Supplementary Document for Adaptive Estimation in Two-way Sparse Reduced-rank Regression

Zhuang Ma, Zongming Ma and Tingni Sun

University of Pennsylvania and University of Maryland

We provide the technical proofs for all theorems in this supplementary document.

6. Proofs

6.1 Proof of Theorem 1

We analyze each step of Algorithm 1 to prove Theorem 1. Throughout the proof, some useful lemmas on tail probabilities will be stated without proof.

Analysis of $V_{(0)}$. We first study the property of the right singular vector matrix $V^{(0)}$ obtained in the column-thresholding step of Stage I. For $0 < a_- < 1 < a_+$, define

$$J_{(0)}^{\pm} = \Big\{ j : \|XA_{*j}\|^2 \geq \widetilde{\sigma}^2 a_{\mp} \alpha \sqrt{n \log(p \vee m)} \Big\}.$$

More specifically, let $a_- = 0.1$ and $a_+ = 2$ in the proof. Recall that $\alpha = \sqrt{12}$ and $\widetilde{\sigma} = 2\sigma$.

Lemma 1. *[Stage I column selection] With probabbility at least* $1 - 4(p \vee m)^{-2}$,

$$J_{(0)}^- \subset J_{(0)} \subset J_{(0)}^+$$

Proof of Lemma 1. Due to Gaussianity, $\|Y_{*j}^{(0)}\|^2/\widetilde{\sigma}^2$ follows a non-central χ^2 distribution with n degrees of freedom and noncentrality parameter $\|XA_{*j}\|^2/\widetilde{\sigma}^2$. By Lemma 2,

$$\begin{aligned}
P(J_{(0)}^- \not\subset J_{(0)}) &\leq \sum_{j \in J_{(0)}^-} P\Big\{\|Y_{*j}^{(0)}\|^2 < \widetilde{\sigma}^2(n + \alpha\sqrt{n\log(p \vee m)})\Big\} \\
&\leq mP\Big\{\|Y_{*j}^{(0)}\|^2 < \widetilde{\sigma}^2 n + \|XA_{*j}\|^2 - \widetilde{\sigma}^2\alpha(a_+ - 1)\sqrt{n\log(p \vee m)} \;\Big|\; j \in J_{(0)}^-\Big\} \\
&\leq 2m\exp\Big(-\frac{\alpha^2(a_+ - 1)^2 n\log(p \vee m)}{4(\sqrt{n} + (a_+\alpha)^{1/2}(n\log(p \vee m))^{1/4})^2}\Big) \\
&\leq 2(p \vee m)^{-2}.
\end{aligned}$$

Similarly, it is proved that $J_{(0)} \subset J_{(0)}^+$ holds with probability at least $1 - 2(p \vee m)^{-2}$.

□

Lemma 2. *Let* X *follow a non-central chi-square distribution* $\chi_n^2(\lambda)$ *with*

n degrees of freedom and non-centrality parameter λ. Then

$$P\Big\{X \geq (n+\lambda) + 2(\sqrt{n}+\sqrt{\lambda})s\Big\} \leq \Big(1 + \frac{1}{\sqrt{2}s}\Big)\exp(-s^2), \quad \textit{if } 0 \leq s \leq \frac{1}{2}n^{9/16},$$
$$P\Big\{X \leq (n+\lambda) - 2(\sqrt{n}+\sqrt{\lambda})s\Big\} \leq 2\exp(-s^2), \quad \textit{if } 0 \leq s \leq \frac{1}{2}n^{1/2}.$$

Lemma 3. *Let X be an $n \times m$ matrix with iid standard Gaussian entries. Then for any $t > 0$,*

$$P\Big\{\|X\| > \sqrt{n} + \sqrt{m} + t\Big\} \leq \exp(-t^2/2).$$

Lemma 4. *[Stage I subspace estimation] With probability at least $1 - 3(p \vee m)^{-2}$,*

$$\begin{aligned}
\|VV' - V_{(0)}{V_{(0)}}'\| &\leq \frac{C_1\widetilde{\sigma}}{d}\Big\{\sqrt{n} + \sqrt{k} + 2\sqrt{\log(p \vee m)} + \sqrt{k\sqrt{n\log(p \vee m)}}\Big\}, \\
\|VV' - V_{(0)}{V_{(0)}}'\|_F &\leq \frac{C_2\widetilde{\sigma}}{d}\Big\{\sqrt{r}(\sqrt{n} + \sqrt{k} + 2\sqrt{\log(p \vee m)}) + \sqrt{k\sqrt{n\log(p \vee m)}}\Big\}.
\end{aligned}$$

Proof of Lemma 4. We study the upper bounds in the event where $J_{(0)}^{-} \subset J_{(0)} \subset J_{(0)}^{+}$ holds. We may reorder the columns of matrices such that $XA - \widetilde{Y}^{(0)}$ is of the following form

$$XA - \widetilde{Y}^{(0)} = \begin{pmatrix} -Z_{*J_{(0)}} & UDV'_{*J_{(0)}^c} \end{pmatrix}$$

Lemma 3 provides an upper bound for $\|Z_{*J_{(0)}}\|$ as follows

$$\|Z_{*J_{(0)}}\| \leq \widetilde{\sigma}(\sqrt{n} + \sqrt{J_{(0)}} + 2\sqrt{\log(p \vee m)}) \leq \widetilde{\sigma}(\sqrt{n} + \sqrt{k} + 2\sqrt{\log(p \vee m)})$$

with probability at least $1 - (p \vee m)^2$, since $|J_{(0)}| \leq |J_{(0)}^+| = k$. Moreover, it holds that, in the event of $J_{(0)}^- \subset J_{(0)}$,

$$\|U\Delta V'_{*J_{(0)}^c}\|^2 \leq \|\Delta V'_{*(J_{(0)}^-)^c}\|_F^2 \leq \widetilde{\sigma}^2 a_- \alpha k \sqrt{n \log(p \vee m)}.$$

Thus, we have

$$\|XA - \widetilde{Y}^{(0)}\| \leq \widetilde{\sigma}(\sqrt{n} + \sqrt{k} + 2\sqrt{\log(p \vee m)}) + \widetilde{\sigma}\sqrt{a_- \alpha k \sqrt{n \log(p \vee m)}}$$

and the desired results then follows from the $\sin\theta$ theorem. □

Analysis of $U_{(1)}$.

Lemma 5. *[Stage I Regression] Under the condition of Theorem 1, there exists a constant C depending only on $\kappa_\pm(s_*), c_*$ and c_0, such that with probability at least $1 - (p \vee m)^{-1}$,*

$$\|U_{(1)}U'_{(1)} - UU'\|_F \leq C\sqrt{s}\lambda/d.$$

Proof of Lemma 5. Let $U_* \in \mathbb{R}^{n\times r}$ be the left singular vector matrix of $XAV_{(0)} = UDV'V_{(0)}$. Under condition (12), $V'V_{(0)}$ is an $r \times r$ matrix of full rank, and so the column space of U^* is the same as the column space of U; i.e., $U_*U_*' = UU'$. By Wedin's $\sin\theta$ Theorem (Wedin, 1972),

$$\|U_{(1)}U_{(1)}' - UU'\|_F = \|U_{(1)}U_{(1)}' - U_*U_*'\|_F \leq \frac{\|XB_{(1)} - XAV_{(0)}\|_F}{\sigma_r(XAV_{(0)})},$$

where $\sigma_r(XAV_{(0)})$ is the r^{th} singular value of $XAV_{(0)}$.

Since for any unit vector x,

$$\begin{aligned}
\|V'V_{(0)}x\|^2 &= x'V_{(0)}'VV'V_{(0)}x \\
&= 1 - x'V_{(0)}'(VV' - V_{(0)}V_{(0)}')V_{(0)}x \\
&\geq 1 - \|VV' - V_{(0)}V_{(0)}'\|.
\end{aligned}$$

Thus, we have $\sigma_r^2(V'V_{(0)}) = \min_{\|x\|=1}\|V'V_{(0)}x\|^2 \geq 1 - \|VV' - V_{(0)}V_{(0)}'\|$. When c_0 is small enough, $\|VV' - V_{(0)}V_{(0)}'\|$ is sufficiently small by Lemma 4. So there exists a constant c' such that $\sigma_r(V'V_{(0)}) > c'$. Note that $XAV_{(0)} = XAVV'V_{(0)}$, and so

$$\sigma_r(XAV_{(0)}) \geq \sigma_r(XAV)\sigma_r(V'V_{(0)}) \geq \delta_r c',$$

where the last inequality holds under condition (12) since $\sigma_r(XAV) =$

$\sigma_r(XA) = \delta_r$. Further note that

$$\|XB_{(1)} - XAV_{(0)}\|_F \leq \kappa_+(2s)\|B_{(1)} - AV_{(0)}\|_F \leq \kappa_+(s_*)\|B_{(1)} - AV_{(0)}\|_F$$

and that $\delta_r \geq \kappa_-(s)\sigma_r(A) \geq \kappa_-(s_*)d$, the desired result then follows from Part (ii) of Theorem 3 with $\eta = 1/(p \vee m)$. □

Analysis of $V_{(1)}$. Recall

$$J_{(1)} = J_{(0)} \cup \left\{ j : \|U_{(1)}{}'Y^{(2)}_{*j}\|^2 \geq \beta\widetilde{\sigma}^2(r + 2\sqrt{3r\log(p \vee m)} + 6\log(p \vee m)) \right\}.$$

For $b_- < b_+$, define

$$J^{\pm}_{(1)} = \left\{ j : \|XA_{*j}\|^2 \geq \widetilde{\sigma}^2 b_{\mp}(r + 2\sqrt{3r\log(p \vee m)} + 6\log(p \vee m)) \right\}.$$

More specifically, let $b_+ = 4.5$ and $b_- = 0.002$ in the proof. Recall that $\beta = 1.1$.

Lemma 6. *Let X follow a chi-square distribution χ^2_n with n degrees of freedom. Then for any $t > 0$*

$$P(X > n + 2\sqrt{n}t + 2t^2) < \exp(-t^2)$$

Lemma 7. *[Stage II column selection] Assume* $\|U_{(1)}U'_{(1)} - UU'\| < c$ *for some small positive constant* $c < 0.05$. *With probabbility at least* $1 - 2(p \vee m)^{-2}$,

$$J^-_{(1)} \subset J_{(1)} \subset J^+_{(1)}$$

Proof of Lemma 7. For $j \in J^-_{(1)} \setminus J_{(0)}$,

$$\begin{aligned} \|U'_{(1)}Y^{(2)}_{*j}\| &= \|U'_{(1)}(UDV'_{*j} + Z^{(2)}_{*j})\| \\ &\geq \|U'_{(1)}UDV'_{*j}\| - \|U'_{(1)}Z^{(2)}_{*j}\| \end{aligned}$$

The first term is

$$\begin{aligned} \|U'_{(1)}UDV'_{*j}\|^2 &\geq \|XA_{*j}\|^2(1 - \|U_{(1)}U'_{(1)} - UU'\|) \geq \|XA_{*j}\|^2(1-c) \\ &\geq \widetilde{\sigma}^2(1-c)b_+(r + 2\sqrt{3r\log(p \vee m)} + 6\log(p \vee m)) \end{aligned}$$

Since $U'_{(1)}Z^{(2)}_{*j} \sim N(0, \widetilde{\sigma}^2 I_r)$, it follows from Lemma 6 that

$$\|U'_{(1)}Z^{(2)}_{*j}\|^2 \leq \widetilde{\sigma}^2(r + 2\sqrt{3r\log(p \vee m)} + 6\log(p \vee m)),$$

with probability at least $1-(p\vee m)^{-3}$. Thus, in the same event, we have

$$\begin{aligned}\|U'_{(1)}Y^{(2)}_{*j}\| &\geq (\sqrt{(1-c)b_+}-1)\widetilde{\sigma}\Big\{r+2\sqrt{3r\log(p\vee m)}+6\log(p\vee m)\Big\}^{1/2}\\ &\geq \beta^{1/2}\widetilde{\sigma}(r+2\sqrt{3r\log(p\vee m)}+6\log(p\vee m))^{1/2},\end{aligned}$$

due to $(\sqrt{(1-c)b_+}-1)^2>(\sqrt{0.95\times 4.5}-1)^2>1.1=\beta$. Hence, we have $j\in J_{(1)}$. So it holds that $J^-_{(1)}\subset J_{(1)}$ with probability at least $1-(p\vee m)^{-2}$. Similarly, we have $J_{(1)}\subset J^+_{(1)}$ with probability at least $1-(p\vee m)^{-2}$, due to $(\sqrt{(1+c)b_-}+1)^2<1.1=\beta$. □

Lemma 8. *[Stage II subspace estimation] Suppose $\|U_{(1)}U'_{(1)}-UU'\|_F<c'_1$ for a sufficiently small positive constant c'_1. Then there exists a constant C depending only on $\kappa_\pm(s_*),\gamma$ and c'_1 such that with probability at least $1-(p\vee m)^{-1}$,*

$$\|V_{(1)}V'_{(1)}-VV'\|_F\leq C\sigma\sqrt{(k+s)(r+\log(p\vee m))}/d$$

Proof of Lemma 8.

$$\|V_{(1)}V'_{(1)}-VV'\|_F\leq\frac{\|U_{(1)}U'_{(1)}\widetilde{Y}^{(1)}-U_{(1)}U'_{(1)}XA\|_F}{\sigma_r(U_{(1)}U'_{(1)}XA)}. \tag{14}$$

We first upper bound the numerator

$$
\begin{aligned}
& \|U_{(1)}U'_{(1)}\widetilde{Y}^{(1)} - U_{(1)}U'_{(1)}XA\|_F \\
\leq\ & \|U'_{(1)}(\widetilde{Y}^{(1)}_{*J_{(1)}} - XA_{*J_{(1)}})\|_F + \|U_{(1)}U'_{(1)}XA_{*J^c_{(1)}}\|_F \\
\leq\ & \|U'_{(1)}(\widetilde{Y}^{(1)}_{*J^{(1)}} - XA_{*J_{(1)}})\|_F + \|(U_{(1)}U'_{(1)} - UU')XA_{*J^c_{(1)}}\|_F + \|UU'XA_{*(J^-_{(1)})^c}\|_F \\
\leq\ & \widetilde{\sigma}(\sqrt{rk} + \sqrt{\log(p\vee m)}) + d\|(U_{(1)}U'_{(1)} - UU'\| + \widetilde{\sigma}\sqrt{k}\sqrt{b_+(r + 2\sqrt{3r\log(p\vee m)} + 6\log(p\vee m))} \\
\leq\ & C\sigma\sqrt{(k+s)(r+\log(p\vee m))} \qquad (15)
\end{aligned}
$$

To lower bound the denominator, we apply Weyl's theorem to obtain

$$
\begin{aligned}
\sigma_r(U_{(1)}U'_{(1)}XA) &\geq \sigma_r(UU'XA) - \|U_{(1)}U'_{(1)}XA - UU'XA\|_{\mathrm{op}} \\
&\geq \delta_r - \|U_{(1)}U'_{(1)} - UU'\|_{\mathrm{op}}\|XA\|_{\mathrm{op}}.
\end{aligned}
$$

Note that $\delta_r \geq \kappa_-(s_*)d$, $\|XA\|_{\mathrm{op}} \leq \kappa_+(s_*)\gamma d$ and that $\|U_{(1)}U'_{(1)} - UU'\|_{\mathrm{op}} \leq \|U_{(1)}U'_{(1)} - UU'\|_F \leq c'_1$. Thus, for sufficiently small value of c'_1, we obtain

$$
\sigma_r(U_{(1)}U'_{(1)}XA) \geq C^{-1}d, \qquad (16)
$$

where $C > 0$ is a constant depending only on $\kappa_\pm(s_*), \gamma$ and c'_1. Combining (14) – (16), we complete the proof. □

Proof of Theorem 1.

Proof. By the definition of $\widehat{A}$, we have

$$\begin{aligned}\|\widehat{A} - A\|_F &= \|B_{(2)}V_{(1)}' - AVV'\|_F \\ &\leq \|B_{(2)}V_{(1)}' - AV_{(1)}V_{(1)}'\|_F + \|AV_{(1)}V_{(1)}' - AVV'\|_F \\ &\leq \|V_{(1)}\|_{\mathrm{op}}\|B_{(2)} - AV_{(1)}\|_F + \|A\|_{\mathrm{op}}\|V_{(1)}V_{(1)}' - VV'\|_F.\end{aligned}$$

Assembling the bounds in all lemmas,

$$\|\widehat{A} - A\|_F^2 \lesssim \sigma^2(k+s)(r + \log(p \vee m)) \tag{17}$$

The desired upper bound on other Schatten norm losses is a consequence of (17) and the inequality $\|\widehat{A} - A\|_{\mathrm{s}_q}^2 \leq (2r)^{2/q-1}\|\widehat{A} - A\|_F^2$ for all $q \in [1, 2]$.

□

6.2 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 $\mathbb{R}^{m\times n}$, the volume of K is $\mathrm{vol}(K) = \int_K \mathrm{d}\mu$ where $\mathrm{d}\mu$ is the usual Lebesgue measure on $\mathbb{R}^{m\times n}$ by taking the product measure of the Lebesgue measures of individual entries. With these definitions, we state the following variant of Fano's lemma (Ibragimov and Has'minskii, 1981; Birgé, 1983; Tsybakov, 2009). This version has been established as Proposition 1 in Ma and Wu (2015). It will be

used repeatedly in the proof of the lower bounds. Throughout the proof, we denote $\kappa_+(2s)$ by κ_+.

Proposition 1. *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 in ρ is at least ϵ. Define the* Kullback-Leibler diameter *of T by*

$$d_{\mathrm{KL}}(T) \triangleq \sup_{\theta,\theta' \in T} D(P_\theta \,||\, P_{\theta'}). \tag{18}$$

Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\rho^2(\hat{\theta}(X), \theta)] \geq \sup_{T \subset \Theta} \sup_{\epsilon > 0} \frac{\epsilon^2}{4} \left(1 - \frac{d_{\mathrm{KL}}(T) + \log 2}{\log \mathcal{M}(T, \rho, \epsilon)}\right). \tag{19}$$

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

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\|\hat{\theta}(X) - \theta\|^2] \geq \sup_{T \subset \Theta} \sup_{\epsilon > 0} \frac{\epsilon^2}{4} \left(1 - \frac{d_{\mathrm{KL}}(T) + \log 2}{\log \frac{\mathrm{vol}(T)}{\mathrm{vol}(B_{\|\cdot\|}(\epsilon))}}\right). \tag{20}$$

We first prove an oracle version of the lower bound. One can think of it as an lower bound for the minimax risk when we know that the nonzero entries of the coefficient matrix $A \in \mathbb{R}^{p \times m}$ are restricted to the top–left $s \times r$ block (or the top left $r \times k$ block).

Lemma 9. *Let $\Theta_0(s, r, r, d, \gamma) \subset \Theta(s, k, r, d, \gamma)$ be the sub-collection of all matrices whose nonzero entries are in the top left $s \times r$ block. Suppose $\sigma = 1$. There exists a positive constant c that depends only on κ_+ and γ, such that for any $q \in [1, 2]$, the minimax risk for estimating A over Θ_0 satisfies*

$$\inf_{\widehat{A}} \sup_{\Theta_0} \mathbb{E} L_q(A, \widehat{A}) \geq c \left[(r^{2/q-1} d^2) \wedge (r^{2/q} s) \right].$$

Similarly, let $\Theta_0'(r, k, r, d, \gamma) \subset \Theta(s, k, r, d, \gamma)$ be the sub-collection of all matrices whose nonzero entries are in the top left $r \times k$ block. Under the same conditions, we have

$$\inf_{\widehat{A}} \sup_{\Theta_0'} \mathbb{E} L_q(A, \widehat{A}) \geq c \left[(r^{2/q-1} d^2) \wedge (r^{2/q} k) \right].$$

Proof. In what follows, we focus on proving the first claim and the second claim follows from essentially the same argument.

By a simple sufficiency argument, we can reduce to model (1) with $p = s$ and $m = r$, which we assume in the rest of this proof without loss of generality.

Let $A_0 = \mathrm{diag}(1, \ldots, 1) \in \mathbb{R}^{s \times r}$. Moreover, for any δ and any $q \in [1, 2]$, let $B_{\mathrm{S}_q}(\delta) = \{A \in \mathbb{R}^{s \times r} : \|A\|_{\mathrm{s}_q} \leq \delta\}$ denote the Schatten-q ball with radius

δ in $\mathbb{R}^{s\times r}$. For some constant $a > 0$ to be specified later, define

$$T(a) = \frac{\gamma d}{2} A_0 + B_{\mathrm{S}_2}(\sqrt{a}) = \left\{ \frac{\gamma d}{2} A_0 + M : M \in B_{\mathrm{S}_2}(\sqrt{a}) \right\}. \tag{21}$$

For any $A_1, A_2 \in T(a)$, we have

$$D(P_{A_1} || P_{A_2}) = \frac{1}{2} \|XA_1 - XA_2\|_{\mathrm{S}_2}^2 \le \frac{1}{2} \|X\|_{\mathrm{op}}^2 \|A_1 - A_2\|_{\mathrm{S}_2}^2 \le 2\kappa_+^2 a.$$

Here, the last inequality holds since $\|X\|_{\mathrm{op}} \le \kappa_+$ under the assumption that $X \in \mathbb{R}^{s\times r}$ and $\|A_1 - A_2\|_{\mathrm{S}_2}^2 \le 4a$ by definition (21). So

$$d_{\mathrm{KL}}(T(a)) \le 2\kappa_+^2 a. \tag{22}$$

By the inverse Santalo's inequality (see, e.g., Lemma 3 of Ma and Wu (2015)), for some universal constants c_0,

$$\begin{aligned} \mathrm{vol}(T(a))^{\frac{1}{sr}} &= \mathrm{vol}(B_{\mathrm{S}_2}(\sqrt{a}))^{\frac{1}{sr}} = \sqrt{a} \cdot \mathrm{vol}(B_{\mathrm{S}_2}(1))^{\frac{1}{sr}} \\ &\ge \sqrt{a} \cdot \frac{c_0}{\mathbb{E}\,\|Z\|_{\mathrm{S}_2}} \qquad (23) \\ &\ge \sqrt{a} \cdot \frac{c_0'}{\sqrt{sr}}. \qquad (24) \end{aligned}$$

In (23), Z is a $s \times r$ matrix with i.i.d. $N(0,1)$ entries. The inequality in (24) holds since by Jensen's inequality, $\mathbb{E}\,\|Z\|_{\mathrm{S}_2} \le \sqrt{\mathbb{E}\,\|Z\|_{\mathrm{S}_2}^2} = \sqrt{sr}$.

On the other hand, by Urysohn's inequality (see, e.g., Eq.(19) of Ma and Wu (2015)), for any $\epsilon > 0$ and $q \in [1, 2]$,

$$\mathrm{vol}(B_{\mathrm{S}_q}(\epsilon))^{\frac{1}{sr}} \le \frac{\epsilon \mathbb{E}\,\|Z\|_{\mathrm{S}_{q'}}}{\sqrt{sr}} \le \frac{\epsilon r^{\frac{1}{q'}} \mathbb{E}\|Z\|_{\mathrm{op}}}{\sqrt{sr}} \le 2\epsilon r^{\frac{1}{2}-\frac{1}{q}}.$$

Here, $\frac{1}{q'} + \frac{1}{q} = 1$ and Z is a $s \times r$ matrix with i.i.d. $N(0, 1)$ entries. The last inequality is due to Gordon's inequality (see, e.g., Davidson and Szarek (2001)): $\mathbb{E}\|Z\|_{\mathrm{op}} \le \sqrt{s} + \sqrt{r} \le 2\sqrt{s}$.

Now let

$$a = \left(\frac{\gamma \wedge 2 - 1}{2}\right)^2 \left(sr \wedge d^2\right), \quad \text{and} \quad \epsilon = \frac{c_0'}{2\kappa_+} \sqrt{a}\, r^{\frac{1}{q}-\frac{1}{2}}. \tag{25}$$

Then for any $A \in T(a)$ and any $i \in [r]$, $|\sigma_i(A) - \frac{\gamma}{2}d| \le \sqrt{a} \le \frac{\gamma \wedge 2 - 1}{2} d$, and so $\sigma_i(A) \in [d, \gamma d]$ and $T(a) \subset \Theta_0(s, r, d, \gamma)$. Applying Proposition 1 with $T(a)$ and ϵ in (21) and (25), we obtain a lower bound on the order of ϵ^2. This completes the proof. □

Lemma 10. *Let $s \ge r$ be positive integers. There exist a matrix $W \in \mathbb{R}^{s \times r}$ and two absolute constants $c_0 \in (\frac{1}{2}, 1)$ and $c_1 > 0$ such that $\|W\|_{\mathrm{F}} \le 1$ and for any subset $B \subset [s]$ such that $|B| \ge c_0 s$, $\|W_{B*}\|_{\mathrm{s}_q} \ge c_1 r^{\frac{1}{q}-\frac{1}{2}}$ for any $q \in [1, 2]$.*

Proof. We divide the proof into two cases, namely when $s \ge 25$ and when

$s < 25$.

1° When $s \geq 25$, let $Z \in \mathbb{R}^{s\times r}$ have i.i.d. $N(0,1)$ entries. Then $\|Z\|_{\mathrm{F}}^2 \sim \chi_{sr}^2$, and Laurent and Massart (2000, Eq.(4.3)) implies that

$$\mathbb{P}\left\{\|Z\|_{\mathrm{F}}^2 \geq sr + 2s\sqrt{r} + 2s\right\} \leq \mathrm{e}^{-s}.$$

Moreover, for any $c_0 > \frac{1}{2}$,

$$\begin{aligned}
&\mathbb{P}\left\{\exists B \subset [s], \text{ s.t. } |B| = c_0 s \text{ and } \sigma_r(Z_{B*}) < \sqrt{c_0 s} - \sqrt{r} - \frac{1}{2}\sqrt{c_0 s}\right\} \\
&\leq \sum_{B\subset[s],|B|=c_0 s} \mathbb{P}\left\{\sigma_r(Z_{B*}) < \sqrt{c_0 s} - \sqrt{r} - \frac{1}{2}\sqrt{c_0 s}\right\} \\
&\leq \binom{s}{(1-c_0)s} \mathrm{e}^{-c_0 s/4} \\
&\leq \exp\left\{-s\left[\frac{c_0}{4} + (1-c_0)\log(1-c_0)\right]\right\}.
\end{aligned}$$

Here, the first inequality is due to the union bound, the second inequality is due to the Davidson-Szarek bound, and the last inequality holds since for any $\alpha \in (\frac{1}{2}, 1)$, $\binom{s}{\alpha s} = \binom{s}{(1-\alpha)s} \leq (\frac{\mathrm{e}}{1-\alpha})^{(1-\alpha)s}$. If we set $c_0 \geq 0.96$, then the multiplier $\frac{c_0}{4} + (1-c_0)\log(1-c_0) \geq 0.1$.

So when $c_0 = 0.96$ and $s \geq 25$, the sum of the right hand sides of the last two displays is less than 1. Thus, there exists a deterministic matrix Z_0 on which both events happen. Now define $W = Z_0/\|Z_0\|_{\mathrm{F}}$. Then $\|W\|_{\mathrm{F}} = 1$

by definition, and for any $B \subset [s]$ with $|B| = c_0 s$,

$$\begin{aligned}
\|W_{B*}\|_{\mathrm{s}_q} &\geq r^{1/q} \sigma_r(W_{B*}) \\
&= r^{1/q} \sigma_r((Z_0)_{B*}) / \|Z_0\|_{\mathrm{F}} \\
&\geq r^{1/q} \frac{\frac{1}{2}\sqrt{c_0 s} - \sqrt{r}}{\sqrt{sr + 2s\sqrt{r} + 2r}} \\
&\geq c_1 r^{1/q - 1/2}.
\end{aligned}$$

Note that the last inequality holds with an absolute constant c_1 when $r \leq \frac{1}{8} c_0 s$. When $r > \frac{1}{8} c_0 s$, we can always let $\tilde{r} = \frac{1}{8} c_0 r \leq \frac{1}{8} c_0 s$ and repeat the above arguments on the $s \times \tilde{r}$ submatrix of Z consisting of its first $\tilde{r}$ columns, and the conclusion continues to hold with a modified absolute constant c_1. This completes the proof for all subsets B with $|B| = c_0 s$. The claim continues to hold for all $|B| \geq c_0 s$ since the Schatten-q norm of a submatrix is always no smaller than the the whole matrix.

2° When $s < 25$, we have $r < 25$ since $r \leq s$ always holds. Let $W = \left[\frac{1}{\sqrt{s}} \mathbf{1}_s \quad \mathbf{0} \right] \in \mathbb{R}^{s \times r}$, i.e., the first column of W consists of s entries all equal to $1/\sqrt{s}$ and the rest are all zeros. So W is rank one. It is straightforward to verify the desired conclusion holds since for any $B \subset [s]$, $\|W_{B*}\|_{\mathrm{s}_q} = \|W_{B*}\|_{\mathrm{F}} = \sqrt{|B|/s}$. This completes the proof. □

Lemma 11. *Let $a = d^2 \wedge s \log \frac{\mathrm{e}p}{s}$. There exist three positive constants c_1, c_2, c_3 that depend only on γ and κ_+, and a subset $\Theta_1 \subset \Theta(s, k, r, d, \gamma)$,*

such that $c_3 \le c_2/3$, $d_{\mathrm{KL}}(\Theta_1) \le c_3 a$ *and that for any* $q \in [1,2]$,

$$\log \mathcal{M}(\Theta_1, \|\cdot\|_{\mathrm{s}_q}, c_1\sqrt{a}\, r^{1/q-1/2}) \ge c_2 s \log \frac{\mathrm{e}p}{s},$$

where d_{KL} *is the Kullback–Leibler diameter and* $\mathcal{M}$ *is the packing number defined in Proposition 1.*

Similarly, for $b = d^2 \wedge k \log \frac{\mathrm{e}m}{k}$, *there is another subset* $\Theta' \subset \Theta(s,k,r,d,\gamma)$ *such that* $d_{\mathrm{KL}}(\Theta_1') \le c_3 b$ *and that for any* $q \in [1,2]$,

$$\log \mathcal{M}(\Theta_1', \|\cdot\|_{\mathrm{s}_q}, c_1\sqrt{b}\, r^{1/q-1/2}) \ge c_2 k \log \frac{\mathrm{e}m}{k}.$$

Proof. Let us focus on the first claim and we shall remark on how to establish the second claim at the end of this proof.

Let $W \in \mathbb{R}^{(s-r)\times r}$ satisfy the conclusion of Lemma 10 and define $s_0 = (1-c_0)(s-r)$. Let $\mathcal{B} = \{B_1, \ldots, B_N\}$ be a maximal set consisting of subsets of $[p]\backslash[r]$ with cardinality $s-r$ and for any $B_i \ne B_j$, $|B_i \cap B_j| \le s_0$. By Lemma A.3 of Rigollet and Tsybakov (2011) and Lemma 2.9 of Tsybakov (2009), there exists an absolute positive constant c_2' such that

$$\log N \ge c_2'(s-r) \log \frac{\mathrm{e}(p-r)}{s-r}.$$

Now for each $B_i \in \mathcal{B}$, define $W^{(i)} \in \mathbb{R}^{m\times n}$ by setting the submatrix $W^{(i)}_{B_i[r]} =$

W and filling the remaining entries with zeros. Then for any $i \neq j$, $|B_i \cap B_j| \leq s_0$, and so there exists a set $B_{ij} \subset [s]$ with $|B_{ij}| \geq s - r - s_0 = c_0(s-r)$, such that

$$\|W^{(i)} - W^{(j)}\|_{\mathrm{s}_q} \geq \|W_{B_{ij}*}\|_{\mathrm{s}_q} \geq c_1' r^{1/q-1/2},$$

where c_1' is an absolute constant due to Lemma 10.

Define $M_0 = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{p\times m}$ and for some positive constant $c_1'' \leq \frac{\gamma\wedge 2-1}{2} \wedge \sqrt{\frac{c_2'}{6\kappa_+^2}}$, let

$$\Theta_1 = \left\{ A^{(i)} = \frac{\gamma d}{2} M_0 + c_1'' \sqrt{a}\, W^{(i)} : i = 1, \ldots, N \right\}.$$

Note that each $A^{(i)}$ has s nonzero rows and r nonzero columns. Moreover, for $i \in [N]$, and $j \in [r]$

$$\left| \sigma_j(A^{(i)}) - \sigma_j(\frac{\gamma d}{2} M_0) \right| \leq \|A^{(i)} - \frac{\gamma d}{2} M_0\|_{\mathrm{op}} = c_1'' \sqrt{a} \|W^{(i)}\|_{\mathrm{op}} \leq c_1'' \sqrt{a} \|W^{(i)}\|_{\mathrm{F}} \leq \frac{\gamma \wedge 2 - 1}{2} d.$$

Here, the second last inequality holds since $\|W^{(i)}\|_{\mathrm{op}} \leq \|W^{(i)}\|_{\mathrm{F}} \leq 1$, and the last inequality holds since $c_1'' \leq \frac{\gamma\wedge 2-1}{2}$ and $\sqrt{a} \leq d$. Since $\sigma_j(\frac{\gamma d}{2} M_0) = \frac{\gamma d}{2}$ for all $j \in [r]$, and so $\sigma_j(A^{(i)}) \in [d, \gamma d]$ for all $j \in [r]$ and $i \in [N]$. Thus, $\Theta_1 \subset \Theta(s, r, d, \gamma)$.

For any $i \neq j$, $D(P_{A^{(i)}}||P_{A^{(j)}}) = \frac{1}{2}\|XA^{(i)} - XA^{(j)}\|_{\mathrm{F}}^2 \leq (c_1''\kappa_+)^2 a$, and

$$\|A^{(i)} - A^{(j)}\|_{\mathrm{s}_q} \geq c_1'' c_1' \sqrt{a}\, r^{1/q-1/2}.$$

Hence, for $c_1 = c_1' c_1''$, $c_2 = c_2'/2$ and $c_3 = (c_1''\kappa_+)^2$, $d_{\mathrm{KL}}(\mathcal{F}_0) \leq c_3 a$ and

$$\log \mathcal{M}(\Theta_1, \|\cdot\|_{\mathrm{s}_q}, c_1\sqrt{a}\, r^{1/q-1/2}) \geq c_2'(s-r)\log\frac{\mathrm{e}(p-r)}{s-r} \geq c_2 s \log\frac{\mathrm{e}p}{s}.$$

Here, the second inequality holds since $s \geq 2r$ and $\frac{p-r}{s-r} \geq \frac{p}{s}$. Moreover, by our choice of c_3, it is guaranteed that $c_3 \leq c_2/3$. This completes the proof of the first claim.

To establish the second claim, we note that Lemma 10 continues to hold if we replace s with k and W with W'. Thus, we could essentially repeat the foregoing arguments to obtain the second claim. This completes the proof. □

Proof of Theorem 2. Throughout the proof, let $c > 0$ denote a generic constant that depends only on γ and κ_+, though its actual value might vary at different occurrences. Note that we only need to prove the lower bounds for $\sigma = 1$, and the case of $\sigma \neq 1$ follows directly from standard scaling argument.

First, by restricting the nonzero entries of any matrix in $\Theta(s, k, r, d, \gamma)$

to the top left $s \times r$ (or $r \times k$) corner, we obtain a minimax lower bound by applying Lemma 9, i.e., for $\Theta = \Theta(s, r, d, \gamma)$ and any $q \in [1, 2]$,

$$\inf_{\widehat{A}} \sup_{\Theta} \mathbb{E}\|\widehat{A} - A\|_{s_q}^2 \geq c(r^{2/q-1}d^2) \wedge (r^{2/q}(s+k)). \tag{26}$$

Here, we have used the fact that for any $a, b, c > 0$,

$$(a \wedge b) \vee (a \wedge c) = a \wedge (b \vee c) \asymp a \wedge (b + c). \tag{27}$$

Next, by Proposition 1, Lemma 11 and (27), we obtain

$$\inf_{\widehat{A}} \sup_{\Theta} \mathbb{E}\|\widehat{A} - A\|_{s_q}^2 \geq c(\sqrt{a}\, r^{1/q-1/2})^2 = c(r^{2/q-1}d^2) \wedge \left(r^{2/q-1}\left(s \log \frac{ep}{s} + k \log \frac{em}{k}\right)\right). \tag{28}$$

Thus, the minimax risk is lower bounded by the maximum of the lower bounds in (26) and (28). Applying (27) again, we complete the proof. □

6.3 A Theorem on Group Lasso

Theorem 3. *Consider the linear model $W = XB + Z$, where W is an $n \times r$ response matrix, X is an $n \times p$ design matrix, B is a $p \times r$ coefficient matrix with s-sparse row support for some $s \geq 1$, and Z is an $n \times r$ error matrix.*

Let

$$\widehat{B} = \arg\min_{B \in \mathbb{R}^{p \times r}} \|W - XB\|_F^2/2 + \lambda \|B\|_{2,1},$$

with a given penalty level λ. *Let Condition 1 hold with an absolute constant* $K > 1$ *and positive constants* s_*, c_* *satisfying* (11).

(i) If $2\|X'_{*j}(W - XB)\|_F \leq \lambda$ *for all* j, *then it holds that*

$$\|\widehat{B} - B\|_F \leq \frac{3(1 + (4c_*)^{-1})}{\kappa_-^2(s_*)} \sqrt{s}\lambda. \tag{29}$$

(ii) Assume the error matrix Z *has iid* $N(0, \sigma^2)$ *entries. For any given* $\eta \in (0, 1)$, *if we set*

$$\lambda \geq 2\sigma \max_j \|X_{*j}\|(\sqrt{r} + \sqrt{2\log(p/\eta)}),$$

then (29) holds with probability at least $1 - \eta$.

Proof of Theorem 3. We may rewrite the minimization problem in a vectorized version as follows

$$\min_{B \in \mathbb{R}^{p \times r}} \|\mathrm{vec}(W) - (I_r \otimes X)\mathrm{vec}(B)\|_2^2/2 + \lambda \|B\|_{2,1},$$

where vec is usual vectorization operator and $\otimes$ is the Kronecker product

as defined in (Muirhead, 1982, Section 2.2). In this case, the rows of B form natural groups which are all of size r and $\mathrm{vec}(B)$ satisfies the (s, rs) strong group-sparsity as defined in Huang and Zhang (2010).

We are to prove the desired result by invoking Lemma D.4 of Huang and Zhang (2010). To this end, we first verify that the two conditions of the lemma is satisfied. Note that the penalty level in Huang and Zhang (2010) corresponds to $2\lambda/(nr)$ in our notion, X_{G_j} corresponds to X_{*j}, and the sparse eigenvalues $\rho_+(G_j)$ and $\rho_\pm(rs)$ are identified as

$$\rho_+(G_j) = \|X_{*j}\|^2/(nr), \quad \rho_\pm(rs) = \kappa_\pm^2(s)/(nr).$$

Let $\ell = s_* - s - 1$ and $\lambda_-^2 = \min\{k\lambda^2 : kr \geq \ell r + 1, k \in \mathbb{Z}^+\} = (\ell+1)\lambda^2$. The conditions of Huang and Zhang (2010, Lemma D.4) can be rewritten in our notation as

$$2\|X_{*j}'(W - XB)\|_F \leq \lambda \qquad \text{and} \qquad \frac{\widetilde{\kappa}_+^2(s_*, s_* - s)}{\kappa_-^2(s_*)} \leq \sqrt{\frac{\ell+1}{s}}, \tag{30}$$

where $\widetilde{\kappa}_+^2(s_*, s_* - s) = \sqrt{(\kappa_+^2(s_*) - \kappa_-^2(2s_* - s))(\kappa_+^2(s_* - s) - \kappa_-^2(2s_* - s))}$.

Since by Definition 1, $\kappa_-^2(s) \leq \kappa_-^2(t) \leq \kappa_+^2(t) \leq \kappa_+^2(s)$, $\forall t \leq s$, we obtain

$$\widetilde{\kappa}_+^2(s_*, s_* - s) \leq \kappa_+^2(s_*) - \kappa_-^2(2s_*).$$

Thus, the conditions in (30) are satisfied under the assumption of Theorem 3. Then the conclusion of Huang and Zhang (2010, Lemma D.4) leads to

$$\|\widehat{B} - B\|_F \leq \frac{3}{\kappa_-^2(s_*)}(1 + 1.5\sqrt{s/(\ell+1)})\sqrt{s}\lambda \leq \frac{3(1+(4c_*)^{-1})}{\kappa_-^2(s_*)}\sqrt{s}\lambda.$$

This completes the proof of part (i).

Turning to part (ii), we need to upper bound $2\|X'_{*j}(W - XB)\|_F$. Since $X'_{*j}(W - XB)$ is a vector of length r with iid $N(0, \sigma^2\|X_{*j}\|^2)$ entries, it follows from Laurent and Massart (2000, Eq.(4.3)) that with probability $1 - \eta/p$,

$$\begin{aligned}\|X'_{*j}(W - XB)\|_F^2 &\leq \sigma^2\|X_{*j}\|^2(r + 2\sqrt{r\log(p/\eta)} + 2\log(p/\eta)) \\ &\leq \sigma^2\|X_{*j}\|^2(\sqrt{r} + \sqrt{2\log(p/\eta)})^2.\end{aligned}$$

With probability at least $1 - \eta$, we have $2\|X'_{*j}(W - XB)\|_F \leq \lambda$ for all j and thus (29) holds. □