Supplementary Material for 'Multi-Stage Multi-Task Learning with Reduced Rank'

A. Some Basic Lemmas Used for Proofs

Lemma 3 *Let $\delta_1, \cdots, \delta_n$ be n random variables that are from the Gaussian distribution $\mathcal{N}(0, \sigma)$. Given another sequence $x_1, \cdots, x_n$ which satisfies $x_1^2 + \cdots + x_n^2 = 1$, define a random variable v as*

$$v = \frac{1}{\phi} \sum_{i=1}^{n} x_i \delta_i.$$

Then v follows a Gaussian distribution $\mathcal{N}(0, 1)$.

Lemma 4 *Let x^2 be a chi-squared random variable with k degrees of freedom, then we have*

$$\Pr(x^2 \geq k + c) \leq \exp\left(-\frac{1}{2}\left(c - k \ln\left(1 + \frac{c}{k}\right)\right)\right),$$

where c is a positive constant.

The proofs of Lemma 3 and Lemma 4 can be found in (Chen, Zhou, and Ye 2011). The proof of Lemma 1 can be found in (Zhang et al. 2012).

Lemma 5 *For any matrices $\hat{\mathbf{W}}$ and $\mathbf{W}$ with the same size $d \times m$, we have*

$$\sum_{i=1}^{R} (\sigma_i(\hat{\mathbf{W}}) - \sigma_i(\mathbf{W}))^2 \leq \|\hat{\mathbf{W}} - \mathbf{W}\|_*^2. \tag{11}$$

Lemma 6 *Let $\bar{r}$ be the rank of $\bar{\mathbf{W}}$. For any estimator $\hat{\mathbf{W}}$, we have the following inequalities satisfied:*

$$\sum_{i \in \bar{\mathcal{F}}} \mathbb{I}^2(\sigma_i(\hat{\mathbf{W}}) \geq \tau) \leq \bar{r}, \tag{12}$$

$$\sum_{i \in \bar{\mathcal{F}}^c} \mathbb{I}^2(\sigma_i(\hat{\mathbf{W}}) \geq \tau) \leq \frac{(R - \bar{r})}{\tau^2} \sum_{i \in \bar{\mathcal{F}}^c} \left(\sigma_i(\bar{\mathbf{W}}) - \sigma_i(\hat{\mathbf{W}})\right)^2. \tag{13}$$

Lemma 5 and Lemma 6 reveals the inherent relationships among $\mathbb{I}(\sigma_i(\hat{\mathbf{W}}) \geq \tau)$, $\sigma_i(\hat{\mathbf{W}}) - \sigma_i(\mathbf{W})$, and $\|\hat{\mathbf{W}} - \mathbf{W}\|_*^2$ for any estimator $\hat{\mathbf{W}}$.

B. Proofs in Section and Section

B.1 Proof of Lemma 2 For any non-negative integer $r \leq R$, and matrices $\mathbf{A} \in \mathcal{C}_{r,d}$, $\mathbf{B} \in \mathcal{C}_{r,m}$, we can directly obtain the following result with the equality held in Lemma 1:

$$\|\mathbf{W}\|_{r^+} = \sum_{i=1}^{r} \sigma_i(\mathbf{W}) = \max_{\mathbf{A} \in \mathcal{C}_{r,d}, \mathbf{B} \in \mathcal{C}_{r,m}} \mathrm{tr}(\mathbf{A}\mathbf{W}\mathbf{B}^T).$$

Now we have to show $\max_{\mathbf{A} \in \mathcal{C}_{r,d}, \mathbf{B} \in \mathcal{C}_{r,m}} \mathrm{tr}(\mathbf{A}\mathbf{W}\mathbf{B}^T) = \mathrm{tr}\left(\hat{\mathbf{A}}\mathbf{W}\hat{\mathbf{B}}^T\right)$. Actually, we have

$$\begin{aligned}
\mathrm{tr}\left(\hat{\mathbf{A}}\mathbf{W}\hat{\mathbf{B}}^T\right) &= \mathrm{tr}\left((\mathbf{u}_1, \cdots, \mathbf{u}_r)^T \mathbf{W} (\mathbf{v}_1, \cdots, \mathbf{v}_r)\right) \\
&= \mathrm{tr}\left((\mathbf{u}_1, \cdots, \mathbf{u}_r)^T \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T (\mathbf{v}_1, \cdots, \mathbf{v}_r)\right) \\
&= \mathrm{tr}\left((\mathbf{u}_1, \cdots, \mathbf{u}_r)^T \mathbf{U}\right) \boldsymbol{\Sigma} \left(\mathbf{V}^T (\mathbf{v}_1, \cdots, \mathbf{v}_r)\right) \\
&= \mathrm{tr}\left(\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{\Sigma} \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\right) \\
&= \mathrm{tr}(\mathrm{diag}([\sigma_1(\mathbf{W}), \cdots, \sigma_r(\mathbf{W}), 0, \cdots, 0])) \\
&= \sum_{i=1}^{r} \|\sigma_i(\mathbf{W})\| = \|\mathbf{W}\|_{r^+},
\end{aligned}$$

where $\mathbf{I}_r$ is a $r \times r$ identity matrix. Then we reach the conclusion.

Next, we show $\|\mathbf{W}\|_{r^+}$ is convex with respect to $\mathbf{W}$ and the operator $\|\cdot\|_{r^+}$ is a norm. From the theorem 2.2 in (Chen, Dong, and Chan 2013), we know that the function $f(\mathbf{W}) = \sum_{i=1}^{R} \omega_i \sigma_i(\mathbf{W})$ is convex with respect to $\mathbf{W}$ if and only if the weights ω_i's are decreasingly ordered by $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_R \geq 0$. For $\|\mathbf{W}\|_{r^+}$, we can rewrite

$$\|\mathbf{W}\|_{r^+} = 1 \cdot \sigma_1(\mathbf{W}) + \cdots + 1 \cdot \sigma_r(\mathbf{W}) + 0 \cdot \sigma_{r+1}(\mathbf{W}) + 0 \cdot \sigma_R(\mathbf{W}),$$

where the decreasing order of the weights are satisfied. Therefore, $\|\mathbf{W}\|_{r^+}$ is convex with respect to $\mathbf{W}$. Moreover, for any matrix $\mathbf{W}$, $\mathbf{W}_1$ and $\mathbf{W}_2$, we have: (1) $\|\mathbf{W}\|_{r^+} \geq 0$; (2) $\|\mathbf{W}\|_{r^+} = 0$ if and only if $\mathbf{W} = \mathbf{0}$; (3) $\|c\mathbf{W}\|_{r^+} = |c| \|\mathbf{W}\|_{r^+}$ for any scalar c; (4) $\|\mathbf{W}_1 + \mathbf{W}_2\|_{r^+} \leq \|\mathbf{W}_1\|_{r^+} + \|\mathbf{W}_2\|_{r^+}$ due to the convexity of $\|\cdot\|_{r^+}$. By the definition of norm, we know that $\|\cdot\|_{r^+}$ is a norm, which completes the proof. □

B.2 Proof of Theorem 1 From Eq. (4), we have

$$\begin{aligned}
&\frac{1}{mn} \sum_{i=1}^{m} \|\mathbf{X}_i \hat{\mathbf{w}}_i - \mathbf{y}_i\|_2^2 \\
&\leq \frac{1}{mn} \sum_{i=1}^{m} \|\mathbf{X}_i \mathbf{w}_i - \mathbf{y}_i\|_2^2 + \lambda \|\mathbf{W}\|_* - \lambda \|\hat{\mathbf{W}}\|_* \\
&\quad + \lambda \mathrm{tr}\left(\hat{\mathbf{A}}_t \hat{\mathbf{W}} \hat{\mathbf{B}}_t^T\right) - \lambda \mathrm{tr}\left(\hat{\mathbf{A}}_t \mathbf{W} \hat{\mathbf{B}}_t^T\right).
\end{aligned} \tag{14}$$

Based on the property of the trace, we have

$$\mathrm{tr}\left(\hat{\mathbf{A}}_t \mathbf{W} \hat{\mathbf{B}}_t^T\right) = \mathrm{tr}\left(\mathbf{W} \hat{\mathbf{B}}_t^T \hat{\mathbf{A}}_t\right). \tag{15}$$

Then, we have

$$\begin{aligned}
&\frac{1}{mn} \sum_{i=1}^{m} \|\mathbf{X}_i \hat{\mathbf{w}}_i - \bar{f}_i\|_2^2 \\
&\leq \frac{1}{mn} \sum_{i=1}^{m} \|\mathbf{X}_i \mathbf{w}_i - \bar{f}_i\|_2^2 + \lambda(\|\mathbf{W}\|_* - \|\hat{\mathbf{W}}\|_*) \\
&\quad + \lambda \mathrm{tr}\left((\hat{\mathbf{W}} - \mathbf{W}) \hat{\mathbf{B}}_t^T \hat{\mathbf{A}}_t\right) + \sum_{i=1}^{m} \langle \hat{\mathbf{w}}_i - \mathbf{w}_i, \mathbf{X}_i \delta_i \rangle.
\end{aligned} \tag{16}$$

We first compute the upper bound of $\sum_{i=1}^{m} \langle \hat{\mathbf{w}}_i - \mathbf{w}_i, \mathbf{X}_i \delta_i \rangle$. Define a set of random events $\{\mathcal{A}_i\}$ as

$$\mathcal{A}_i = \{\|\mathbf{X}_i \delta_i\|_2 \leq \lambda\}, \forall i \in \mathbb{N}_m.$$

For each $\mathcal{A}_i$, define a set of random variables $\{v_{ij}\}$ as

$$v_{ij} = \frac{1}{\phi} \sum_{k=1}^{n} x_{jk}^i \delta_{ik}, j \in \mathbb{N}_d,$$

where x_{jk}^i denotes the (j, k)-th entry of the data matrix $\mathbf{X}_i$. Since $\mathbf{X}_i$ is normalized, the diagonal elements of $\mathbf{X}_i^T \mathbf{X}_i$ are ones, and thus $\{v_{i1}, \cdots, v_{id}\}$ are i.i.d. Gaussian variables following $\mathcal{N}(0, 1)$ by Lemma 3. Then we can verify that $\sum_{j=1}^{d} v_{ij}^2$ is a chi-squared random variable with d degree of freedom. By choosing λ according to Theorem 1, we have

$$\begin{aligned}\Pr(\frac{2}{mn}\|\mathbf{X}_i\delta_i\|_2 > \lambda) =&\Pr(\sum_{j=1}^{d}(\sum_{k=1}^{n} x_{jk}^i\delta_{ik})^2 > \frac{\lambda^2 m^2 n^2}{4})\\ =&\Pr(\sum_{j=1}^{d} v_{ij}^2 > d + c)\\ \leq&\exp(-\frac{1}{2}\mu_d^2(c)),\end{aligned}$$

where $\mu_d(c) = \sqrt{c - d\ln(1+\frac{c}{d})}$ and the last inequality holds due to Lemma 4. Let $\mathcal{A} = \bigcap_{i=1}^m \mathcal{A}_i$ and denote by $\mathcal{A}_i^c$ the complement of each event $\mathcal{A}_i$. It follows that

$$\Pr(\mathcal{A}) \geq 1 - \Pr(\bigcup_{i=1}^m \mathcal{A}_i^c) \geq 1 - m\exp(-\frac{1}{2}\mu_d^2(c)).$$

Under the event $\mathcal{A}$, we can derive a bound on $\sum_{i=1}^m \langle \hat{\mathbf{w}}_i - \mathbf{w}_i, \mathbf{X}_i\delta_i\rangle$ as

$$\begin{aligned}\sum_{i=1}^m\langle\hat{\mathbf{w}}_i - \mathbf{w}_i, \mathbf{X}_i\delta_i\rangle \leq& \sum_{i=1}^m \|\hat{\mathbf{w}}_i - \mathbf{w}_i\|_2\|\mathbf{X}_i\delta_i\|_2\\ \leq&\lambda\sum_{i=1}^m \|\hat{\mathbf{w}}_i - \mathbf{w}_i\|_2\\ \leq&\sqrt{m}\lambda\|\hat{\mathbf{W}} - \mathbf{W}\|_*. \end{aligned} \quad (17)$$

Next, we examine the bound for the trace term $\mathrm{tr}\left((\hat{\mathbf{W}} - \mathbf{W})\hat{\mathbf{B}}_t^T\hat{\mathbf{A}}_t\right)$. By using Lemma 1, we have

$$\begin{aligned}\lambda\mathrm{tr}\left((\hat{\mathbf{W}} - \mathbf{W})\hat{\mathbf{B}}_t^T\hat{\mathbf{A}}_t\right) \quad\leq&\quad \lambda\sum_{i=1}^{r_l^+}\sigma_i(\hat{\mathbf{W}} - \mathbf{W})\\ =&\quad \lambda\|\hat{\mathbf{W}} - \mathbf{W}\|_{r_l^+}. \end{aligned} \quad (18)$$

Combining Eq. (16), Eq. (17) and Eq. (18) together with the fact that $\|\mathbf{W}\|_* - \|\hat{\mathbf{W}}\|_* \leq \|\hat{\mathbf{W}} - \mathbf{W}\|_*$, we can reach the conclusion. □

B.3 Proof of Lemma 5 The conclusion can be reached by the following steps as

$$\begin{aligned}&\sum_{i=1}^R(\sigma_i(\hat{\mathbf{W}}) - \sigma_i(\mathbf{W}))^2\\ &= \sum_{i=1}^R\sigma_i^2(\hat{\mathbf{W}}) + \sum_{i=1}^R\sigma_i^2(\mathbf{W}) - \sum_{i=1}^R 2\sigma_i(\hat{\mathbf{W}})\sigma_i(\mathbf{W})\\ &= \|\hat{\mathbf{W}}\|_F^2 + \|\mathbf{W}\|_F^2 - 2\sum_{i=1}^R\sigma_i(\hat{\mathbf{W}})\sigma_i(\mathbf{W})\\ &\leq \|\hat{\mathbf{W}}\|_F^2 + \|\mathbf{W}\|_F^2 - 2\mathrm{tr}(\hat{\mathbf{W}}^T\mathbf{W})\\ &= \|\hat{\mathbf{W}} - \mathbf{W}\|_F^2 \leq \|\hat{\mathbf{W}} - \mathbf{W}\|_*^2,\end{aligned}$$

where the inequality is due to the Von Neumann's trace inequality. □

B.4 Proof of Lemma 6 For $i \in \bar{\mathcal{F}}$, it is easy to see that

$$\sum_{i\in\bar{\mathcal{F}}}\mathbb{I}^2(\sigma_i(\hat{\mathbf{W}}) \geq \tau) \leq |\bar{\mathcal{F}}| = \bar{r}. \quad (19)$$

For $i \in \bar{\mathcal{F}}^c \cap \hat{\mathcal{G}}$, we have $\sigma_i(\bar{\mathbf{W}}) = 0$ and $\sigma_i(\hat{\mathbf{W}}) < \tau$, therefore

$$\begin{aligned}&\sum_{i\in\bar{\mathcal{F}}^c\cap\hat{\mathcal{G}}}\mathbb{I}^2(\sigma_i(\hat{\mathbf{W}}) \geq \tau)\\ &= 0\\ &\leq \frac{|\bar{\mathcal{F}}^c\cap\hat{\mathcal{G}}|}{\tau^2}\sum_{i\in\bar{\mathcal{F}}^c\cap\hat{\mathcal{G}}}\left(\sigma_i(\bar{\mathbf{W}}) - \sigma_i(\hat{\mathbf{W}})\right)^2.\end{aligned} \quad (20)$$

For $i \in \bar{\mathcal{F}}^c \cap \hat{\mathcal{G}}^c$, we have $\sigma_i(\bar{\mathbf{W}}) = 0$ and $\sigma_i(\hat{\mathbf{W}}) \geq \tau$, therefore we also have

$$\begin{aligned}&\sum_{i\in\bar{\mathcal{F}}^c\cap\hat{\mathcal{G}}^c}\mathbb{I}^2(\sigma_i(\hat{\mathbf{W}}) \geq \tau)\\ &\leq |\bar{\mathcal{F}}^c\cap\hat{\mathcal{G}}^c|\\ &\leq \frac{|\bar{\mathcal{F}}^c\cap\hat{\mathcal{G}}^c|}{\tau^2}\sum_{i\in\bar{\mathcal{F}}^c\cap\hat{\mathcal{G}}}\left(\sigma_i(\bar{\mathbf{W}}) - \sigma_i(\hat{\mathbf{W}})\right)^2.\end{aligned} \quad (21)$$

Combing Eqs. (19)-(21), we reach the conclusion. □

B.5 Proof of Theorem 2 Let $\mathbf{W} = \bar{\mathbf{W}}$ and set $\Delta = \hat{\mathbf{W}} - \bar{\mathbf{W}}$. By Assumption 1, we have

$$\kappa^2\|\hat{\mathbf{W}} - \bar{\mathbf{W}}\|_*^2 \leq \frac{1}{mn}\|\mathcal{X}\mathcal{D}(\hat{\mathbf{W}}) - \mathcal{D}(\bar{\mathbf{F}})\|_F^2. \quad (22)$$

Let $\lambda_i^{(l)} = \lambda\mathbb{I}(\sigma_i(\hat{\mathbf{W}}_\star^{(l)}) \geq \tau)$, we can rewrite the last term in Eq. (9) as

$$\begin{aligned}&\lambda\|\hat{\mathbf{W}} - \bar{\mathbf{W}}\|_{r_l^+}\\ &= \sum_{i=1}^R\lambda_i^{(l)}\sigma_i(\hat{\mathbf{W}} - \bar{\mathbf{W}})\\ &= \lambda\sum_{i=1}^R\mathbb{I}(\sigma_i(\hat{\mathbf{W}}_\star^{(l)}) \geq \tau)\sigma_i(\hat{\mathbf{W}} - \bar{\mathbf{W}})\\ &= \lambda\sum_{i\in\bar{\mathcal{F}}}\mathbb{I}(\sigma_i(\hat{\mathbf{W}}_\star^{(l)}) \geq \tau)\sigma_i(\hat{\mathbf{W}} - \bar{\mathbf{W}})\\ &\quad + \lambda\sum_{i\in\bar{\mathcal{F}}^c}\mathbb{I}(\sigma_i(\hat{\mathbf{W}}_\star^{(l)}) \geq \tau)\sigma_i(\hat{\mathbf{W}} - \bar{\mathbf{W}}).\end{aligned} \quad (23)$$

By combining Lemmas 5 and 6 with Eq. (23), we have

$$\begin{aligned}&\lambda\|\hat{\mathbf{W}} - \bar{\mathbf{W}}\|_{r_l^+}\\ &\leq \lambda\sqrt{\bar{r} + \frac{R - \bar{r}\sum_{i\in\bar{\mathcal{F}}^c}\left(\sigma_i(\bar{\mathbf{W}}) - \sigma_i(\hat{\mathbf{W}}_\star^{(l)})\right)^2}{\tau^2}}\sqrt{\|\hat{\mathbf{W}} - \bar{\mathbf{W}}\|_F^2}\\ &\leq \left(\lambda\sqrt{\bar{r}} + \frac{\lambda\sqrt{R - \bar{r}}}{\tau}\|\hat{\mathbf{W}}_\star^{(l)} - \bar{\mathbf{W}}\|_*\right)\|\hat{\mathbf{W}} - \bar{\mathbf{W}}\|_F\\ &\leq \left(\lambda\sqrt{\bar{r}} + \frac{\lambda\sqrt{R - \bar{r}}}{\tau}\|\hat{\mathbf{W}}_\star^{(l)} - \bar{\mathbf{W}}\|_*\right)\|\hat{\mathbf{W}} - \bar{\mathbf{W}}\|_*,\end{aligned} \quad (24)$$

where the first inequality holds due to the Cauchy-Schwarz inequality and Lemma 6, and the second inequality is due to Lemma 5 and a fact that $\sqrt{a^2 + b^2} \leq a + b$ for all $a, b \geq 0$.

Now, by substituting Eq. (24) into Eq. (9) and combining

Eq. (22), we obtain

$$
\begin{aligned}
&\|\hat{\mathbf{W}}_\star^{(l+1)} - \bar{\mathbf{W}}\|_* \\
&\leq \frac{\lambda\sqrt{R-\bar{r}}}{\tau\kappa^2}\|\hat{\mathbf{W}}_\star^{(l)} - \bar{\mathbf{W}}\|_* + \frac{\lambda(\sqrt{\bar{r}}+1+\sqrt{m})}{\kappa^2} \\
&\leq a^l\|\hat{\mathbf{W}}^{(0)} - \bar{\mathbf{W}}\|_* + b\frac{1-a^l}{1-a} \\
&\leq a^l\|\hat{\mathbf{W}}^{(0)} - \bar{\mathbf{W}}\|_* + \frac{b}{1-a},
\end{aligned}
$$

where $a = \frac{\lambda\sqrt{R-\bar{r}}}{\tau\kappa^2} < 1$, $b = \frac{\lambda(\sqrt{\bar{r}}+1+\sqrt{m})}{\kappa^2}$. □