Supplementary Material for Sparse Reduced Rank Huber Regression in High Dimensions

A An Alternating Direction Method of Multipliers Algorithm

In this section, we develop an alternating direction method of multipliers (ADMM) algorithm for solving the convex relaxation (4), which allows us to decouple some of the terms that are difficult to optimize jointly (Eckstein and Bertsekas, 1992; Boyd et al., 2010). More specifically, the convex program is equivalent to

$$\begin{aligned} \underset{\mathbf{A},\mathbf{Z},\mathbf{W}\in\mathbb{R}^{p\times q},\mathbf{D}\in\mathbb{R}^{n\times q}}{\text{minimize}} \quad & \left\{\frac{1}{n}\ell_\tau\left(\mathbf{Y}-\mathbf{D}\right)+\lambda\left(\|\mathbf{W}\|_*+\gamma\|\mathbf{Z}\|_{1,1}\right)\right\}, \\ \text{subject to} \quad & \begin{pmatrix}\mathbf{D}\\ \mathbf{Z}\\ \mathbf{W}\end{pmatrix}=\begin{pmatrix}\mathbf{X}\\ \mathbf{I}\\ \mathbf{I}\end{pmatrix}\mathbf{A}. \end{aligned} \tag{12}$$

With some abuse of notation, let $\mathbf{B}=(\mathbf{B}_D,\mathbf{B}_Z,\mathbf{B}_W)^{\mathrm{T}}$, $\widetilde{\mathbf{X}}=(\mathbf{X},\mathbf{I},\mathbf{I})^{\mathrm{T}}$, and $\boldsymbol{\Omega}=(\mathbf{D},\mathbf{Z},\mathbf{W})^{\mathrm{T}}$. The scaled augmented Lagrangian of (12) takes the form

$$\mathcal{L}_\rho(\mathbf{A},\mathbf{D},\mathbf{Z},\mathbf{W},\mathbf{B})=\frac{1}{n}\ell_\tau(\mathbf{Y}-\mathbf{D})+\lambda\left(\|\mathbf{W}\|_*+\gamma\|\mathbf{Z}\|_{1,1}\right)+\frac{\rho}{2}\|\boldsymbol{\Omega}-\widetilde{\mathbf{X}}\mathbf{A}+\mathbf{B}\|_{\mathrm{F}}^2,$$

where $\mathbf{A},\mathbf{D},\mathbf{Z},\mathbf{W}$ are the primal variables, and $\mathbf{B}$ is the dual variable.

The ADMM algorithm requires the following updates:

1. $\mathbf{A}^{t+1}\leftarrow\underset{\mathbf{A}}{\text{argmin}}\ \mathcal{L}_\rho(\mathbf{A},\mathbf{D}^t,\mathbf{W}^t,\mathbf{Z}^t,\mathbf{B}^t)$.

2. $\mathbf{D}^{t+1}\leftarrow\underset{\mathbf{D}}{\text{argmin}}\ \mathcal{L}_\rho(\mathbf{A}^{t+1},\mathbf{D},\mathbf{W}^t,\mathbf{Z}^t,\mathbf{B}^t)$.

3. $\mathbf{W}^{t+1}\leftarrow\underset{\mathbf{W}}{\text{argmin}}\ \mathcal{L}_\rho(\mathbf{A}^{t+1},\mathbf{D}^{t+1},\mathbf{W},\mathbf{Z}^t,\mathbf{B}^t)$.

4. $\mathbf{Z}^{t+1} \leftarrow \underset{\mathbf{Z}}{\text{argmin}}\ \mathcal{L}_\rho(\mathbf{A}^{t+1}, \mathbf{D}^{t+1}, \mathbf{W}^{t+1}, \mathbf{Z}, \mathbf{B}^t)$.

5. $\mathbf{B}^{t+1} \leftarrow \mathbf{B}^t + \rho(\mathbf{X}(\mathbf{A}^{t+1}) - \boldsymbol{\Omega}^{t+1})$.

Update for A: To obtain an update for $\mathbf{A}$, we solve the following optimization problem

$$\underset{\mathbf{A}}{\text{minimize}}\ \left\|\boldsymbol{\Omega} + \mathbf{B} - \widetilde{\mathbf{X}}\mathbf{A}\right\|_{\text{F}}^2.$$

Thus, we obtain $\widehat{\mathbf{A}} = (\widetilde{\mathbf{X}}^{\text{T}}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^{\text{T}}(\boldsymbol{\Omega} + \mathbf{B})$.

Update for Z: To obtain an update for $\mathbf{Z}$, we need to solve the following optimization problem

$$\underset{\mathbf{Z}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{Z} - (\mathbf{A} - \mathbf{B}_Z)\|_{\text{F}}^2 + \frac{\lambda\gamma}{\rho}\|\mathbf{Z}\|_{1,1}.$$

Thus, we have $\widehat{\mathbf{Z}} = S(\mathbf{A} - \mathbf{B}_Z, \lambda\gamma/\rho)$.

Update for W: To obtain an update for $\mathbf{W}$, it amounts to solving

$$\underset{\mathbf{W}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{W} - (\mathbf{A} - \mathbf{B}_W)\|_{\text{F}}^2 + \frac{\lambda}{\rho}\|\mathbf{W}\|_*.$$

Let $\mathbf{A} - \mathbf{B}_W = \sum_{j=1}^{\min\{p,q\}} \omega_j \mathbf{a}_j \mathbf{b}_j^{\text{T}}$ be the singular value decomposition of $\mathbf{A} - \mathbf{B}_W$. Then, we obtain $\widehat{\mathbf{W}} = \sum_{j=1}^{\min\{p,q\}} \max(\omega_j - \lambda/\rho, 0)\, \mathbf{a}_j \mathbf{b}_j^{\text{T}}$.

Update for D: We solve the following problem to obtain an update for $\mathbf{D}$:

$$\underset{\mathbf{D}}{\text{minimize}}\ \frac{1}{n}\ell_\tau(\mathbf{Y} - \mathbf{D}) + \frac{\rho}{2}\|\mathbf{D} - (\mathbf{X}\mathbf{A} - \mathbf{B}_D)\|_{\text{F}}^2.$$

For notational convenience, let $\mathbf{C} = \mathbf{X}\mathbf{A} - \mathbf{B}_D$. We can solve the above problem element-wise:

$$\underset{D_{ij}}{\text{minimize}}\ \frac{1}{n}\ell_\tau(Y_{ij} - D_{ij}) + \frac{\rho}{2}(D_{ij} - C_{ij})^2.$$

Recall the Huber loss function from Definition 1 that there are two cases.

First, we assume that $|Y_{ij} - D_{ij}| \leq \tau$. Then, the above optimization problem reduces to

$$\underset{D_{ij}}{\text{minimize}}\ \frac{1}{2n}\left(Y_{ij} - D_{ij}\right)^2 + \frac{\rho}{2}(D_{ij} - C_{ij})^2.$$

Thus, we have $\widehat{D}_{ij} = (Y_{ij} + n\rho C_{ij})/(1 + n\rho)$. Substituting this into the constraint $|Y_{ij} - D_{ij}| \leq \tau$, we have $|[n\rho(Y_{ij} - C_{ij})]/(1 + n\rho)| \leq \tau$. Thus, $\widehat{D}_{ij} = (Y_{ij} + n\rho C_{ij})/(1 + n\rho)$ if $|[n\rho(Y_{ij} - C_{ij})]/(1 + n\rho)| \leq \tau$.

Next, we assume that $|Y_{ij} - D_{ij}| > \tau$. To obtain an estimate of D_{ij} in this case, we solve

$$\underset{D_{ij}}{\text{minimize}}\ \frac{\tau}{n}|Y_{ij} - D_{ij}| + \frac{\rho}{2}(D_{ij} - C_{ij})^2.$$

Let $H_{ij} = Y_{ij} - D_{ij}$. By a change of variable, we consider solving

$$\underset{H_{ij}}{\text{minimize}}\ \frac{1}{2}(Y_{ij} - C_{ij} - H_{ij})^2 + \frac{\tau}{n\rho}|H_{ij}|,$$

which yields the solution $\widehat{H}_{ij} = S(Y_{ij} - C_{ij}, \tau/(n\rho))$. Thus, we have $\widehat{D}_{ij} = Y_{ij} - S(Y_{ij} - C_{ij}, \tau/(n\rho))$.

Algorithm 2 summarizes the ADMM algorithm for solving (12). Since the term $(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}})^{-1}$ can be calculated before Step 2 in Algorithm 2, the computational bottleneck in each iteration of Algorithm 2 is the same as that of Algorithm 1.

B Proof of Lemma 1

Proof. The proposed Huber loss function can be written as

$$\mathcal{L}_\tau(\mathbf{A}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}\ell_\tau(Y_{ik} - \mathbf{X}_{i\cdot}^{\mathrm{T}}\mathbf{A}_{\cdot k}).$$

Let

$$\mathbf{T}_{i\tau} = \text{diag}\{1(|Y_{i1} - \mathbf{X}_{i\cdot}^{\mathrm{T}}\mathbf{A}_{\cdot 1}| \leq \tau), \ldots, 1(|Y_{iq} - \mathbf{X}_{i\cdot}^{\mathrm{T}}\mathbf{A}_{\cdot q}| \leq \tau)\}.$$

It can be shown that the pseudo Hessian takes the form

$$\mathbf{H}_\tau(\mathbf{A}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{T}_{i\tau} \otimes \mathbf{X}_{i\cdot}\mathbf{X}_{i\cdot}^{\mathrm{T}},$$

Algorithm 2 An ADMM Algorithm for Solving (12).

```
1. Initialize the parameters:
    (a) primal variables $\mathbf{A}, \mathbf{D}, \mathbf{Z}$, and $\mathbf{W}$ to the zero matrix.
    (b) dual variables $\mathbf{B}_D, \mathbf{B}_Z$, and $\mathbf{B}_W$ to the zero matrix.
    (c) constants $\rho > 0$ and $\epsilon > 0$.
2. Iterate until the stopping criterion $\|\mathbf{A}^t - \mathbf{A}^{t-1}\|_\mathrm{F}^2 / \|\mathbf{A}^{t-1}\|_\mathrm{F}^2 \leq \epsilon$ is met, where $\mathbf{A}^t$ is the value of $\mathbf{A}$ obtained at the $t$th iteration:
    (a) Update $\mathbf{A}, \mathbf{Z}, \mathbf{W}, \mathbf{D}$:
        i. $\mathbf{A} = (\widetilde{\mathbf{X}}^\mathrm{T}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\mathrm{T}(\mathbf{\Omega} + \mathbf{B})$.
        ii. $\mathbf{Z} = S(\mathbf{A} - \mathbf{B}_Z, \lambda\gamma/\rho)$. Here $S$ denote the soft-thresholding operator, applied element-wise to a matrix: $S(A_{ij}, b) = \mathrm{sign}(A_{ij}) \max(|A_{ij}| - b, 0)$.
        iii. $\mathbf{W} = \sum_j \max(\omega_j - \lambda/\rho, 0)\, \mathbf{a}_j \mathbf{b}_j^\mathrm{T}$, where $\sum_j \omega_j \mathbf{a}_j \mathbf{b}_j^\mathrm{T}$ is the singular value decomposition of $\mathbf{A} - \mathbf{B}_W$.
        iv. $\mathbf{C} = \mathbf{X}\mathbf{A} - \mathbf{B}_D$. Set
            $D_{ij} = \begin{cases} (Y_{ij} + n\rho C_{ij})/(1 + n\rho), & \text{if } |n\rho(Y_{ij} - C_{ij})/(1 + n\rho)| \leq \tau, \\ Y_{ij} - S(Y_{ij} - C_{ij}, \tau/(n\rho)), & \text{otherwise.} \end{cases}$
    (b) Update $\mathbf{B}_D, \mathbf{B}_Z, \mathbf{B}_W$:
        i. $\mathbf{B}_D = \mathbf{B}_D + \mathbf{D} - \mathbf{X}\mathbf{A}$;   ii. $\mathbf{B}_Z = \mathbf{B}_Z + \mathbf{Z} - \mathbf{A}$;   iii. $\mathbf{B}_W = \mathbf{B}_W + \mathbf{W} - \mathbf{A}$.
```

where $\otimes$ is the kronecker product between two matrices. For notational convenience, let

$$\widetilde{\mathbf{T}}_{i\tau} = \mathrm{diag}\{1(|Y_{i1} - \mathbf{X}_{i\cdot}^\mathrm{T}\mathbf{A}_{\cdot 1}| > \tau), \ldots, 1(|Y_{iq} - \mathbf{X}_{i\cdot}^\mathrm{T}\mathbf{A}_{\cdot q}| > \tau)\}.$$

Let $\widetilde{\mathbf{u}} = \text{vec}(\mathbf{U})$. For any $(\mathbf{U}, \mathbf{A}) \in \mathcal{C}(m, \xi, \eta)$, we have

$$\begin{aligned} \widetilde{\mathbf{u}}^{\mathrm{T}} \mathbf{H}_{\tau}(\mathbf{A}) \widetilde{\mathbf{u}} &= \widetilde{\mathbf{u}}^{\mathrm{T}} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{T}_{i\tau} \otimes \mathbf{X}_{i\cdot} \mathbf{X}_{i\cdot}^{\mathrm{T}} \right) \widetilde{\mathbf{u}} \\ &= \widetilde{\mathbf{u}}^{\mathrm{T}} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_q \otimes \mathbf{X}_{i\cdot} \mathbf{X}_{i\cdot}^{\mathrm{T}} \right) \widetilde{\mathbf{u}} - \widetilde{\mathbf{u}}^{\mathrm{T}} \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathbf{T}}_{i\tau} \otimes \mathbf{X}_{i\cdot} \mathbf{X}_{i\cdot}^{\mathrm{T}} \right) \widetilde{\mathbf{u}} \qquad (13) \\ &= \| \widetilde{\mathbf{S}}^{1/2} \widetilde{\mathbf{u}} \|_2^2 - \widetilde{\mathbf{u}}^{\mathrm{T}} \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathbf{T}}_{i\tau} \otimes \mathbf{X}_{i\cdot} \mathbf{X}_{i\cdot}^{\mathrm{T}} \right) \widetilde{\mathbf{u}}, \end{aligned}$$

where $\widetilde{\mathbf{S}} = n^{-1} \sum_{i=1}^{n} \mathbf{I}_q \otimes \mathbf{X}_{i\cdot} \mathbf{X}_{i\cdot}^{\mathrm{T}}$. We now obtain an upper bound for each element in $\widetilde{\mathbf{T}}_{i\tau}$. For $1 \leq j \leq q$,

$$\begin{aligned} 1(|Y_{ij} - \mathbf{X}_{i\cdot}^{\mathrm{T}} \mathbf{A}_{\cdot j}| > \tau) &= 1(|Y_{ij} - \mathbf{X}_{i\cdot}^{\mathrm{T}} \mathbf{A}_{\cdot j}^* + \mathbf{X}_{i\cdot}^{\mathrm{T}} \mathbf{A}_{\cdot j}^* - \mathbf{X}_{i\cdot}^{\mathrm{T}} \mathbf{A}_{\cdot j}| > \tau) \\ &\leq 1(|E_{ij}| > \tau/2) + 1(|\mathbf{X}_{i\cdot}^{\mathrm{T}} (\mathbf{A}_{\cdot j}^* - \mathbf{A}_{\cdot j})| > \tau/2). \qquad (14) \end{aligned}$$

Moreover, we have

$$\begin{aligned} 1(|\mathbf{X}_{i\cdot}^{\mathrm{T}} (\mathbf{A}_{\cdot j}^* - \mathbf{A}_{\cdot j})| > \tau/2) &= 1 \left(\frac{2}{\tau} |\mathbf{X}_{i\cdot}^{\mathrm{T}} (\mathbf{A}_{\cdot j}^* - \mathbf{A}_{\cdot j})| > 1 \right) \\ &\leq \frac{2}{\tau} |\mathbf{X}_{i\cdot}^{\mathrm{T}} (\mathbf{A}_{\cdot j}^* - \mathbf{A}_{\cdot j})| \qquad (15) \\ &\leq \frac{2\eta}{\tau} \max_{1 \leq i \leq n} \|\mathbf{X}_{i\cdot}\|_{\infty} \qquad (16) \\ &\leq \frac{2\eta}{\tau}. \qquad (17) \end{aligned}$$

where the second inequality holds by Holder's inequality and the condition that $\|\mathbf{A}_{\cdot j}^* - \mathbf{A}_{\cdot j}\|_1 \leq \eta$. Let $\mathbf{u}_j$ be the jth column of $\mathbf{U}$. Since, $\widetilde{\mathbf{T}}_{i\tau}$ is a diagonal matrix, we obtain

$$\begin{aligned} &\widetilde{\mathbf{u}}^{\mathrm{T}} \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathbf{T}}_{i\tau} \otimes \mathbf{X}_{i\cdot} \mathbf{X}_{i\cdot}^{\mathrm{T}} \right) \widetilde{\mathbf{u}} \\ &= \sum_{j=1}^{q} \mathbf{u}_j^{\mathrm{T}} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_{i\cdot} \mathbf{X}_{i\cdot}^{\mathrm{T}} \cdot 1(|Y_{ij} - \mathbf{X}_{i\cdot}^{\mathrm{T}} \mathbf{A}_{\cdot j}| > \tau) \right) \mathbf{u}_j \\ &\leq \sum_{j=1}^{q} \mathbf{u}_j^{\mathrm{T}} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_{i\cdot} \mathbf{X}_{i\cdot}^{\mathrm{T}} \cdot 1(|E_{ij}| > \tau/2) \right) \mathbf{u}_j \qquad (18) \\ &\quad + \sum_{j=1}^{q} \mathbf{u}_j^{\mathrm{T}} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_{i\cdot} \mathbf{X}_{i\cdot}^{\mathrm{T}} \cdot 1(|\mathbf{X}_{i\cdot}^{\mathrm{T}} (\mathbf{A}_{\cdot j}^* - \mathbf{A}_{\cdot j})| > \tau/2) \right) \mathbf{u}_j \\ &\leq \frac{2\eta}{\tau} \| \widetilde{\mathbf{S}}^{1/2} \widetilde{\mathbf{u}} \|_2^2 + \max_{1 \leq i \leq n} \sum_{j=1}^{q} (\mathbf{X}_{i\cdot}^{\mathrm{T}} \mathbf{u}_j)^2 \cdot \max_{1 \leq j \leq q} \left(\frac{1}{n} \sum_{i=1}^{n} 1(|E_{ij}| > \tau/2) \right), \end{aligned}$$

where the first inequality holds by (14) and the last inequality holds by (15).

By Lemma 11, for any $1 \le j \le q$ and $t > 0$, we have

$$\frac{1}{n}\sum_{i=1}^{n} 1(|E_{ij}| > \tau/2) \le (2/\tau)^{1+\delta}\nu_\delta + \sqrt{t/n}$$

with probability at least $1 - \exp(-2t)$. Moreover, for any $1 \le i \le n$, we have

$$\sum_{j=1}^{q} |\mathbf{X}_{i\cdot}^{\mathrm{T}}\mathbf{u}_j| \le \|\mathbf{X}_{i\cdot}^{\mathrm{T}}\|_\infty \|\widetilde{\mathbf{u}}\|_1 \le (1+\xi)\|\widetilde{\mathbf{u}}_{\mathcal{S}}\|_1 \le (1+\xi)\sqrt{m}\|\widetilde{\mathbf{u}}_{\mathcal{S}}\|_2.$$

Thus, combining the above with (13) and (18), we have

$$\widetilde{\mathbf{u}}^{\mathrm{T}}\mathbf{H}_\tau(\mathbf{A})\widetilde{\mathbf{u}} \ge \|\widetilde{\mathbf{S}}^{1/2}\widetilde{\mathbf{u}}\|_2^2 - \frac{2\eta}{\tau}\|\widetilde{\mathbf{S}}^{1/2}\widetilde{\mathbf{u}}\|_2^2 - (1+\xi)^2 m\left[(2/\tau)^{1+\delta}\nu_\delta + \sqrt{t/n}\right].$$

Consequently, picking $\tau \ge \max(8\eta, C(m\nu_\delta)^{1/(1+\delta)})$, $t = \log(pq)/2$, and $n > C'(m^2\log(pq))$ for sufficiently large C and C', we have

$$\widetilde{\mathbf{u}}^{\mathrm{T}}\mathbf{H}_\tau(\mathbf{A})\widetilde{\mathbf{u}} \ge \frac{3}{4}\kappa_{\text{lower}} - m(1+\xi)^2\left[(2/\tau)^{1+\delta}\nu_\delta + \sqrt{t/n}\right] \ge \frac{1}{2}\kappa_{\text{lower}},$$

with probability at least $1 - (pq)^{-1}$.

The upper bound $\widetilde{\mathbf{u}}^{\mathrm{T}}\mathbf{H}_\tau(\mathbf{A})\widetilde{\mathbf{u}} \le \kappa_{\text{upper}}$ can be obtained similarly.

□

C Proof of Theorem 1

Recall from (4) that the optimization problem takes the form

$$\underset{\mathbf{A}}{\text{minimize}}\left\{\mathcal{L}_\tau(\mathbf{A}) + \lambda\left(\|\mathbf{A}\|_* + \gamma\|\mathbf{A}\|_{1,1}\right)\right\}, \tag{19}$$

where we use the notation $\mathcal{L}_\tau(\mathbf{A}) = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{q}\ell_\tau(Y_{ik} - \mathbf{X}_{i\cdot}^{\mathrm{T}}\mathbf{A}_{.k})$ for convenience throughout the proof. We start with stating some facts and notation.

Let $\mathbf{A} \in \mathbb{R}^{p\times q}$ be a rank r matrix with singular value decomposition $\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{U} \in \mathbb{R}^{p\times r}$, $\mathbf{V} \in \mathbb{R}^{q\times r}$, and $\boldsymbol{\Lambda} \in \mathbb{R}^{r\times r}$. The sub-differential of the nuclear norm is then given by (see, for instance, Recht et al., 2010)

$$\partial\|\mathbf{A}\|_* = \left\{\mathbf{U}\mathbf{V}^{\mathrm{T}} + \mathbf{W} : \mathbf{W} \in \mathbb{R}^{p\times q}, \mathbf{U}^{\mathrm{T}}\mathbf{W} = \mathbf{0}, \mathbf{W}\mathbf{V} = \mathbf{0}, \|\mathbf{W}\|_2 \le 1\right\}. \tag{20}$$

Let $\mathcal{F}(r) = \{\mathbf{A} \in \mathbb{R}^{p\times q} : \mathrm{rank}(\mathbf{A}) \le r\}$ be the algebraic variety of matrices with rank at most r. Then, the tangent space at $\mathbf{A}$ with respect to $\mathcal{F}(r)$ is given by

$$T(\mathbf{A}) = \left\{\mathbf{U}\mathbf{W}_1^{\mathrm{T}} + \mathbf{W}_2\mathbf{V}^{\mathrm{T}} : \mathbf{W}_1 \in \mathbb{R}^{q\times r}, \mathbf{W}_2 \in \mathbb{R}^{p\times r}\right\},$$

where $T(\mathbf{A})$ can be interpreted as a subspace in $\mathbb{R}^{p\times q}$ (Chandrasekaran et al., 2012). We now state a connection between the sub-differential of the nuclear norm and its tangent space. Let $\mathcal{P}_{T(\mathbf{A})}$ denote the projection operator onto $T(\mathbf{A})$. Then, it can be shown that the following relationship holds

$$\widetilde{\mathbf{N}} \in \partial\|\mathbf{A}\|_* \qquad \text{if and only if} \qquad \mathcal{P}_{T(\mathbf{A})}(\widetilde{\mathbf{N}}) = \mathbf{U}\mathbf{V}^{\mathrm{T}}, \;\; \|\mathcal{P}_{T(\mathbf{A})^\perp}\widetilde{\mathbf{N}}\|_2 \le 1.$$

In addition, we define several quantities that will be used in the proof. For any convex loss function $\mathcal{L}_\tau(\cdot)$, the Bregman divergence between $\widehat{\mathbf{A}}$ and $\mathbf{A}^*$ is

$$D_{\mathcal{L}}(\widehat{\mathbf{A}}, \mathbf{A}^*) = \mathcal{L}_\tau(\widehat{\mathbf{A}}) - \mathcal{L}_\tau(\mathbf{A}^*) - \langle \nabla\mathcal{L}_\tau(\mathbf{A}^*), \widehat{\mathbf{A}} - \mathbf{A}^*\rangle \ge 0.$$

We define the symmetric Bregman divergence as

$$D^s_{\mathcal{L}}(\widehat{\mathbf{A}}, \mathbf{A}^*) = D_{\mathcal{L}}(\widehat{\mathbf{A}}, \mathbf{A}^*) + D_{\mathcal{L}}(\mathbf{A}^*, \widehat{\mathbf{A}}) = \langle \nabla\mathcal{L}_\tau(\widehat{\mathbf{A}}) - \nabla\mathcal{L}_\tau(\mathbf{A}^*), \widehat{\mathbf{A}} - \mathbf{A}^*\rangle \ge 0 \tag{21}$$

The proof involves obtaining an upper bound and a lower bound for the symmetric Bregman divergence. To this end, we state some technical lemmas that will be used in the proof.

Lemma 2. Assume that the covariates are standardized such that $\max_{i,j}|X_{ij}| = 1$ and that E_{ik} is such that $v_\delta = \mathbb{E}(|E_{ik}|^{1+\delta}) < \infty$. Pick $\tau \ge C_1\{nv_\delta/\log(pq)\}^{\min\{1/2, 1/(1+\delta)\}}$, we have

$$\|\nabla\mathcal{L}_\tau(\mathbf{A}^*)\|_{\infty,\infty} \le C_2 v_\delta^{1/\min(1+\delta,2)}\left(\frac{\log(pq)}{n}\right)^{\min\{1/2,\delta/(1+\delta)\}},$$

with probability at least $1 - (pq)^{-1}$, where C_1 and C_2 are universal constants.

Lemma 3 ($\ell_{1,1}$-Cone Property)**.** Assume that $\|\nabla\mathcal{L}_\tau(\mathbf{A}^*)\|_{\infty,\infty} \le \lambda/2$. Let $\widehat{\mathbf{A}}$ be a solution to (4). We have $\widehat{\mathbf{A}}$ falls in the following $\ell_{1,1}$-cone

$$\big\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}^c}\big\|_{1,1} \le \frac{2\gamma+5}{2\gamma-5}\big\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\big\|_{1,1}.$$

Let $\mathcal{U}$ be the linear space spanned by the columns of $\mathbf{U}$, and $\mathcal{V}$ the linear space spanned by the columns of $\mathbf{V}$. We denote by $\mathcal{U}^\perp$ and $\mathcal{V}^\perp$ the orthogonal complements of $\mathcal{U}$ and $\mathcal{V}$, respectively.

Lemma 4 (Nuclear Cone Property)**.** Assume that $\|\nabla\mathcal{L}_\tau(\mathbf{A}^*)\|_{\infty,\infty} \le \lambda/2$ and $\gamma \ge 1/2$. We have

$$\big\|\mathcal{P}_{\mathcal{U}^\perp}(\widehat{\mathbf{A}} - \mathbf{A}^*)\mathcal{P}_{\mathcal{V}^\perp}\big\|_* \le \big\|\mathcal{P}_{\mathcal{U}}(\widehat{\mathbf{A}} - \mathbf{A}^*)\mathcal{P}_{\mathcal{V}}\big\|_* + (\gamma+0.5)\big\|\big(\widehat{\mathbf{A}} - \mathbf{A}^*\big)_{\mathcal{S}}\big\|_{1,1}.$$

Lemma 5 (Restricted Strong Convexity)**.** Under the same conditions as in Lemma 1, for matrices $(\mathbf{A},\mathbf{U}) \in \mathcal{C}(m,\xi,\eta)$, we have

$$\mathcal{D}^s_{\mathcal{L}}(\mathbf{A},\mathbf{A}^*) \ge \frac{\kappa_{\text{lower}}}{2}\|\mathbf{A} - \mathbf{A}^*\|_{\text{F}}^2,$$

with probability at least $1-(pq)^{-1}$.

To prove Theorem 1, we obtain upper and lower bounds for the symmetric Bregman divergence, respectively.

Proof. **Upper bound under Frobenius norm:** By the first order optimality condition of (4), there exists $\widetilde{\mathbf{N}} \in \partial\|\widehat{\mathbf{A}}\|_*$ and $\widetilde{\mathbf{\Gamma}} \in \partial\|\widehat{\mathbf{A}}\|_{1,1}$ such that

$$\nabla\mathcal{L}_\tau(\widehat{\mathbf{A}}) + \lambda(\widetilde{\mathbf{N}} + \gamma\widetilde{\mathbf{\Gamma}}) = \mathbf{0}. \tag{22}$$

Substituting (22) into (21), we have

$$\begin{aligned} D^s_{\mathcal{L}}(\widehat{\mathbf{A}},\mathbf{A}^*) &= \langle -\lambda\widetilde{\mathbf{N}} - \lambda\gamma\widetilde{\mathbf{\Gamma}} - \nabla\mathcal{L}_\tau(\mathbf{A}^*), \widehat{\mathbf{A}} - \mathbf{A}^*\rangle \\ &= \underbrace{\langle\nabla\mathcal{L}_\tau(\mathbf{A}^*), \mathbf{A}^* - \widehat{\mathbf{A}}\rangle}_{\mathrm{I}_1} + \underbrace{\lambda\langle\widetilde{\mathbf{N}}, \mathbf{A}^* - \widehat{\mathbf{A}}\rangle}_{\mathrm{I}_2} + \underbrace{\lambda\gamma\langle\widetilde{\mathbf{\Gamma}}, \mathbf{A}^* - \widehat{\mathbf{A}}\rangle}_{\mathrm{I}_3}. \end{aligned} \tag{23}$$

Upper bound on I_1: By the Holder's inequality, we have

$$\begin{aligned}
\mathrm{I}_1 &\leq \|\nabla \mathcal{L}_\tau(\mathbf{A}^*)\|_{\infty,\infty} \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} \\
&\leq \frac{\lambda}{2} \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} \\
&= \frac{\lambda}{2} \left(\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\|_{1,1} + \|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}^c}\|_{1,1} \right) \\
&\leq \frac{2\lambda\gamma}{2\gamma - 5} \|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\|_{1,1},
\end{aligned} \tag{24}$$

where the last inequality holds by Lemma 3.

Upper bound on I_2: By the Holder's inequality, we have

$$\begin{aligned}
\mathrm{I}_2 &\leq \lambda \|\widetilde{\mathbf{N}}\|_{\infty,\infty} \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} \\
&\leq \lambda \|\widetilde{\mathbf{N}}\|_2 \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} \\
&\leq 2\lambda \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} \\
&\leq \frac{8\lambda\gamma}{2\gamma - 5} \|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\|_{1,1},
\end{aligned} \tag{25}$$

where the second inequality holds by the fact that $\|\widetilde{\mathbf{N}}\|_2 \leq 2$, and the last inequality holds by Lemma 3.

Upper bound on I_3: Similarly, by Holder's inequality and using the fact that $\|\widetilde{\mathbf{\Gamma}}\|_{\infty,\infty} \leq 1$, we obtain

$$\mathrm{I}_3 \leq \lambda\gamma \|\widetilde{\mathbf{\Gamma}}\|_{\infty,\infty} \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} \leq \lambda\gamma \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} \leq \frac{4\lambda\gamma^2}{2\gamma - 5} \|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\|_{1,1}, \tag{26}$$

where the last inequality holds by Lemma 3.

Thus, substituting (24), (25), and (26) into (23), we obtain

$$D_{\mathcal{L}}^s(\widehat{\mathbf{A}}, \mathbf{A}^*) \leq \frac{4\gamma^2 + 10\gamma}{2\gamma - 5} \lambda \|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\|_{1,1} \leq \frac{4\gamma^2 + 10\gamma}{2\gamma - 5} \lambda \sqrt{s} \|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\|_{\mathrm{F}}, \tag{27}$$

where $s \le r s_u s_v$ is the sparsity parameter of $\mathbf{A}^*$, that is $s = |\mathrm{supp}(\mathbf{A}^*)|$.

Next, we employ Lemma 5 to obtain a lower bound for the symmetric Bregman divergence. Lemma 5 requires the matrix $\mathbf{A} \in \mathcal{C}(m, \xi, \eta)$. To this end, we construct the matrix $\widehat{\mathbf{A}}_\eta = \mathbf{A}^* + \zeta(\widehat{\mathbf{A}} - \mathbf{A}^*)$ such that $\|\widehat{\mathbf{A}}_\eta - \widehat{\mathbf{A}}^*\|_{1,1} \le \eta$ for some $\eta > 0$. If $\|\widehat{\mathbf{A}} - \mathbf{A}^*\| < \eta$, we set $\zeta = 1$, so $\widehat{\mathbf{A}}_\eta = \widehat{\mathbf{A}}$. Otherwise, we pick $\zeta \in (0,1)$ such that $\|\widehat{\mathbf{A}}_\eta - \mathbf{A}^*\|_{1,1} = \eta$. By Lemma 3, it can be shown that $\widehat{\mathbf{A}}_\eta$ falls in an ℓ_1-cone, and thus, $\widehat{\mathbf{A}}_\eta \in \mathcal{C}(m, \xi, \eta)$ with

$$\|(\widehat{\mathbf{A}}_\eta - \mathbf{A}^*)_{\mathcal{S}^c}\|_{1,1} \le \frac{2\gamma+5}{2\gamma-5}\|(\widehat{\mathbf{A}}_\eta - \mathbf{A}^*)_{\mathcal{S}}\|_{1,1} \qquad \text{and} \qquad \|\widehat{\mathbf{A}}_\eta - \mathbf{A}^*\|_{1,1} \le \eta. \tag{28}$$

Therefore, by Lemma 5, we have

$$D^s_{\mathcal{L}}(\widehat{\mathbf{A}}_\eta, \mathbf{A}^*) \ge \frac{\kappa_{\text{lower}}}{2}\|\widehat{\mathbf{A}}_\eta - \mathbf{A}^*\|^2_{\mathrm{F}}. \tag{29}$$

By Lemma A.1 of Sun et al. (2018),

$$D^s_{\mathcal{L}}(\widehat{\mathbf{A}}_\eta, \mathbf{A}^*) \le \zeta D^s_{\mathcal{L}}(\widehat{\mathbf{A}}, \mathbf{A}^*). \tag{30}$$

Combining (29) and (30) yields

$$\|\widehat{\mathbf{A}}_\eta - \mathbf{A}^*\|^2_{\mathrm{F}} \le \zeta \kappa^{-1}_{\text{lower}} \frac{8\gamma^2 + 20\gamma}{2\gamma-5}\lambda\sqrt{s}\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{\mathrm{F}}.$$

Since $\widehat{\mathbf{A}} - \mathbf{A}^* = \zeta^{-1}(\widehat{\mathbf{A}}_\eta - \mathbf{A}^*)$, this yields

$$\|\widehat{\mathbf{A}}_\eta - \mathbf{A}^*\|_{\mathrm{F}} \le \kappa^{-1}_{\text{lower}} \frac{8\gamma^2 + 20\gamma}{2\gamma-5}\lambda\sqrt{s}.$$

Finally, by (28), we have

$$\|\widehat{\mathbf{A}}_\eta - \mathbf{A}^*\|_{1,1} \le \frac{4\gamma\sqrt{s}}{2\gamma-5}\|(\widehat{\mathbf{A}}_\eta - \mathbf{A}^*)_{\mathcal{S}}\|_{\mathrm{F}} \le \kappa^{-1}_{\text{lower}} \frac{4\gamma}{2\gamma-5}\frac{8\gamma^2 + 20\gamma}{2\gamma-5}\lambda s < \eta,$$

where the last inequality holds by the assumption that $n > Cs^2\log(pq)$ for some sufficiently large constant $C > 0$. By the construction of $\widehat{\mathbf{A}}_\eta$, since $\|\widehat{\mathbf{A}}_\eta - \mathbf{A}^*\|_{1,1} < \eta$, we have $\widehat{\mathbf{A}}_\eta = \widehat{\mathbf{A}}$, implying

$$\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{\mathrm{F}} \leq \kappa_{\text{lower}}^{-1} \frac{8\gamma^2 + 20\gamma}{2\gamma - 5} \lambda\sqrt{s}.$$

Upper bound under nuclear norm: Next, we establish an upper bound for $\widehat{\mathbf{A}} - \mathbf{A}^*$ under the nuclear norm. Recall that $s = |\text{supp}(\mathbf{A}^*)|$. We have shown previously that $\widehat{\mathbf{A}}$ is in $\mathcal{C}$. Applying Lemma 4, we can bound $\|\mathcal{P}_{\mathcal{U}^\perp}(\widehat{\mathbf{A}} - \mathbf{A}^*)\mathcal{P}_{\mathcal{V}^\perp}\|_*$ as

$$\begin{aligned}
\big\|\mathcal{P}_{U^\perp}(\widehat{\mathbf{A}} - \mathbf{A}^*)\mathcal{P}_{V^\perp}\big\|_* &\leq \big\|\mathcal{P}_{\mathcal{U}}(\widehat{\mathbf{A}} - \mathbf{A}^*)\mathcal{P}_{\mathcal{V}}\big\|_* + (\gamma + 0.5)\big\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\big\|_{1,1} \\
&\leq \sqrt{r}\big\|\mathcal{P}_{\mathcal{U}}(\widehat{\mathbf{A}} - \mathbf{A}^*)\mathcal{P}_{\mathcal{V}}\big\|_{\mathrm{F}} + (\gamma + 0.5)\sqrt{s}\big\|\widehat{\mathbf{A}} - \mathbf{A}^*\big\|_{\mathrm{F}} \\
&\lesssim \kappa_{\text{lower}}^{-1} \frac{4\gamma^2 + 10\gamma}{2\gamma - 5} \lambda\sqrt{s}\big\{\sqrt{r} \vee (\gamma + 0.5)\sqrt{s}\big\}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\big\|\widehat{\mathbf{A}} - \mathbf{A}^*\big\|_* &\leq \big\|\mathcal{P}_{T_*}(\widehat{\mathbf{A}} - \mathbf{A}^*)\big\|_* + \big\|\mathcal{P}_{T_*^\perp}(\widehat{\mathbf{A}} - \mathbf{A}^*)\big\|_* \\
&\lesssim \kappa_{\text{lower}}^{-1} \frac{4\gamma^2 + 10\gamma}{2\gamma - 5} \lambda\sqrt{s}\big\{2\sqrt{r} \vee (\gamma + 0.5)\sqrt{s}\big\} \\
&\leq C_\gamma \kappa_{\text{lower}}^{-1} \lambda\sqrt{s}(\sqrt{r} \vee \sqrt{s}) \\
&\lesssim \kappa_{\text{lower}}^{-1} \lambda\sqrt{s}(\sqrt{r} \vee \sqrt{s}),
\end{aligned}$$

where $C_\gamma = (2\gamma - 5)^{-1}(4\gamma^2 + 10\gamma)\big\{2 \vee (\gamma + 0.5)\big\}$ is a constant depending only on γ.

□

D Proof of Theorem 2

For any probability distributions P and Q, let $D(P\|Q)$ denote the Kullback-Leibler divergence of Q from P. For any subset K of $R^{n\times n}$, the volume of K is $\text{vol}(K) = \int K d\mu$ where $d\mu$ is the usual Lebesgue measure on $R^{m\times n}$ by taking the product measure of the Lebesgue measures of individual entries. With these definitions, we state the following variant of Fano's lemma (Ma and Wu, 2015).

Lemma 6. Let (Θ, ρ) be a metric space and $\{P_\theta : \theta \in \Theta\}$ a collection of probability measures. For any totally bounded $T \subset \Theta$, denote by $\mathcal{M}(T, \rho, \epsilon)$ the ϵ-packing number of T with respect to ρ, i.e., the maximal number of points in T whose pairwise minimum distance is at least ϵ. Define the Kullback-Leibler diameter of T by $d_{\text{KL}}(T) = \sup_{\theta,\theta'} D(P_\theta \| P'_\theta)$. Then

$$\inf \sup \mathbb{E}[\|\widehat{\theta} - \theta\|^2] \geq \sup_{T\subset\Theta} \sup_{\epsilon>0} \frac{\epsilon^2}{4} \left(1 - \frac{d_{\text{KL}}(T) + \log 2}{\log \mathcal{M}(T, \rho, \epsilon)}\right).$$

In particular, if $\Theta \subseteq \mathbb{R}^d$ and $\|\cdot\|$ is some norm then

$$\inf \sup \mathbb{E}[\|\widehat{\theta} - \theta\|^2] \geq \sup_{T\subset\Theta} \sup_{\epsilon>0} \frac{\epsilon^2}{4} \left(1 - \frac{d_{\text{KL}}(T) + \log 2}{\log \frac{\text{vol}(T)}{\text{vol}B_{\|\cdot\|}(\epsilon)}}\right).$$

We first use Lemma 6 to prove an oracle result in the sense that we know the locations of the sparse signals. In order to apply the Lemma 6, we need bound three quantities $d_{\text{KL}}(T)$, $\text{vol}(T)$ and $\text{vol}B_{\|\cdot\|}(\epsilon)$. Now consider $\mathcal{F}_0 \subseteq \mathcal{F}$ such that $\mathbf{u}_j$'s $((j-1)s_u + 1)$-th to js_u-th elements, s_u elements in total, are nonzeros and $\mathbf{v}_j$'s $((j-1)s_v + 1)$-th to js_v-th elements, s_v elements in total, are nonzeros, while all other entries are zeros. In this case, we may assume for simplicity that $p = rs_u$ and $q = rs_v$. In this case, only $d = rs_v s_v$, instead of $r^2 s_u s_v$, of the entries are nonzeros. Denote the support by $\mathcal{A}$.

We then proceed in two cases by assuming $s_v = 1$ and $s_u = 1$ respectively.

Case 1: $s_v = 1$. The for any a, let $B_{\text{F}}(a) = B_{\|\cdot\|_{\text{F}}}(a) = \{\mathbf{A} \in \mathbb{R}^{p\times q} : \|\mathbf{A}\|_{\text{F}} \leq a, \mathbf{A}_{\mathcal{A}^c} = 0\}$. Denote $B_{\text{F}}(a)$ by $T(a)$. It is easy to see that $T(a) \subset \mathcal{F}$.

Then for any $\mathbf{A}_1, \mathbf{A}_2 \in T(a)$, we have by the Condition 3 with support size rs_u and an arbitrary ξ

$$D(P_{\mathbf{A}_1} \| P_{\mathbf{A}_2}) = \frac{1}{2}\|\mathbf{X}\mathbf{A}_1 - \mathbf{X}\mathbf{A}_2\|_{\text{F}}^2 \leq \frac{1}{2} n\kappa_{\text{upper}} \|\mathbf{A}_1 - \mathbf{A}_2\|_{\text{F}}^2 \leq 2\kappa_{\text{upper}} na^2.$$

Thus the diameter satisfies that $d_{\text{KL}}(T(a)) \leq 2\kappa_{\text{upper}} na^2$. To obtain a lower bound for $\text{vol}(T(a))$, we apply the inverse Santaló's inequality which implies, for some numerical constant c_0, that

$$\mathrm{vol}(T(a))^{1/(rs_u)} = a\,\mathrm{vol}(B_{\mathrm{F}}(1))^{1/(rs_u)} \geq a\,\frac{c_0}{\mathbb{E}\|\mathbf{Z}_{\mathcal{A}}\|_{\mathrm{F}}} \geq a\,\frac{c_0}{\sqrt{rs_u}},$$

where $\mathbf{Z}_{\mathcal{A}} \in \mathbb{R}^{rs_u \times rs_v}$ is a random matrix with entires in $\mathcal{A}$ following i.i.d. standard normals while all entries in $\mathcal{A}^c$ are zeroes. The last inequality follows from Jensen's inequality. To obtain an upper bound for $\mathrm{vol}(B_{\|\|_{\mathrm{F}}}(\epsilon))$, we have

$$\mathrm{vol}(B_{\mathrm{F}}(\epsilon))^{1/(rs_u)} \leq \frac{\epsilon}{\sqrt{rs_u}}.$$

Consequently we have

$$\frac{(T(a))^{1/(rs_u)}}{\mathrm{vol}(B_{\mathrm{F}}(\epsilon))^{1/(rs_u)}} \geq \frac{ac_0}{\epsilon}.$$

For $\mathrm{vol}(B_{\|\|_*}(\epsilon))$, we use Urysohn's inequality to obtain that

$$\frac{\mathrm{vol}(B_*(\epsilon))^{1/(rs_u)}}{\mathrm{vol}(B_{\mathrm{F}}(a))^{1/(rs_u)}} \leq \frac{\epsilon\mathbb{E}\|\mathbf{Z}_{\mathcal{A}}\|_{s_\infty}}{a\sqrt{rs_u}} = \frac{\epsilon\mathbb{E}\|\mathbf{Z}_{\mathcal{A}}\|_2}{a\sqrt{rs_u}} \leq \frac{4\epsilon}{a}\,\frac{\sqrt{s_u}+\sqrt{1}}{\sqrt{rs_u}}\sqrt{\log r} \leq \frac{8\epsilon}{a\sqrt{r}}\sqrt{\log r},$$

where s_∞ is the Schattern infinity norm and the last second inequality uses a generalized Gordon's inequality, that is, Lemma 8.

Let

$$a^2 = \frac{rs_u}{n}, \text{ and } \epsilon = c_0' c_0 a,$$

where c_0', depending on κ_{upper}, is a small enough constant such that

$$\frac{d_{\mathrm{KL}}(T(a)) + \log 2}{\log \frac{\mathrm{vol}(T(a))}{\mathrm{vol} B_{\|\cdot\|}(\epsilon)}} \leq \frac{1}{2}.$$

Then applying Lemma 6 with $T(a)$ and ϵ specified previously and optimizing ϵ, we obtain that

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}} \mathbb{E}\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{\mathrm{F}}^2 \geq c\frac{rs_u}{n},$$

where c is a constant only depending on κ_{upper}. For the nuclear norm lower bound, taking

$$a^2 = \frac{rs_u}{n},\ \epsilon = c_0' a\sqrt{r}/\sqrt{\log r},$$

implies

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}} \mathbb{E}\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_*^2 \geq c\frac{r^2 s_u}{n\log r}.$$

Case 2: $s_u = 1$. Similarly in this case, we can obtain that

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}} \mathbb{E}\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{\mathrm{F}}^2 \geq c\frac{rs_v}{n},$$

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}} \mathbb{E}\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_*^2 \geq c\frac{r^2 s_v}{n\log r}.$$

Combining two cases, we shall obtain that

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}} \mathbb{E}\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{\mathrm{F}}^2 \geq c\frac{r(s_u + s_v)}{n},$$

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}} \mathbb{E}\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_*^2 \geq c\frac{r^2(s_u + s_v)}{n\log r}.$$

Now we move to the general case that we do not know the locations of the sparse signals. In this case we will pay a log factor for searching the sparse signals. To prove the general result, we first need the following lemma.

Lemma 7. Suppose $2(r-1)(s_u \vee s_v) \leq p \wedge q$. There exis a subclass $\mathcal{F}_0$ of $\mathcal{F}$ and a positive constant c_1 such that

$$\log \mathcal{M}(\mathcal{F}_0, \|\cdot\|_{\mathrm{F}}, \sqrt{r}a) \geq c_1 r\left(s_u \log\left(\frac{ep}{s_u}\right) + s_v \log\left(\frac{eq}{s_v}\right)\right),$$

$$\log \mathcal{M}(\mathcal{F}_0, \|\cdot\|_*, ra) \geq c_1 r\left(s_u \log\left(\frac{ep}{s_v}\right) + s_v \log\left(\frac{eq}{s_v}\right)\right).$$

Proof of Lemma 7. Let us focus on the first inequality. Other inequalities follow from similar arguments.

We first consider the case where $s_v = 1$. We start by constructing a subclass $\mathcal{F}_0$ of $\mathcal{F}$. Let $\mathbf{u} \in \mathbb{R}^p$ be an rs_u-sparse vector. We consider all such u's such that the supports $\mathcal{A}_i$, $\mathcal{A}_j$ of any two of them satisfies $|\mathcal{A}_i \cap \mathcal{A}_j| = 0$, that is, they are disjoint. Let $\mathcal{A}_1, \ldots, \mathcal{A}_N$ be a maximal set consisting subsets of $[p]$ with cardinality rs_u such that any two of them are disjoint. Then, we have at least

$$\log N \geq \log \left\{ \binom{p}{s_u} \binom{p - s_u}{s_u} \cdots \binom{p - (r-1)s_u}{s_u} \right\}.$$

Now since $2(r-1)(s_u \vee s_v) \leq p \wedge q$ and by the binomial coefficient bound that for any $k \leq p$

$$\left(\frac{p}{k}\right)^k \leq \binom{p}{k} \leq \left(\frac{ep}{k}\right)^k,$$

we obtain

$$\log N \geq \sum_{j=1}^{r} s_u \log \left(\frac{p}{2s_u} \right) \geq rs_u \log \left(\frac{p}{2s_u} \right) \gtrsim rs_u \log \left(\frac{ep}{s_u} \right),$$

where $\gtrsim$ in the last inequality means $\geq$ up to a universal constant.

Now denote the support of the first s_u nonzeros in $\mathbf{u}$ by $\mathcal{S}_1$, the support of the second s_u nonzeros in $\mathbf{u}$ by $\mathcal{S}_2, \cdots$, the support of the r-th s_u nonzeros in $\mathbf{u}$ by $\mathcal{S}_r$. Construct $\mathbf{u}_j$ such that $(\mathbf{u}_j)_{\mathcal{S}_j} = u_{\mathcal{S}_j} = (1/\sqrt{s_u}, \ldots, 1/\sqrt{s_u})^{\mathrm{T}}$, the vector of all $1/\sqrt{s_u}$'s, and $(\mathbf{u}_j)_{\mathcal{S}_j^c} = 0$, for $1 \leq j \leq r$. Let $\mathbf{v}_j = \mathbf{e}_j$ for $1 \leq j \leq d$, where $\mathbf{e}_j$ is the standard unit vector. Let $\mathbf{\Lambda} = \mathrm{diag}(a) \in \mathbb{R}^{r \times r}$. Let $\mathcal{F}_0$ be the family of all such $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$'s. It is easy to see that $\mathcal{F}_0 \subseteq \mathcal{F}$.

For every pair $\mathbf{A}_i, \mathbf{A}_j \in \mathcal{F}_0$, we have

$$\|\mathbf{A}_i - \mathbf{A}_j\|_{\mathrm{F}} = \sqrt{\|\mathbf{A}_i - \mathbf{A}_j\|_{\mathrm{F}}^2} = \sqrt{2r}a.$$

Let $\mathcal{R}_i, \mathcal{R}_j \subseteq [p]$ be the supports of $\mathbf{A}_i$ and $\mathbf{A}_j$'s nonzero rows. Since the nuclear norm is unitarily invariant, we can switch rows of $\mathbf{A}_i - \mathbf{A}_j$ such that it becomes

$$\begin{bmatrix} (\mathbf{A}_i)_{\mathcal{R}_i *} \\ -(\mathbf{A}_j)_{\mathcal{R}_j *} \\ 0 \end{bmatrix},$$

where $\mathbf{A}_{\mathcal{R}*}$ consists of the rows of $\mathbf{A}$ that are in $\mathcal{R}$. Then by Lemma 9, we have

$$\|\mathbf{A}_i - \mathbf{A}_j\|_* \geq \|(\mathbf{A}_i)_{\mathcal{R}_i^*}\|_* = ra.$$

Therefore we have

$$\log \mathcal{M}(\mathcal{F}_0, \|\cdot\|_{\mathrm{F}}, \sqrt{r}a) \geq c_1' r s_u \log\left(\frac{ep}{s_u}\right),$$
$$\log \mathcal{M}(\mathcal{F}_0, \|\cdot\|_*, ra) \geq c_1' r s_u \log\left(\frac{ep}{s_u}\right).$$

Similarly for the case of $s_u = 1$, we have

$$\log \mathcal{M}(\mathcal{F}_0, \|\cdot\|_{\mathrm{F}}, \sqrt{r}a) \geq c_1' r s_v \log\left(\frac{eq}{s_v}\right),$$
$$\log \mathcal{M}(\mathcal{F}_0, \|\cdot\|_*, ra) \geq c_1' r s_v \log\left(\frac{eq}{s_v}\right).$$

Combining both cases completes the proof.

□

Now let $\mathcal{F}_0$ be the same subclass as constructed in Lemma 7. Then for any $\mathbf{A}_1, \mathbf{A}_2 \in \mathcal{F}_0$, we have by Condition 3 with an arbitrary ξ and $m = rs_u$ that

$$D(P_{\mathbf{A}_1}\|P_{\mathbf{A}_2}) = \frac{1}{2}\|\mathbf{X}\mathbf{A}_1 - \mathbf{X}\mathbf{A}_2\|_{\mathrm{F}}^2 \leq \frac{1}{2}n\kappa_{\mathrm{upper}}\|\mathbf{A}_1 - \mathbf{A}_2\|_{\mathrm{F}}^2 \leq 2n\kappa_{\mathrm{upper}} ra^2.$$

Thus the KL diameter is $d_{\mathrm{KL}}(\mathcal{F}_0) \leq 2n\kappa_{\mathrm{upper}} nra^2$. In Lemma 6, take

$$\epsilon = \sqrt{r}a, \quad a^2 = \frac{c'(s_u \log(ep/s_u) + s_v \log(eq/s_v))}{n}.$$

where c' only depending on κ_{upper}. Now setting c' to be small enough such that

$$\frac{d_{\text{KL}}(T) + \log 2}{\log \mathcal{M}(T, \|\cdot\|_{\text{F}}, \epsilon)} \leq \frac{1}{2},$$

we obtain

$$\inf_{\widehat{A}} \sup_{A} \mathbb{E}\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{\text{F}}^2 \geq cr \left(s_u \log\left(\frac{ep}{s_u}\right) + s_v \log\left(\frac{eq}{s_v}\right)\right),$$

where c is a constant only depending on κ_{upper}. In a similar argument, we shall obtain

$$\inf_{\widehat{A}} \sup_{A} \mathbb{E}\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_*^2 \geq cr^2 \left(s_u \log\left(\frac{ep}{s_u}\right) + s_v \log\left(\frac{eq}{s_v}\right)\right).$$

where c is a constant only depending on κ_{upper}.

Combining the oracle case and sparse case, we finish the proof.

D.1 Technical Lemmas

We present two technical lemmas in the appendix. Our first result is a generalized Gordon's inequality.

Lemma 8 (A Generalized Gordon's Inequality)**.** Let $r \geq 1$ and $\mathbf{Z}_{\mathcal{A}} \in \mathbb{R}^{rs_u \times rs_v}$ be a random matrix with entires in $\mathcal{A}$ following i.i.d. standard normals while all entries in $\mathcal{A}^c$ are zeroes. We have

$$\mathbb{E}\|\mathbf{Z}_{\mathcal{A}}\|_2 \leq 4\sqrt{s_u \log r} + 4\sqrt{s_v \log r}.$$

Proof of Lemma 8. This is proved using the Slepian inequality and the properties of Chi-square distributions. The proof follows a similar strategy for proving Gordon's inequality and thus is omitted. □

The following lemma is taken from Bhatia (2013).

Lemma 9. For any unitarily invariant norm $\|\cdot\|$, we have

$$\|[\mathbf{A}, \mathbf{B}]\| \geq \|\mathbf{A}\| \vee \|\mathbf{B}\|.$$

E Proof of Lemmas in Appendix C

E.1 Proof of Lemma 2

Proof. To obtain an upper bound for $\|\nabla\mathcal{L}_\tau(\mathbf{A}^*)\|_{\infty,\infty}$, we first obtain an upper bound for a single element of the gradient and then use a union bound argument to obtain an upper bound for the max norm. Recall from (19) that $\mathcal{L}_\tau(\mathbf{A}^*) = \ell_\tau(\mathbf{Y} - \mathbf{X}\mathbf{A}^*)/n$ and note that $E_{ik} = Y_{ik} - \mathbf{X}_{i\cdot}^{\mathrm{T}}\mathbf{A}_{\cdot k}^*$, where $\mathbf{X}_{i\cdot}$ and $\mathbf{A}_{\cdot k}^*$ are the ith row of $\mathbf{X}$ and kth column of $\mathbf{A}^*$, respectively. Taking the gradient of $\mathcal{L}_\tau(\mathbf{A}^*)$ with respect to A_{jk}^*, we obtain

$$\{\nabla\mathcal{L}_\tau(\mathbf{A}^*)\}_{jk} = -\frac{1}{n}\sum_{i=1}^n X_{ij}\left\{E_{ik}1(|E_{ik}| \le \tau) + \tau 1(E_{ik} > \tau) - \tau 1(E_{ik} < -\tau)\right\}. \tag{31}$$

It remains to obtain an upper bound for (31). To this end, we define the quantity

$$\psi(u) = u1(|u| \le 1) + 1(u > 1) - 1(u < -1).$$

We will consider two cases: (i) $0 < \delta \le 1$ and (ii) $\delta > 1$. When $0 < \delta \le 1$, it can be verified that $\psi(u)$ has the following lower and upper bounds for all $u \in \mathbb{R}$

$$-\log\left(1 - u + |u|^{1+\delta}\right) \le \psi(u) \le \log\left(1 + u + |u|^{1+\delta}\right). \tag{32}$$

Using the notation $\psi(u)$, the gradient can be rewritten as

$$\{\nabla\mathcal{L}_\tau(\mathbf{A}^*)\}_{jk} = -\frac{\tau}{n}\sum_{i=1}^n X_{ij}\psi(E_{ik}/\tau).$$

Next, we obtain an upper bound for $X_{ij}\psi(E_{ik}/\tau)$. By (32), we have

$$\begin{aligned} X_{ij}\psi(E_{ik}/\tau) \le\ & 1(X_{ij} \ge 0)X_{ij}\log\left(1 + E_{ik}/\tau + |E_{ik}/\tau|^{1+\delta}\right) \\ & - 1(X_{ij} < 0)X_{ij}\log\left(1 - E_{ik}/\tau + |E_{ik}/\tau|^{1+\delta}\right). \end{aligned}$$

Since only one of the two terms on the upper bound is nonzero, we have

$$\begin{aligned}
&\exp\{X_{ij}\psi(E_{ik}/\tau)\}\\
&\leq \left(1+E_{ik}/\tau+|E_{ik}/\tau|^{1+\delta}\right)^{1(X_{ij}\geq 0)X_{ij}}+\left(1-E_{ik}/\tau+|E_{ik}/\tau|^{1+\delta}\right)^{-1(X_{ij}<0)X_{ij}}\\
&\leq 1+\left(E_{ik}/\tau+|E_{ik}/\tau|^{1+\delta}\right)X_{ij},
\end{aligned}$$

where the last inequality follows from the inequality $(1+u)^v \leq 1+uv$ for $u \geq -1$ and $0 < v \leq 1$. Using the above inequality, we obtain

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\sum_{i=1}^{n} X_{ij}\psi(E_{ik}/\tau)\right\}\right] &= \prod_{i=1}^{n}\mathbb{E}\left[\exp\left\{X_{ij}\psi(E_{ik}/\tau)\right\}\right]\\
&\leq \prod_{i=1}^{n}\mathbb{E}\left[\left\{1+(E_{ik}/\tau)X_{ij}+|E_{ik}/\tau|^{1+\delta}X_{ij}\right\}\right]\\
&\leq \prod_{i=1}^{n}\mathbb{E}\left[\left\{1+|E_{ik}/\tau|^{1+\delta}\right\}\right] \qquad (33)\\
&= \prod_{i=1}^{n}\left\{1+v_\delta/\tau^{1+\delta}\right\}\\
&\leq \exp\left(nv_\delta/\tau^{1+\delta}\right),
\end{aligned}$$

where the second inequality holds using the fact that $\mathbb{E}[E_{ik}]=0$ and that $\max_{i,j}|X_{ij}|=1$, and the last inequality holds by the fact that $1+u \leq \exp(u)$.

Recall that $\{\nabla\mathcal{L}_\tau(\mathbf{A}^*)\}_{jk} = -\tau n^{-1}\sum_{i=1}^{n} X_{ij}\psi(E_{ik}/\tau)$. By the Markov's inequality and (33), for any $z>0$, we have

$$\begin{aligned}
\mathbb{P}\left(-\{\nabla\mathcal{L}_\tau(\mathbf{A}^*)\}_{jk}\geq v_\delta\tau z\right) &= \mathbb{P}\left(\sum_{i=1}^{n} X_{ij}\psi(E_{ik}/\tau)\geq nv_\delta z\right)\\
&\leq \frac{\mathbb{E}\left\{\exp\left(\sum_{i=1}^{n} X_{ij}\psi(E_{ik}/\tau)\right)\right\}}{\exp(nv_\delta z)}\\
&\leq \exp\left\{-nv_\delta(z-\tau^{-(1+\delta)})\right\}\\
&\leq \exp\left\{-nv_\delta z/2\right\},
\end{aligned}$$

where the last inequality holds by picking $\tau \geq (2/z)^{1/(1+\delta)}$. Similarly, it can be shown that

$\mathbb{P}\left(\{\nabla \mathcal{L}_\tau(\mathbf{A}^*)\}_{jk} \geq v_\delta \tau z\right) \leq \exp\{-n v_\delta z/2\}$. Then, by the union bound, we have

$$\begin{aligned}\mathbb{P}\left(\|\nabla \mathcal{L}_\tau(\mathbf{A}^*)\|_{\infty,\infty} \geq v_\delta \tau z\right) &\leq \sum_{j=1}^{p}\sum_{k=1}^{q} \mathbb{P}\left(|\{\nabla \mathcal{L}_\tau(\mathbf{A}^*)\}_{jk}| \geq v_\delta \tau z\right) \\ &\leq 2pq\exp(-n v_\delta z/2).\end{aligned} \tag{34}$$

Picking $z = (6/v_\delta)\log(pq)/n$ and $\tau \geq \{(n v_\delta)/(3\log(pq))\}^{1/(1+\delta)}$, we obtain

$$\mathbb{P}\left(\|\nabla \mathcal{L}_\tau(\mathbf{A}^*)\|_{\infty,\infty} \geq v_\delta \tau z\right) \leq \frac{1}{pq},$$

implying

$$\|\nabla \mathcal{L}_\tau(\mathbf{A}^*)\|_{\infty,\infty} \leq 6^{\delta/(1+\delta)}(2v_\delta)^{1/(1+\delta)}\left(\frac{\log(pq)}{n}\right)^{\delta/(1+\delta)}$$

with probability at least $1-(pq)^{-1}$.

For $\delta > 1$, instead of the inequality in (32), we use

$$-\log\left(1-u+|u|^2\right) \leq \psi(u) \leq \log\left(1+u+|u|^2\right).$$

Following a similar argument, we arrive at

$$\|\nabla \mathcal{L}_\tau(\mathbf{A}^*)\|_{\infty,\infty} \leq 12^{1/2} v_\delta^{1/2}\left(\frac{\log(pq)}{n}\right)^{1/2}$$

with probability at least $1-(pq)^{-1}$. We obtain the desired results by combining both cases when $0 < \delta \leq 1$ and $\delta > 1$.

□

E.2 Proof of Lemma 3

Proof. Recall that $\mathcal{S}$ is the support of $\mathbf{A}^*$. Under the condition that $\|\nabla \mathcal{L}_\tau(\mathbf{A}^*)\|_{\infty,\infty} \leq \lambda/2$, we will show that

$$\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}^c}\|_{1,1} \leq \frac{2\gamma+5}{2\gamma-5}\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\|_{1,1}.$$

By the first order optimality condition of (4), there exists $\widetilde{\mathbf{N}} \in \partial \|\widehat{\mathbf{A}}\|_*$ and $\widetilde{\mathbf{\Gamma}} \in \partial \|\widehat{\mathbf{A}}\|_{1,1}$ such that

$$\langle \nabla \mathcal{L}_\tau(\widehat{\mathbf{A}}) + \lambda(\widetilde{\mathbf{N}} + \gamma \widetilde{\mathbf{\Gamma}}), \widehat{\mathbf{A}} - \mathbf{A}^* \rangle = 0. \tag{35}$$

From (21), we have $D^s_{\mathcal{L}}(\widehat{\mathbf{A}}, \mathbf{A}^*) = \langle \nabla \mathcal{L}_\tau(\widehat{\mathbf{A}}) - \nabla \mathcal{L}_\tau(\mathbf{A}^*), \widehat{\mathbf{A}} - \mathbf{A}^* \rangle \geq 0$, implying

$$\langle \nabla \mathcal{L}_\tau(\widehat{\mathbf{A}}), \widehat{\mathbf{A}} - \mathbf{A}^* \rangle \geq \langle \nabla \mathcal{L}_\tau(\mathbf{A}^*), \widehat{\mathbf{A}} - \mathbf{A}^* \rangle. \tag{36}$$

Substituting (36) into (35), we obtain

$$\langle \nabla \mathcal{L}_\tau(\mathbf{A}^*) + \lambda(\widetilde{\mathbf{N}} + \gamma \widetilde{\mathbf{\Gamma}}), \widehat{\mathbf{A}} - \mathbf{A}^* \rangle \leq 0,$$

or equivalently,

$$\underbrace{\langle \nabla \mathcal{L}_\tau(\mathbf{A}^*), \widehat{\mathbf{A}} - \mathbf{A}^* \rangle}_{\mathrm{I}_1} + \underbrace{\lambda \langle \widetilde{\mathbf{N}}, \widehat{\mathbf{A}} - \mathbf{A}^* \rangle}_{\mathrm{I}_2} + \underbrace{\lambda\gamma \langle \widetilde{\mathbf{\Gamma}}, \widehat{\mathbf{A}} - \mathbf{A}^* \rangle}_{\mathrm{I}_3} \leq 0, \tag{37}$$

It remains to obtain lower bounds for $\mathrm{I}_1, \mathrm{I}_2$, and I_3.

Lower bound for I_1: By the Holder's inequality and the condition that $\|\nabla \mathcal{L}_\tau(\mathbf{A}^*)\|_{\infty,\infty} \leq \lambda/2$, we can lower bound I_1 by

$$\mathrm{I}_1 \geq -\|\nabla \mathcal{L}_\tau(\mathbf{A}^*)\|_{\infty,\infty} \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} \geq -(\lambda/2)\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1}. \tag{38}$$

Lower bound for I_2: Similarly, by the Holder's inequality, we have

$$\mathrm{I}_2 \geq -\lambda \|\widetilde{\mathbf{N}}\|_{\infty,\infty} \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} \geq -\lambda \|\widetilde{\mathbf{N}}\|_2 \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} \geq -2\lambda \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1}, \tag{39}$$

were the second inequality holds using the fact that $\|\widetilde{\mathbf{N}}\|_{\infty,\infty} \leq \|\widetilde{\mathbf{N}}\|_2$ and the last inequality holds by $\|\widetilde{\mathbf{N}}\|_2 \leq 2$.

Lower bound for I_3: By the definition of the subgradient of an ℓ_1 norm, we have $\langle \widetilde{\mathbf{\Gamma}}, \widehat{\mathbf{A}} \rangle =$

$\|\widehat{\mathbf{A}}\|_{1,1}$ and that $\|\widetilde{\mathbf{\Gamma}}\|_{\infty,\infty} \leq 1$. Thus, we have

$$\begin{aligned}
\mathrm{I}_3 &= \lambda\gamma\langle \widetilde{\mathbf{\Gamma}}_{\mathcal{S}}, (\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\rangle + \lambda\gamma\langle \widetilde{\mathbf{\Gamma}}_{\mathcal{S}^c}, (\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}^c}\rangle \\
&\geq -\lambda\gamma\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\|_{1,1} + \lambda\gamma\langle \widetilde{\mathbf{\Gamma}}_{\mathcal{S}^c}, (\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}^c}\rangle \\
&\geq -\lambda\gamma\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\|_{1,1} + \lambda\gamma\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}^c}\|_{1,1},
\end{aligned} \tag{40}$$

where the second inequality follows from Holder's inequality and the last inequality follows from the fact that $\langle \widetilde{\mathbf{\Gamma}}_{\mathcal{S}^c}, \widehat{\mathbf{A}}_{\mathcal{S}^c}\rangle = \|\widehat{\mathbf{A}}_{\mathcal{S}^c}\|_{1,1}$ and that $\mathbf{A}^*_{\mathcal{S}^c} = \mathbf{0}$.

Substituting (38), (39), and (40) into (37), we obtain

$$-(\lambda/2)\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} - 2\lambda\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} - \lambda\gamma\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\|_{1,1} + \lambda\gamma\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}^c}\|_{1,1} \leq 0.$$

After rearranging the terms, we have

$$\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}^c}\|_{1,1} \leq \frac{2\gamma + 5}{2\gamma - 5}\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\|_{1,1}.$$

□

E.3 Proof of Lemma 4

Proof. From (35)–(38) in the proof of Lemma 3, there exists $\widetilde{\mathbf{N}} \in \partial\|\widehat{\mathbf{A}}\|_*$ and $\widetilde{\mathbf{\Gamma}} \in \partial\|\widehat{\mathbf{A}}\|_{1,1}$ such that

$$\langle \nabla\mathcal{L}_\tau(\mathbf{A}^*), \widehat{\mathbf{A}} - \mathbf{A}^*\rangle + \lambda\langle \widetilde{\mathbf{N}}, \widehat{\mathbf{A}} - \mathbf{A}^*\rangle + \lambda\gamma\langle \widetilde{\mathbf{\Gamma}}, \widehat{\mathbf{A}} - \mathbf{A}^*\rangle \leq 0.$$

Moreover, by monotonicity of subdifferentials of convex functions, $\langle -\lambda(\widetilde{\mathbf{N}} - \mathbf{N}), \widehat{\mathbf{A}} - \mathbf{A}^*\rangle \leq 0$, where $\mathbf{N} \in \partial\|\mathbf{A}^*\|_*$. Combining the above inequalities, we have

$$\underbrace{\lambda\langle \mathbf{N}, \widehat{\mathbf{A}} - \mathbf{A}^*\rangle}_{\mathrm{II}_1} + \underbrace{\lambda\gamma\langle \widetilde{\mathbf{\Gamma}}, \widehat{\mathbf{A}} - \mathbf{A}^*\rangle}_{\mathrm{II}_2} + \underbrace{\langle \nabla\mathcal{L}(\mathbf{A}^*), \widehat{\mathbf{A}} - \mathbf{A}^*\rangle}_{\mathrm{II}_3} \leq 0. \tag{41}$$

Lower bound for II_1: Recall the sub-differential of the nuclear norm in (20). From (20), the subdifferential $\mathbf{N}$ can be written as

$$\mathbf{N} = \mathbf{U}\mathbf{V}^{\mathrm{T}} + \mathcal{P}_{\mathcal{U}^\perp}\mathbf{W}\mathcal{P}_{\mathcal{V}^\perp}, \text{ where } \|\mathbf{W}\|_2 \leq 1.$$

We choose $\mathbf{W}$ such that $\langle \mathcal{P}_{\mathcal{U}^\perp}\mathbf{W}\mathcal{P}_{\mathcal{V}^\perp}, \widehat{\mathbf{A}} - \mathbf{A}^*\rangle = \|\mathcal{P}_{\mathcal{U}^\perp}\widehat{\mathbf{A}}\mathcal{P}_{\mathcal{V}^\perp}\|_*$, and this implies that

$$\begin{aligned}
\mathrm{II}_1 &= \lambda\big\langle \mathbf{U}\mathbf{V}^{\mathrm{T}} + \mathcal{P}_{\mathcal{U}^\perp}\mathbf{W}\mathcal{P}_{\mathcal{V}^\perp}, \widehat{\mathbf{A}} - \mathbf{A}^*\big\rangle \\
&= \lambda\big\langle \mathbf{U}\mathbf{V}^{\mathrm{T}}, \mathcal{P}_{\mathcal{U}}(\widehat{\mathbf{A}} - \mathbf{A}^*)\mathcal{P}_{\mathcal{V}}\big\rangle + \lambda\big\langle \mathcal{P}_{\mathcal{U}^\perp}\mathbf{W}\mathcal{P}_{\mathcal{V}^\perp}, \widehat{\mathbf{A}}\big\rangle \\
&\geq \lambda\big\|\mathcal{P}_{\mathcal{U}^\perp}\widehat{\mathbf{A}}\mathcal{P}_{\mathcal{V}^\perp}\big\|_* - \lambda\big\|\mathcal{P}_{\mathcal{U}}(\widehat{\mathbf{A}} - \mathbf{A}^*)\mathcal{P}_{\mathcal{V}}\big\|_*.
\end{aligned}$$

Lower bound for II_2: using a similar argument to the proof of Lemma 3, we have

$$\mathrm{II}_2 \geq -\lambda\gamma\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}}\|_{1,1} + \lambda\gamma\|(\widehat{\mathbf{A}} - \mathbf{A}^*)_{\mathcal{S}^c}\|_{1,1}.$$

Lower bound for II_3: using a similar argument to the proof of Lemma 3, we obtain that

$$\mathrm{II}_3 \geq -\big\|\nabla\mathcal{L}_\tau(\mathbf{A}^*)\big\|_{\infty,\infty}\big\|\widehat{\mathbf{A}} - \mathbf{A}^*\big\|_{1,1} \geq -\frac{\lambda}{2}\big\|\widehat{\mathbf{A}} - \mathbf{A}^*\big\|_{1,1}.$$

Therefore, combining the lower bounds for II_1, II_2 and II_3 into (41), we obtain

$$\begin{aligned}
&\lambda\big\|\mathcal{P}_{\mathcal{U}^\perp}(\widehat{\mathbf{A}} - \mathbf{A}^*)\mathcal{P}_{\mathcal{V}^\perp}\big\|_* - \lambda\big\|\mathcal{P}_{\mathcal{U}}(\widehat{\mathbf{A}} - \mathbf{A}^*)\mathcal{P}_{\mathcal{V}}\big\|_* - \lambda\gamma\big\|\big(\widehat{\mathbf{A}} - \mathbf{A}^*\big)_{\mathcal{S}}\big\|_{1,1} \\
&+ \lambda\gamma\big\|\big(\widehat{\mathbf{A}} - \mathbf{A}^*\big)_{\mathcal{S}^c}\big\| - (\lambda/2)\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{1,1} \leq 0.
\end{aligned}$$

By the assumption that $\gamma \geq 1/2$, the above equation simplifies to

$$\big\|\mathcal{P}_{\mathcal{U}^\perp}(\widehat{\mathbf{A}} - \mathbf{A}^*)\mathcal{P}_{\mathcal{V}^\perp}\big\|_* \leq \big\|\mathcal{P}_{\mathcal{U}}(\widehat{\mathbf{A}} - \mathbf{A}^*)\mathcal{P}_{\mathcal{V}}\big\|_* + (\gamma + 0.5)\big\|\big(\widehat{\mathbf{A}} - \mathbf{A}^*\big)_{\mathcal{S}}\big\|_{1,1}.$$

□

E.4 Proof of Lemma 5

Proof. Recall that

$$D^s_{\mathcal{L}}(\mathbf{A}, \mathbf{A}^*) = \langle \nabla\mathcal{L}_\tau(\mathbf{A}) - \nabla\mathcal{L}_\tau(\mathbf{A}^*), \mathbf{A} - \mathbf{A}^*\rangle.$$

Let $\mathbf{\Delta} = \mathbf{A} - \mathbf{A}^*$. It can be shown that

$$D^s_{\mathcal{L}}(\mathbf{A}, \mathbf{A}^*) \geq \kappa_-(\mathbf{H}_\tau(\mathbf{A}), \xi, \eta)\|\mathbf{A} - \mathbf{A}^*\|_{\mathrm{F}}^2.$$

By Lemma 1, we have $\kappa_-(\mathbf{H}_\tau(\mathbf{A}), \xi, \eta) \geq \kappa_{\text{lower}}/2$ with probability $1 - (pq)^{-1}$. Thus,

$$D_{\mathcal{L}}^s(\mathbf{A}, \mathbf{A}^*) \geq \frac{\kappa_{\text{lower}}}{2} \|\mathbf{A} - \mathbf{A}^*\|_{\mathrm{F}}^2.$$

□

F Technical Lemmas

Lemma 10 (Hoeffding's Inequality)**.** Let $Z_1, \ldots, Z_n$ be independent random variables such that $\mathbb{E}(Z_i) = \mu$ and $a \leq Z_i \leq b$. Then, for any $z > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i \geq z + \mu\right) \leq \exp(-2nz^2/(b-a)^2).$$

Lemma 11. Let $X_1, \ldots, X_n$ be independent random variables with

$$\mathbb{E}(X_i) = 0 \qquad \text{and} \qquad v_\delta = \max_i \; \mathbb{E}(|X_i|^{1+\delta}) < \infty \text{ for } \delta > 0.$$

For any $t \geq 0$ and $\tau > 0$, we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n 1(|X_i| > \tau/2) \geq (2/\tau)^{1+\delta} v_\delta + \sqrt{t/n}\right) \leq \exp(-2t).$$

Proof. We first obtain an upper bound for $\mathbb{E}(n^{-1}\sum_{i=1}^n 1(|X_i| > \tau/2))$. By the Markov's inequality, we have

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n 1(|X_i| > \tau/2)\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{P}(|X_i| > \tau/2) = \frac{1}{n}\sum_{i=1}^n \mathbb{P}\left(|X_i|^{1+\delta} > (\tau/2)^{1+\delta}\right) \leq (2/\tau)^{1+\delta} v_\delta.$$

Let $Z_i = 1(|X_i| > \tau/2)$, $\mu = E(Z_i)$, and $z = \sqrt{t/n}$. Note that $0 \leq Z_i \leq 1$. By Lemma 10, we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n 1(|X_i| > \tau/2) \geq (2/\tau)^{1+\delta} v_\delta + \sqrt{t/n}\right) \leq \exp(-2t),$$

as desired. □