Theta}) - \Phi(\boldsymbol{\Theta}^*) \le \Lambda, \tag{A.32}$$

we have

$$\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_* \le \left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_F + 2\min\left\{\frac{\Lambda}{\eta}, R\right\}. \tag{A.33}$$

Defining the error matrix $\Delta := \boldsymbol{\Theta} - \boldsymbol{\Theta}^*$, (A.32) implies

$$\mathcal{L}_n(\boldsymbol{\Theta}^* + \Delta) + \lambda\|\boldsymbol{\Theta}^* + \Delta\|_1 + \mathcal{R}_\eta(\boldsymbol{\Theta}^* + \Delta) \le \mathcal{L}_n(\boldsymbol{\Theta}^*) + \lambda\|\boldsymbol{\Theta}^*\|_1 + \mathcal{R}_\eta(\boldsymbol{\Theta}^*) + \Lambda.$$

By subtracting $\langle\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*), \Delta\rangle$ from both sides gives

$$\begin{aligned}&\mathcal{T}(\boldsymbol{\Theta}^* + \Delta, \boldsymbol{\Theta}^*) + \mathcal{R}_\eta(\boldsymbol{\Theta}^* + \Delta) - \mathcal{R}_\eta(\boldsymbol{\Theta}^*)\\ &\le -\lambda\|\boldsymbol{\Theta}^* + \Delta\| + \lambda\|\boldsymbol{\Theta}^*\|_1 - \langle\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*), \Delta\rangle + \Lambda.\end{aligned}$$

Then the RSC condition (19) together with hölder's inequality implies

$$\begin{aligned}&\alpha_1\|\Delta\|_F^2 - \tau_1\frac{d_1+d_2}{n}\|\Delta\|_*^2 + \mathcal{R}_\eta(\boldsymbol{\Theta}^* + \Delta) - \mathcal{R}_\eta(\boldsymbol{\Theta}^*)\\ &\le -\lambda(\|\boldsymbol{\Theta}^* + \Delta\|_1 - \|\boldsymbol{\Theta}^*\|_1) - \langle\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*), \Delta\rangle + \omega\\ &\le -\lambda(\|S_{\boldsymbol{\Theta}^*}^\perp \circ \Delta\|_1 - \|S_{\boldsymbol{\Theta}^*} \circ \Delta\|_1) + (1-\gamma)\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_\infty\|\Delta\|_1\\ &\quad + \gamma\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_{op}\|\Delta\|_* + \Lambda.\end{aligned}$$

By the choice of $\lambda \ge 2(1-\gamma)\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_\infty$ and $\eta \ge 4\gamma\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_{op}$, we have

$$\begin{aligned}&\alpha_1\|\Delta\|_F^2 - \tau_1\frac{d_1+d_2}{n}\|\Delta\|_*^2 + \mathcal{R}_\eta(\boldsymbol{\Theta}^* + \Delta) - \mathcal{R}_\eta(\boldsymbol{\Theta}^*)\\ \le&\frac{3\lambda}{2}\|S_{\boldsymbol{\Theta}^*} \circ \Delta\|_1 - \frac{\lambda}{2}\|S_{\boldsymbol{\Theta}^*}^\perp \circ \Delta\|_1 + \frac{\eta}{4}\|\Delta\|_* + \Lambda\\ \le&\frac{3\lambda}{2}\|S_{\boldsymbol{\Theta}^*} \circ \Delta\|_1 - \frac{\lambda}{2}\|S_{\boldsymbol{\Theta}^*}^\perp \circ \Delta\|_1 + \frac{\eta}{4}\|\Delta\|_* + \Lambda.\end{aligned}$$

Note that $\alpha_1 > 0$ and $\|\varDelta\|_* \leq \|\boldsymbol{\Theta}\|_* + \|\boldsymbol{\Theta}^*\|_* \leq 2R/\omega$ and combining the choice $\eta \geq \frac{8\tau_1 R(d_1+d_2)}{n\omega}$, we have

$$\mathcal{R}_\eta(\boldsymbol{\Theta}^* + \varDelta) - \mathcal{R}_\eta(\boldsymbol{\Theta}^*) \leq \frac{3\lambda}{2}\|S_{\boldsymbol{\Theta}^*} \circ \varDelta\|_1 - \frac{\lambda}{2}\|S^{\perp}_{\boldsymbol{\Theta}^*} \circ \varDelta\|_1 + \frac{\eta}{2}\|\varDelta\|_* + \varLambda. \tag{A.34}$$

On the other hand, according to (A.7) in Lemma 4, we have

$$\mathcal{R}_\eta(\boldsymbol{\Theta}^*) - \mathcal{R}_\eta(\boldsymbol{\Theta}^* + \varDelta) \leq \eta(\|\varDelta'\|_* - \|\varDelta''\|_*), \tag{A.35}$$

where $\varDelta'$ and $\varDelta''$ are defined in Lemma 4 with $\varDelta = \varDelta' + \varDelta''$. Combining (A.34) and (A.35) and rearranging the terms on both sides then gives

$$\frac{\lambda}{2}\|S^{\perp}_{\boldsymbol{\Theta}^*} \circ \varDelta\|_1 + \frac{\eta}{2}\|\varDelta''\|_* \leq \frac{3\lambda}{2}\|S_{\boldsymbol{\Theta}^*} \circ \varDelta\|_1 + \frac{3\eta}{2}\|\varDelta'\|_* + \varLambda,$$

and consequently, using the trivial bound $\|\varDelta\|_* \leq 2R$ we have

$$\begin{aligned}\|\varDelta\|_* &\leq \frac{3\lambda}{\eta}\|S_{\boldsymbol{\Theta}^*} \circ \varDelta\|_1 + 4\|\varDelta'\|_* + \frac{2\varLambda}{\eta} \\ &\leq \left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)\|\varDelta\|_F + 2\min\left\{\frac{\varLambda}{\eta}, R\right\},\end{aligned}$$

which verified (A.33). Note that by the optimality of $\hat{\Theta}$, we have $\varPhi(\hat{\boldsymbol{\Theta}}) - \varPhi(\boldsymbol{\Theta}^*) \leq 0$, and by the assumption (A.31), we also have

$$\varPhi(\boldsymbol{\Theta}^t) \leq \varPhi(\hat{\boldsymbol{\Theta}}) + \varLambda \leq \varPhi(\boldsymbol{\Theta}^*) + \varLambda,$$

hence, (A.32) is satisfied and from (A.33) we have

$$\begin{aligned}\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_* &\leq \left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F \\ \|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^*\|_* &\leq \left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)\|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^*\|_F + 2\min\left\{\frac{\varLambda}{\eta}, R\right\}.\end{aligned}$$

By the triangle inequality, we then have

$$\begin{aligned}\|\boldsymbol{\Theta}^t - \hat{\boldsymbol{\Theta}}\|_* &\leq \|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_* + \|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^*\|_* \\ &\leq \left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)\left(\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F + \|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^*\|_F\right) + 2\min\left\{\frac{\varLambda}{\eta}, R\right\} \\ &\leq \left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)\left(2\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F + \|\boldsymbol{\Theta}^t - \hat{\boldsymbol{\Theta}}\|_F\right) + 2\min\left\{\frac{\varLambda}{\eta}, R\right\}.\end{aligned}$$

This completes the proof of the lemma. □

Lemma 8 *Under the conditions of Theorem 3, suppose that there exists a pair* (ω, T) *such that* (A.31) *holds. Define*

$$\bar{\epsilon} := 2\left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)\epsilon_{stat}, \quad \textit{and} \quad \epsilon = 2\min\left\{\frac{\varLambda}{\eta}, R\right\}.$$

Then for any $t \geq T$, *we have*

$$\bar{\mathcal{T}}(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^t) \geq -2\tau_1\frac{d_1+d_2}{n}(\bar{\epsilon} + \epsilon)^2, \tag{A.36}$$

$$\varPhi(\boldsymbol{\Theta}^t) - \varPhi(\hat{\boldsymbol{\Theta}}) \geq \frac{2\alpha_1 - \mu}{4}\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^t\|_F^2 - 2\tau_1\frac{d_1+d_2}{n}(\bar{\epsilon} + \epsilon)^2. \tag{A.37}$$

Proof of Lemma 8. Using the notations above, from Lemma 7 we have

$$\begin{aligned}\|\boldsymbol{\Theta}^t - \hat{\boldsymbol{\Theta}}\|_*^2 &\leq \left[\left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)\|\boldsymbol{\Theta}^t - \hat{\boldsymbol{\Theta}}\|_F + (\bar{\epsilon} + \epsilon)\right]^2 \\ &\leq 2\left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)^2 \|\boldsymbol{\Theta}^t - \hat{\boldsymbol{\Theta}}\|_F^2 + 2(\bar{\epsilon} + \epsilon)^2. \end{aligned} \tag{A.38}$$

By the RSC condition (19) and (A.6) in Lemma 2, we have that

$$\bar{\mathcal{T}}(\boldsymbol{\Theta}^t, \hat{\boldsymbol{\Theta}}) \geq \left(\alpha_1 - \frac{\mu}{2}\right)\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^t\|_F^2 - \tau_1 \frac{d_1 + d_2}{n}\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_*^2. \tag{A.39}$$

Recall that $\Phi(\boldsymbol{\Theta}) = \bar{\mathcal{L}}_n(\boldsymbol{\Theta}) + \lambda\|\boldsymbol{\Theta}\|_* + \eta\|\boldsymbol{\Theta}\|_1$ and by the convexity of $\|\cdot\|_1$ and $\|\cdot\|_*$, we have

$$\lambda(\|\boldsymbol{\Theta}^t\|_1 - \|\hat{\boldsymbol{\Theta}}\|_1 - \langle\partial\|\hat{\boldsymbol{\Theta}}\|_1, \boldsymbol{\Theta}^t - \hat{\boldsymbol{\Theta}}\rangle) \geq 0, \quad \text{and} \tag{A.40}$$

$$\eta(\|\boldsymbol{\Theta}^t\|_* - \|\hat{\boldsymbol{\Theta}}\|_* - \langle\partial\|\hat{\boldsymbol{\Theta}}\|_*, \boldsymbol{\Theta}^t - \hat{\boldsymbol{\Theta}}\rangle) \geq 0. \tag{A.41}$$

Furthermore, the first-order optimality condition for $\hat{\boldsymbol{\Theta}}$ gives

$$\langle\nabla\Phi(\hat{\boldsymbol{\Theta}}), \boldsymbol{\Theta}^t - \hat{\boldsymbol{\Theta}}\rangle \geq 0. \tag{A.42}$$

Summing (A.39)-(A.42) gives

$$\Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}}) \geq \left(\alpha_1 - \frac{\mu}{2}\right)\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^t\|_F^2 - \tau_1\frac{d_1 + d_2}{n}\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_*^2.$$

Applying Lemma 7 and (A.38) to bound the term $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^t\|_*^2$ and utilizing the assumption

$$\frac{2\alpha_1 - \mu}{8} \geq \left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)^2 \frac{\tau_1(d_1 + d_2)}{n},$$

we then have

$$\Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}}) \geq \frac{2\alpha_1 - \mu}{4}\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^t\|_F^2 - 2\tau_1\frac{d_1 + d_2}{n}(\bar{\epsilon} + \epsilon)^2,$$

which establishes (A.37). Moreover, applying(A.38) into (A.39) directly and switching $\boldsymbol{\Theta}^t$ and $\hat{\boldsymbol{\Theta}}$ gives

$$\begin{aligned}\bar{\mathcal{T}}(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^t) \geq& \left(\alpha_1 - \frac{\mu}{2}\right)\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^t\|_F^2 \\ &- \tau_1\frac{d_1 + d_2}{n}\left(2\left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)^2\|\boldsymbol{\Theta}^t - \hat{\boldsymbol{\Theta}}\|_F^2 + 2(\bar{\epsilon} + \epsilon)^2\right) \\ \geq& - 2\tau_1\frac{d_1 + d_2}{n}(\bar{\epsilon} + \epsilon)^2, \end{aligned}$$

which establishes (A.36). □

Lemma 9 *Under the conditions of Lemma 8, and suppose in addition that* $\alpha_2 \leq \xi/2$ *and* $\frac{(2\alpha_1 - \mu)^2}{8\xi} \geq \frac{32v}{n}$. *Then for any* $t \geq T$, *we have*

$$\Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}}) \leq \kappa^{t-T}\left(\Phi(\boldsymbol{\Theta}^T) - \Phi(\hat{\boldsymbol{\Theta}})\right) + \frac{2\delta(\bar{\epsilon}^2 + \epsilon^2)}{1 - \kappa},$$

where κ *and* δ *are defined in* (21) *and* (22).

Proof of Lemma 9. We first define

$$\Psi_t(\boldsymbol{\Theta}) := \bar{\mathcal{L}}_n(\boldsymbol{\Theta}^t) + \langle \nabla \bar{\mathcal{L}}_n(\boldsymbol{\Theta}^t), \boldsymbol{\Theta} - \boldsymbol{\Theta}^t \rangle + \frac{\xi}{2}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^t\|_F^2 + \lambda\|\boldsymbol{\Theta}\|_1 + \eta\|\boldsymbol{\Theta}\|_*,$$

which is equivalent to the objective function (16) minimized over the constraint set Ω at iteration t. For any $c \in [0,1]$, let $\boldsymbol{\Theta}_c := c\hat{\boldsymbol{\Theta}} + (1-c)\boldsymbol{\Theta}^t$, due to the convexity of the constraint set Ω, we have $\boldsymbol{\Theta}_c \in \Omega$. Therefore, by the optimality of $\boldsymbol{\Theta}^{t+1}$, we have

$$\begin{aligned}\Psi_t(\boldsymbol{\Theta}^{t+1}) \le \Psi_t(\boldsymbol{\Theta}_c) =& \bar{\mathcal{L}}_n(\boldsymbol{\Theta}^t) + \langle \nabla \bar{\mathcal{L}}_n(\boldsymbol{\Theta}^t), c\hat{\boldsymbol{\Theta}} - c\boldsymbol{\Theta}^t \rangle + \frac{c^2\xi}{2}\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^t\|_F^2 \\ &+ \lambda\|\boldsymbol{\Theta}_c\|_1 + \eta\|\boldsymbol{\Theta}_c\|_*.\end{aligned}$$

Then by the convexity of norms and (A.36), we have

$$\begin{aligned}\Psi_t(\boldsymbol{\Theta}^{t+1}) \le& (1-c)\bar{\mathcal{L}}_n(\boldsymbol{\Theta}^t) + c\bar{\mathcal{L}}_n(\hat{\boldsymbol{\Theta}}) + 2c\tau_1\frac{d_1+d_2}{n}(\bar{\epsilon}+\epsilon)^2 + \frac{c^2\xi}{2}\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^t\|_F^2 \\ &+ c\lambda\|\hat{\boldsymbol{\Theta}}\|_1 + (1-c)\lambda\|\boldsymbol{\Theta}^t\|_1 + c\lambda\|\hat{\boldsymbol{\Theta}}\|_* + (1-c)\lambda\|\boldsymbol{\Theta}^t\|_* \\ \le& (1-c)\Phi(\boldsymbol{\Theta}^t) + c\Phi(\hat{\boldsymbol{\Theta}}) + 2c\tau_1\frac{d_1+d_2}{n}(\bar{\epsilon}+\epsilon)^2 + \frac{c^2\xi}{2}\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^t\|_F^2 \\ \le& \Phi(\boldsymbol{\Theta}^t) - c\left(\Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}})\right) + 2\tau_1\frac{d_1+d_2}{n}(\bar{\epsilon}+\epsilon)^2 + \frac{c^2\xi}{2}\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^t\|_F^2. \quad \text{(A.43)}\end{aligned}$$

On the other hand, by the concaveness of $\mathcal{Q}_\eta$, we have $\mathcal{Q}_\eta(\boldsymbol{\Theta}) - \mathcal{Q}_\eta(\boldsymbol{\Theta}') - \langle \nabla\mathcal{Q}_\eta(\boldsymbol{\Theta}'), \boldsymbol{\Theta} - \boldsymbol{\Theta}'\rangle \le 0$. By the RSM condition (20), we also have

$$\begin{aligned}\bar{\mathcal{T}}(\boldsymbol{\Theta}^{t+1}, \boldsymbol{\Theta}^t) &\le \alpha_2\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^t\|_F^2 + \tau_2\frac{d_1+d_2}{n}\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^t\|_*^2 \\ &\le \frac{\xi}{2}\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^t\|_F^2 + \tau_2\frac{d_1+d_2}{n}\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^t\|_*^2,\end{aligned}$$

where the las inequality follows from the assumption $\alpha_2 \le \xi/2$. Adding $\lambda\|\boldsymbol{\Theta}^{t+1}\|_1 + \eta\|\boldsymbol{\Theta}^{t+1}\|_*$ on both sides yields

$$\begin{aligned}\Phi(\boldsymbol{\Theta}^{t+1}) \le& \bar{\mathcal{L}}_n(\boldsymbol{\Theta}^{t+1}) + \langle \nabla\bar{\mathcal{L}}_n(\boldsymbol{\Theta}^t), \boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^t\rangle + \frac{\xi}{2}\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^t\|_F^2 \\ &+ \tau_2\frac{d_1+d_2}{n}\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^t\|_*^2 + \lambda\|\boldsymbol{\Theta}^{t+1}\|_1 + \eta\|\boldsymbol{\Theta}^{t+1}\|_* \\ =& \Psi_t(\boldsymbol{\Theta}^{t+1}) + \tau_2\frac{d_1+d_2}{n}\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^t\|_*^2. \quad \text{(A.44)}\end{aligned}$$

Combining (A.43) with (A.44) gives

$$\begin{aligned}\Phi(\boldsymbol{\Theta}^{t+1}) \le& \Phi(\boldsymbol{\Theta}^t) - c\left(\Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}})\right) + 2\tau_1\frac{d_1+d_2}{n}(\bar{\epsilon}+\epsilon)^2 \\ &+ \frac{c^2\xi}{2}\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^t\|_F^2 + \tau_2\frac{d_1+d_2}{n}\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^t\|_*^2. \quad \text{(A.45)}\end{aligned}$$

Define $\Delta^t := \boldsymbol{\Theta}^t - \hat{\boldsymbol{\Theta}}$, by the triangle inequality we have

$$\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^t\|_*^2 \le (\|\Delta^{t+1}\|_* + \|\Delta^t\|_*)^2 \le 2\|\Delta^{t+1}\|_*^2 + 2\|\Delta^t\|_*^2.$$

Then combined with (A.45), we have

$$\begin{aligned}\Phi(\boldsymbol{\Theta}^{t+1}) \le& \Phi(\boldsymbol{\Theta}^t) - c\left(\Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}})\right) + 2\tau_1\frac{d_1+d_2}{n}(\bar{\epsilon}+\epsilon)^2 \\ &+ \frac{c^2\xi}{2}\|\Delta^t\|_F^2 + \tau_2\frac{d_1+d_2}{n}\left(2\|\Delta^{t+1}\|_*^2 + 2\|\Delta^t\|_*^2\right).\end{aligned}$$

Then applying (A.38) to bound the nuclear norms, we have

$$\begin{aligned}\Phi(\boldsymbol{\Theta}^{t+1}) \leq& \Phi(\boldsymbol{\Theta}^t) - c\left(\Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}})\right) + 10\tau_2 \frac{d_1+d_2}{n}(\bar{\epsilon}+\epsilon)^2 \\ &+ \left(\frac{c^2\xi}{2} + 4\tau_2\frac{d_1+d_2}{n}\left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)^2\right)\|\Delta^t\|_F^2 \\ &+ 4\tau_2\frac{d_1+d_2}{n}\left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)^2\|\Delta^{t+1}\|_F^2.\end{aligned}$$

For the sake of simplification of notation, we define $\zeta := \tau_2\frac{d_1+d_2}{n}(\bar{\epsilon}+\epsilon)^2$, introduce the shorthand $\boldsymbol{\varepsilon}_t = \Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}})$ and $\upsilon = \tau_2\frac{d_1+d_2}{n}\left(\frac{3\lambda\sqrt{s}}{\eta} + 4\sqrt{2r}\right)^2$. Applying (A.37) and subtracting $\Phi(\hat{\boldsymbol{\Theta}})$ from both sides of the above inequality, we have

$$\boldsymbol{\varepsilon}_{t+1} \leq (1-c)\boldsymbol{\varepsilon}_t + \frac{c^2\xi + 8\upsilon}{\alpha_1 - \mu/2}(\boldsymbol{\varepsilon}_t + 2\zeta) + \frac{8\upsilon}{\alpha_1-\mu/2}(\boldsymbol{\varepsilon}_{t+1} + 2\zeta) + 10\zeta.$$

By the choice of $c = \frac{2\alpha_1-\mu}{4\xi} \in (0,1)$, we have

$$\begin{aligned}\left(1 - \frac{8\upsilon}{\alpha_1-\mu/2}\right)\boldsymbol{\varepsilon}_{t+1} \leq& \left(1 - \frac{2\alpha_1-\mu}{8\xi} + \frac{8\upsilon}{\alpha_1-\mu/2}\right)\boldsymbol{\varepsilon}_t \\ &+ 2\left(\frac{2\alpha_1-\mu}{8\xi} + \frac{16\upsilon}{\alpha_1-\mu/2} + 5\right)\zeta.\end{aligned}$$

According to the definition of κ and δ in (21) and (22), the above inequality can be rewritten as

$$\boldsymbol{\varepsilon}_{t+1} \leq \kappa\boldsymbol{\varepsilon}_t + \delta(\bar{\epsilon}+\epsilon)^2.$$

Note that by the assumption $\xi \geq \max\{\frac{2\alpha_1-\mu}{4}, 2\alpha_2\}$ ensures $\kappa \in (0,1)$. Finally, by iterating the procedure, we have

$$\begin{aligned}\boldsymbol{\varepsilon}_t &\leq \kappa^{t-T}\boldsymbol{\varepsilon}_T + \delta(\bar{\epsilon}+\epsilon)^2(1+\kappa+\kappa^2+\cdots+\kappa^{t-T-1}) \\ &\leq \kappa^{t-T}\boldsymbol{\varepsilon}_T + \frac{\delta(\bar{\epsilon}+\epsilon)^2}{1-\kappa} \leq \kappa^{t-T}\boldsymbol{\varepsilon}_T + \frac{2\delta(\bar{\epsilon}^2+\epsilon^2)}{1-\kappa},\end{aligned}$$

which completes the proof of this lemma. □

Proof of Theorem 3. Follows the lines in Agarwal et al. (2012a); Loh and Wainwright (2015), we first prove that

$$\Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}}) \leq \tau^2, \qquad \forall t \geq T^*(\tau).$$

We divide iterations $t \geq 0$ into a series of epochs $[T_j, T_{j+1})$, and define the associated tolerances $\Lambda_0 > \Lambda_1 > \cdots$ such that

$$\Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}}) \leq \Lambda_j, \qquad \forall t \geq T_j. \tag{A.46}$$

For the first iteration, define $\Lambda_0 = \Phi(\boldsymbol{\Theta}^0) - \Phi(\hat{\boldsymbol{\Theta}})$ and recall $\epsilon(\Lambda_j) = 2\min\left\{\frac{\Lambda_j}{\eta}, R\right\}$, we define $\epsilon(\Lambda_0) = R$ and $T_0 = 0$, applying Lemma 9 yields

$$\Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}}) \leq \kappa^t\left(\Phi(\boldsymbol{\Theta}^0) - \Phi(\hat{\boldsymbol{\Theta}})\right) + \frac{2\delta(\bar{\epsilon}^2+4R^2)}{1-\kappa}, \qquad \forall t \geq 0.$$

Let $\Lambda_1 := \frac{4\delta(\bar{\epsilon}^2+4R^2)}{1-\kappa}$, for $T_1 := \left\lceil \frac{\log(2\Lambda_0/\Lambda_1)}{\log(1/\kappa)} \right\rceil$, since $\kappa \in (0,1)$, we have

$$\Phi(\boldsymbol{\Theta}^t) - \Phi(\hat{\boldsymbol{\Theta}}) \leq \frac{\Lambda_1}{2} + \frac{2\delta(\bar{\epsilon}^2+4R^2)}{1-\kappa} \leq \Lambda_1 \leq \frac{8\delta}{1-\kappa}\max\{\bar{\epsilon}^2, 4R^2\}, \qquad \forall t \geq T_1.$$

For $j \geq 1$, now we define

$$\varLambda_j := \frac{2\delta(\bar{\epsilon}^2 + \epsilon^2(\varLambda_j))}{1-\kappa}, \qquad \text{and} \tag{A.47}$$

$$T_{j+1} := \left\lceil \frac{\log(2\varLambda_j/\varLambda_{j+1})}{\log(1/\kappa)} \right\rceil + T_j. \tag{A.48}$$

Again, by Lemma 9, we have

$$\varPhi(\boldsymbol{\Theta}^t) - \varPhi(\hat{\boldsymbol{\Theta}}) \leq \kappa^{t-T_j}\left(\varPhi(\boldsymbol{\Theta}^{T_j}) - \varPhi(\hat{\boldsymbol{\Theta}})\right) + \frac{2\delta(\bar{\epsilon}^2 + 4R^2)}{1-\kappa}, \qquad \forall t \geq T_j.$$

Then by the choice of $\varLambda_j$ and T_{j+1}, we have

$$\varPhi(\boldsymbol{\Theta}^t) - \varPhi(\hat{\boldsymbol{\Theta}}) \leq \varLambda_{j+1} \leq \frac{8\delta}{1-\kappa}\max\{\bar{\epsilon}^2, 4R^2\}, \qquad \forall t \geq T_{j+1}.$$

According to Agarwal et al. (2012a), for all $j = 1, 2, \cdots$, the recursion above shows that

$$\varLambda_{j+1} \leq \frac{\varLambda_j}{4^{2^{j-1}}}, \quad \text{and} \quad \frac{\varLambda_{j+1}}{\eta} \leq \frac{R}{4^{2^j}}. \tag{A.49}$$

These statements are used to upper bound the smallest j such that $\varLambda_k \leq \tau^2$. If we are in the first epoch, the claim of the theorem is straightforward from (A.47). Otherwise, by the recursion (A.49) and iteration (A.48), we see that it is sufficient to ensure the error drops below τ^2 after at most

$$k_\tau \geq \log\left(\log\left(\frac{R\eta}{\tau^2}\right)/\log 4\right)/\log 2 + 1 = \log_2\log_2\left(\frac{R\eta}{\tau}\right)$$

epochs. Combining the above bound on k_τ with (A.48), we have $\varPhi(\boldsymbol{\Theta}^t) - \varPhi(\hat{\boldsymbol{\Theta}}) \leq \tau^2$ to hold for all iterations

$$t \geq k_\tau\left(1 + \frac{\log 2}{\log(1/\kappa)}\right) + \frac{\log\left(\frac{\varLambda_0}{\tau^2}\right)}{\log(1/\kappa)} =: T^*(\tau).$$

Therefore, (A.46) is proved, then according to (A.37) in Lemma 8, for $t \geq T^*(\tau)$, we have

$$\begin{aligned}
\frac{2\alpha_1 - \mu}{4}\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^t\|_F^2 &\leq \varPhi(\boldsymbol{\Theta}^t) - \varPhi(\hat{\boldsymbol{\Theta}}) + 2\tau_1\frac{d_1+d_2}{n}(\bar{\epsilon}+\epsilon)^2 \\
&\leq \tau^2 + 2\tau_1\frac{d_1+d_2}{n}(\bar{\epsilon} + \frac{2\tau^2}{\eta})^2 \\
&\leq \tau^2 + 8\tau_1\frac{d_1+d_2}{n}\frac{\tau^4}{\eta^2} + 4\tau_1\frac{d_1+d_2}{n}\bar{\epsilon}^2,
\end{aligned}$$

where we let $\epsilon = \frac{2\tau^2}{\eta}$. Finally, some algebra with the choice of η in (24) proves the theorem. □

B ADMM for solving optimization problem (17).

The objective function in Eq.(17) is jointly convex with respect to all variables, and the variables are separable. Therefore, it can be further solved by using the alternate direction multiplier method (ADMM). By introducing the slack variables $\boldsymbol{Q} = \boldsymbol{\Theta}$ (for $\|\cdot\|_1$), $\boldsymbol{R} = \boldsymbol{\Theta}$ (for $\|\cdot\|_*$), Eq.(17) can be formulated as

$$\begin{aligned}
\mathop{\arg\min}_{\boldsymbol{\Theta}\in\mathbb{R}^{d_1\times d_2}} &\left\{\frac{1}{2}\|\boldsymbol{\Theta} - \boldsymbol{V}\|_F^2 + \frac{\lambda}{\xi}\|\boldsymbol{Q}\|_1 + \frac{\eta}{\xi}\|\boldsymbol{R}\|_*\right\}, \\
&\text{s.t. } \boldsymbol{\Theta} - \boldsymbol{Q} = \boldsymbol{0}, \\
&\quad\ \ \boldsymbol{\Theta} - \boldsymbol{R} = \boldsymbol{0},
\end{aligned} \tag{B.1}$$

where we denote $\boldsymbol{V} = \boldsymbol{\Theta}^t - \frac{\nabla \bar{\mathcal{L}}_n(\boldsymbol{\Theta}^t)}{\xi}$, which can be seen as a known matrix at step t. Then we get the augmented Lagrangian equation as

$$\begin{aligned}\mathcal{L}_\rho(\boldsymbol{\Theta}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{U}^1, \boldsymbol{U}^2) = &\ \frac{1}{2}\|\boldsymbol{\Theta} - \boldsymbol{V}\|_F^2 + \frac{\lambda}{\xi}\|\boldsymbol{Q}\|_1 + \langle \boldsymbol{U}^1, \boldsymbol{\Theta} - \boldsymbol{Q}\rangle \\ &+ \frac{\rho}{2}\|\boldsymbol{\Theta} - \boldsymbol{Q}\|_F^2 + \frac{\eta}{\xi}\|\boldsymbol{R}\|_* + \langle \boldsymbol{U}^2, \boldsymbol{\Theta} - \boldsymbol{R}\rangle + \frac{\rho}{2}\|\boldsymbol{\Theta} - \boldsymbol{R}\|_F^2,\end{aligned} \tag{B.2}$$

where $\boldsymbol{U}^1$ and $\boldsymbol{U}^2$ are augmented lagrangian multipliers. The variables $\boldsymbol{\Theta}$, $\boldsymbol{Q}$ and $\boldsymbol{R}$ are separable in Eq.(B.2). This means that the optimization can be split into sub-problems that can be solved in parallel corresponding to variables $\boldsymbol{\Theta}$, $\boldsymbol{Q}$, and $\boldsymbol{R}$.

Update $\boldsymbol{\Theta}$

When updating $\boldsymbol{\Theta}$, the other variables can be seen as constants. Therefore, $\boldsymbol{\Theta}$ can be updated by

$$\begin{aligned}\boldsymbol{\Theta}^{(t+1)} &= \arg\min_{\boldsymbol{\Theta}} \mathcal{L}_\rho(\boldsymbol{V}^{(t)}, \boldsymbol{Q}^{(t)}, \boldsymbol{R}^{(t)}, \boldsymbol{U}^{1(t)}, \boldsymbol{U}^{2(t)}) \\ &= \arg\min_{\boldsymbol{\Theta}} \frac{1}{2}\|\boldsymbol{\Theta} - \boldsymbol{V}\|_F^2 + \langle \boldsymbol{U}^{1(t)}, \boldsymbol{\Theta} - \boldsymbol{Q}^{(t)}\rangle + \frac{\rho}{2}\|\boldsymbol{\Theta} - \boldsymbol{Q}^{(t)}\|_F^2 \\ &\quad + \langle \boldsymbol{U}^{2(t)}, \boldsymbol{\Theta} - \boldsymbol{R}^{(t)}\rangle + \frac{\rho}{2}\|\boldsymbol{\Theta} - \boldsymbol{R}^{(t)}\|_F^2,\end{aligned} \tag{B.3}$$

where t is the number of iterations. Eq.(B.3) is the closed form and a closed solution can be obtained. Let its derivative be zero, we can get the following equations:

$$\begin{aligned}\mathbf{0} &= (\boldsymbol{\Theta} - \boldsymbol{V}^{(t)}) + \boldsymbol{U}^{1(t)} + \rho(\boldsymbol{\Theta} - \boldsymbol{Q}^{(t)}) + \boldsymbol{U}^{2(t)} + \rho(\boldsymbol{\Theta} - \boldsymbol{R}^{(t)}), \\ \boldsymbol{\Theta} &= (\boldsymbol{I} + 2\rho \boldsymbol{I})^{-1}(\boldsymbol{V}^{(t)} - \boldsymbol{U}^{1(t)} + \rho \boldsymbol{Q}^{(t)} - \boldsymbol{U}^{2(t)} + \rho \boldsymbol{R}^{(t)}).\end{aligned} \tag{B.4}$$

$\boldsymbol{I}$ is a $d_1 \times d_1$ identity matrix.

Update Q

According to the augmented Lagrangian in Eq.(B.2), $\boldsymbol{Q}$ can be updated by

$$\begin{aligned}\boldsymbol{Q}^{(t+1)} &= \arg\min_{\boldsymbol{Q}} \mathcal{L}_\rho(\boldsymbol{\Theta}^{(t+1)}, \boldsymbol{U}^{1(t)}) \\ &= \arg\min_{\boldsymbol{Q}} \left\{ \frac{\lambda}{\xi}\|\boldsymbol{Q}\|_1 + \langle \boldsymbol{U}^{1(t)}, \boldsymbol{\Theta}^{(t+1)} - \boldsymbol{Q}\rangle + \frac{\rho}{2}\|\boldsymbol{\Theta}^{(t+1)} - \boldsymbol{Q}\|_F^2 \right\},\end{aligned} \tag{B.5}$$

which is equivalent to the following problem:

$$\boldsymbol{Q}^{(t+1)} = \arg\min_{\boldsymbol{Q}} \left\{ \frac{1}{2}\|\boldsymbol{Q} - \boldsymbol{\Lambda}_1^{(t+1)}\|_F^2 + \frac{\lambda}{\rho\xi}\|\boldsymbol{Q}\|_1 \right\}, \tag{B.6}$$

where $\boldsymbol{\Lambda}_1^{(t+1)} = \boldsymbol{\Theta}^{(t+1)} + \frac{\boldsymbol{U}^{1(t)}}{\rho}$. Eq.(B.6) can be solved by the l_1 norm proximal operator. That is

$$\boldsymbol{Q}_{ij}^{(t+1)} = \mathcal{S}_{\frac{\lambda}{\rho\xi}}(\boldsymbol{\Lambda}_{1,ij}^{t+1}), \quad \text{for } 1 \le i \le d_1, 1 \le j \le d_2,$$

where the soft-thresholding operator $\mathcal{S}_\tau$ with threshold τ is defined as:

$$\mathcal{S}_\tau(x) = \text{sign}(x) \cdot \max(|x| - \tau, 0).$$

Update R

According to the augmented Lagrangian in Eq.(B.2), $\boldsymbol{R}$ can be updated by

$$\begin{aligned}\boldsymbol{R}^{(t+1)} &= \arg\min_{\boldsymbol{R}} \mathcal{L}_\rho(\boldsymbol{\Theta}^{(t+1)}, \boldsymbol{U}^{2(t)}) \\ &= \arg\min_{\boldsymbol{R}} \left\{ \frac{\eta}{\xi}\|\boldsymbol{R}\|_* + \langle \boldsymbol{U}^{2(t)}, \boldsymbol{\Theta}^{(t+1)} - \boldsymbol{R}\rangle + \frac{\rho}{2}\|\boldsymbol{\Theta}^{(t+1)} - \boldsymbol{R}\|_F^2 \right\},\end{aligned} \tag{B.7}$$

which is equivalent to the following problem:

$$\boldsymbol{R}^{(t+1)} = \arg\min_{\boldsymbol{R}} \left\{ \frac{1}{2}\|\boldsymbol{R} - \boldsymbol{\Lambda}_2^{(t+1)}\|_F^2 + \frac{\eta}{\rho\xi}\|\boldsymbol{R}\|_* \right\}, \tag{B.8}$$

where $\boldsymbol{\Lambda}_2^{(t+1)} = \boldsymbol{\Theta}^{(t+1)} + \frac{\boldsymbol{U}^{2(t)}}{\rho}$. The trace norm proximal operator can solve Eq.(B.8). we use the proximal operator of the trace norm. The steps are as follows:

1. Perform Singular Value Decomposition (SVD) on $\boldsymbol{\Lambda}_2^{(t+1)}$:

$$\boldsymbol{\Lambda}_2^{(t+1)} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T,$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, and $\boldsymbol{\Sigma}$ is a diagonal matrix of singular values σ_i.

2. Apply the soft-thresholding operator to the singular values in $\boldsymbol{\Sigma}$:

$$\mathcal{S}_\tau(\sigma_i) = \max(\sigma_i - \tau, 0),$$

where we take $\tau = \frac{\eta}{\rho\xi}$.

3. Reconstruct the matrix $\boldsymbol{R}^{(t+1)}$ using the thresholded singular values:

$$\boldsymbol{R}^{(t+1)} = \boldsymbol{U}\boldsymbol{\Sigma}_\tau\boldsymbol{V}^T,$$

where $\boldsymbol{\Sigma}_\tau$ is the diagonal matrix with elements $\mathcal{S}_\tau(\sigma_i)$.

Update $\boldsymbol{U}^{1(t+1)}$ and $\boldsymbol{U}^{2(t+1)}$

According to the standard ADMM, the updates of augmented lagrangian multipliers are as follows:

$$\begin{aligned} \boldsymbol{U}^{1(t+1)} &= \boldsymbol{U}^{1(t)} + \rho(\boldsymbol{\Theta}^{(t+1)} - \boldsymbol{Q}^{(t+1)})\ , \\ \boldsymbol{U}^{2(t+1)} &= \boldsymbol{U}^{2(t)} + \rho(\boldsymbol{\Theta}^{(t+1)} - \boldsymbol{R}^{(t+1)})\ . \end{aligned} \tag{B.9}$$

In the above algorithm, there are three tuning parameters ρ, λ, η. The parameter ρ affects mainly the practical speed of the algorithm but less on convergence rate of the estimate. There is no simple way to select optimal ρ, but there are some guideline (Boyd et al. 2011). In our simulations, we fix $\rho = 0.6$ and we use K-fold cross-validation can be used to choose λ, η.

Algorithm 1 ADMM for solving optimization problem (17)

1: **Input:** Initial values $\boldsymbol{\Theta}^{(0)}$, $\boldsymbol{Q}^{(0)}$, $\boldsymbol{R}^{(0)}$, $\boldsymbol{U}^{1(0)}$, $\boldsymbol{U}^{2(0)}$, parameters λ, η, ρ, ξ, maximum iterations T, and tolerance ϵ.
2: **Output:** $\boldsymbol{\Theta}$, $\boldsymbol{Q}$, $\boldsymbol{R}$.
3: **repeat**
4: **Update $\boldsymbol{\Theta}$:**

$$\boldsymbol{\Theta}^{(t+1)} = (\boldsymbol{I} + 2\rho\boldsymbol{I})^{-1}(\boldsymbol{V}^{(t)} - \boldsymbol{U}^{1(t)} + \rho\boldsymbol{Q}^{(t)} - \boldsymbol{U}^{2(t)} + \rho\boldsymbol{R}^{(t)})$$

5: **Update $\boldsymbol{Q}$:**

$$\boldsymbol{Q}^{(t+1)} = \mathcal{S}_{\frac{\lambda}{\rho\xi}}(\boldsymbol{\Theta}^{(t+1)} + \frac{\boldsymbol{U}^{1(t)}}{\rho})$$

6: **Update $\boldsymbol{R}$:**

(1) Perform SVD on $\boldsymbol{\Lambda}_2^{(t+1)} = \boldsymbol{\Theta}^{(t+1)} + \frac{\boldsymbol{U}^{2(t)}}{\rho}$:

$$\boldsymbol{\Lambda}_2^{(t+1)} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

(2) Apply soft-thresholding to $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}_{\frac{\eta}{\rho\xi},ii} = \mathcal{S}_{\frac{\eta}{\rho\xi}}(\sigma_i) = \max(\sigma_i - \frac{\eta}{\rho\xi}, 0),$$

(3) Reconstruct $\boldsymbol{R}^{(t+1)}$:

$$\boldsymbol{R}^{(t+1)} = \boldsymbol{U}\boldsymbol{\Sigma}_{\frac{\eta}{\rho\xi}}\boldsymbol{V}^T$$

7: **Update $\boldsymbol{U}^1$ and $\boldsymbol{U}^2$:**

$$\begin{aligned}\boldsymbol{U}^{1(t+1)} =& \boldsymbol{U}^{1(t)} + \rho(\boldsymbol{\Theta}^{(t+1)} - \boldsymbol{Q}^{(t+1)})\\ \boldsymbol{U}^{2(t+1)} =& \boldsymbol{U}^{2(t)} + \rho(\boldsymbol{\Theta}^{(t+1)} - \boldsymbol{R}^{(t+1)})\end{aligned}$$

8: **Check convergence:**

$$\begin{aligned}\text{primal residual} &= \|\boldsymbol{\Theta}^{(t+1)} - \boldsymbol{Q}^{(t+1)}\|_F + \|\boldsymbol{\Theta}^{(t+1)} - \boldsymbol{R}^{(t+1)}\|_F\\ \text{dual residual} &= \rho\|\boldsymbol{Q}^{(t+1)} - \boldsymbol{Q}^{(t)}\|_F + \rho\|\boldsymbol{R}^{(t+1)} - \boldsymbol{R}^{(t)}\|_F\end{aligned}$$

9: **until** (primal residual $\leq \epsilon$ and dual residual $\leq \epsilon$) or (number of iterations $\geq T$)