Supplementary Material: Multi-Stage Multi-Task Feature Learning

In this supplementary material, we provide detailed proofs for Lemmas 1-4. In our proofs, we use several lemmas (summarized in part B) from Zhang (2010) [26].

We first introduce some notations used in the proof. Define

$$\pi_i(k_i, s_i) = \sup_{\mathbf{v}\in\mathbb{R}^{k_i}, \mathbf{u}\in\mathbb{R}^{s_i}, \mathcal{I}_i, \mathcal{J}_i} \frac{\mathbf{v}^T A^{(i)}_{\mathcal{I}_i,\mathcal{J}_i}\mathbf{u}\|\mathbf{v}\|}{\mathbf{v}^T A^{(i)}_{\mathcal{I}_i,\mathcal{I}_i}\mathbf{v}\|\mathbf{u}\|_\infty}, \tag{16}$$

where $s_i + k_i \leq d$ with $s_i, k_i \geq 1$; $\mathcal{I}_i$ and $\mathcal{J}_i$ are *disjoint* subsets of $\mathbb{N}_d$ with k_i and s_i elements respectively (with some abuse of notation, we also let $\mathcal{I}_i$ be a subset of $\mathbb{N}_d \times \{i\}$, depending on the context.); $A^{(i)}_{\mathcal{I}_i,\mathcal{J}_i}$ is a sub-matrix of $A_i = n^{-1}X_i^T X_i \in \mathbb{R}^{d\times d}$ with rows indexed by $\mathcal{I}_i$ and columns indexed by $\mathcal{J}_i$.

We let $\mathbf{w}_{\mathcal{I}_i}$ be a $d \times 1$ vector with the j-th entry being w_{ji}, if $(j, i) \in \mathcal{I}_i$, and 0, otherwise. We also let $W_{\mathcal{I}}$ be a $d \times m$ matrix with (j, i)-th entry being w_{ji}, if $(j, i) \in \mathcal{I}$, and 0, otherwise.

A. Proofs of Lemmas 1-4

A.1. Proof of Lemma 1

Proof For the j-th entry of $\bar{\boldsymbol{\epsilon}}_i$ ($j \in \mathbb{N}_d$):

$$|\bar{\epsilon}_{ji}| = \frac{1}{n}\left|\left(\mathbf{x}_j^{(i)}\right)^T (X_i\bar{\mathbf{w}}_i - \mathbf{y}_i)\right| = \frac{1}{n}\left|\left(\mathbf{x}_j^{(i)}\right)^T \boldsymbol{\delta}_i\right|,$$

where $\mathbf{x}_j^{(i)}$ is the j-th column of X_i. We know that the entries of $\boldsymbol{\delta}_i$ are independent sub-Gaussian random variables, and $\|1/n\mathbf{x}_j^{(i)}\|^2 = \|\mathbf{x}_j^{(i)}\|^2/n^2 \leq \rho_i^+(1)/n$. According to Lemma 5, we have $\forall t > 0$:

$$\Pr(|\bar{\epsilon}_{ji}| \geq t) \leq 2\exp(-nt^2/(2\sigma^2\rho_i^+(1))) \leq 2\exp(-nt^2/(2\sigma^2\rho^+_{max}(1))).$$

Thus we obtain:

$$\Pr(\|\bar{\Upsilon}\|_{\infty,\infty} \leq t) \geq 1 - 2dm\exp(-nt^2/(2\sigma^2\rho^+_{max}(1))).$$

Let $\eta = 2dm\exp(-nt^2/(2\sigma^2\rho^+_{max}(1)))$ and we can obtain Eq. (10). Eq. (11) directly follows from Lemma 8 and the following fact:

$$\|\mathbf{x}_i\|^2 \leq ay_i \Rightarrow \|X\|_F^2 = \sum_{i=1}^m \|\mathbf{x}_i\|^2 \leq ma\max_{i\in\mathbb{N}_m} y_i.$$

□

A.2 Proof of Lemma 2

Proof The optimality condition of Eq. (2) implies that

$$\frac{2}{n}X_i^T(X_i\hat{\mathbf{w}}_i - \mathbf{y}_i) + \hat{\boldsymbol{\lambda}}_i \odot \mathrm{sign}(\hat{\mathbf{w}}_i) = \mathbf{0},$$

where $\odot$ denotes the element-wise product; $\mathrm{sign}(\mathbf{w}) = [\mathrm{sign}(w_1), \cdots, \mathrm{sign}(w_d)]^T$, where $\mathrm{sign}(w_i) = 1$, if $w_i > 0$; $\mathrm{sign}(w_i) = -1$, if $w_i < 0$; and $\mathrm{sign}(w_i) \in [-1, 1]$, otherwise. We note that $X_i\hat{\mathbf{w}}_i - \mathbf{y}_i = X_i\hat{\mathbf{w}}_i - X_i\bar{\mathbf{w}}_i + X_i\bar{\mathbf{w}}_i - \mathbf{y}_i$ and we can rewrite the above equation into the following form:

$$2A_i\Delta\hat{\mathbf{w}}_i = -2\bar{\boldsymbol{\epsilon}}_i - \hat{\boldsymbol{\lambda}}_i \odot \mathrm{sign}(\hat{\mathbf{w}}_i).$$

Thus, for all $\mathbf{v} \in \mathbb{R}^d$, we have

$$2\mathbf{v}^T A_i \Delta\hat{\mathbf{w}}_i = -2\mathbf{v}^T \bar{\boldsymbol{\epsilon}}_i - \sum_{j=1}^{d} \hat{\lambda}_{ji} v_j \text{sign}(\hat{w}_{ji}). \tag{17}$$

Letting $\mathbf{v} = \Delta\hat{\mathbf{w}}_i$ and noticing that $\Delta\hat{w}_{ji} = \hat{w}_{ji}$ for $(j,i) \notin \bar{\mathcal{F}}_i, i \in \mathbb{N}_m$, we obtain

$$\begin{aligned}
0 \leq 2\Delta\hat{\mathbf{w}}_i^T A_i \Delta\hat{\mathbf{w}}_i &= -2\Delta\hat{\mathbf{w}}_i^T \bar{\boldsymbol{\epsilon}}_i - \sum_{j=1}^{d} \hat{\lambda}_{ji} \Delta\hat{w}_{ji} \text{sign}(\hat{w}_{ji}) \\
&\leq 2\|\Delta\hat{\mathbf{w}}_i\|_1 \|\bar{\boldsymbol{\epsilon}}_i\|_\infty - \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji} \Delta\hat{w}_{ji} \text{sign}(\hat{w}_{ji}) - \sum_{(j,i)\notin\bar{\mathcal{F}}_i} \hat{\lambda}_{ji} \Delta\hat{w}_{ji} \text{sign}(\hat{w}_{ji}) \\
&\leq 2\|\Delta\hat{\mathbf{w}}_i\|_1 \|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji} |\Delta\hat{w}_{ji}| - \sum_{(j,i)\notin\bar{\mathcal{F}}_i} \hat{\lambda}_{ji} |\hat{w}_{ji}| \\
&\leq 2\|\Delta\hat{\mathbf{w}}_i\|_1 \|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji} |\Delta\hat{w}_{ji}| - \sum_{(j,i)\in\mathcal{G}_i} \hat{\lambda}_{ji} |\hat{w}_{ji}| \\
&\leq 2\|\Delta\hat{\mathbf{w}}_i\|_1 \|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{0i} |\Delta\hat{w}_{ji}| - \sum_{(j,i)\in\mathcal{G}_i} \hat{\lambda}_{\mathcal{G}_i} |\hat{w}_{ji}| \\
&= \sum_{(j,i)\in\mathcal{G}_i} (2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty - \hat{\lambda}_{\mathcal{G}_i}) |\hat{w}_{ji}| + \sum_{(j,i)\notin\bar{\mathcal{F}}_i\cup\mathcal{G}_i} 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty |\hat{w}_{ji}| + \sum_{(j,i)\in\bar{\mathcal{F}}_i} (2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \hat{\lambda}_{0i}) |\Delta\hat{w}_{ji}|.
\end{aligned}$$

The last equality above is due to $\mathbb{N}_d \times \{i\} = \mathcal{G}_i \cup (\bar{\mathcal{F}}_i \cup \mathcal{G}_i)^c \cup \bar{\mathcal{F}}_i$ and $\Delta\hat{w}_{ji} = \hat{w}_{ji}, \forall (j,i) \notin \bar{\mathcal{F}}_i \supseteq \mathcal{G}_i$. Rearranging the above inequality and noticing that $2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty < \hat{\lambda}_{\mathcal{G}_i} \leq \hat{\lambda}_{0i}$, we obtain:

$$\sum_{(j,i)\in\mathcal{G}_i} |\hat{w}_{ji}| \leq \frac{2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty} \sum_{(j,i)\notin\bar{\mathcal{F}}_i\cup\mathcal{G}_i} |\hat{w}_{ji}| + \frac{2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \hat{\lambda}_{0i}}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty} \sum_{(j,i)\in\bar{\mathcal{F}}_i} |\Delta\hat{w}_{ji}| \leq \frac{2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \hat{\lambda}_{0i}}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty} \|\Delta\hat{\mathbf{w}}_{\mathcal{G}_i^c}\|_1. \tag{18}$$

Then Lemma 2 can be obtained from the above inequality and the following two inequalities.

$$\max_{i\in\mathbb{N}_m} \frac{2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \hat{\lambda}_{0i}}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty} \leq \frac{2\|\bar{\Upsilon}\|_{\infty,\infty} + \hat{\lambda}_0}{\hat{\lambda}_{\mathcal{G}} - 2\|\bar{\Upsilon}\|_{\infty,\infty}} \text{ and } \sum_{i=1}^{m} x_i y_i \leq \|\mathbf{x}\|_\infty \|\mathbf{y}\|_1.$$

□

A.3 Proof of Lemma 3

Proof According to the definition of $\mathcal{G}$ ($\mathcal{G}_{(\ell)}$), we know that $\bar{\mathcal{F}}_i \cap \mathcal{G}_i = \emptyset$ ($i \in \mathbb{N}_m$) and $\forall (j,i) \in \mathcal{G}$ ($\mathcal{G}_{(\ell)}$), $\hat{\lambda}_{ji}^{(\ell-1)} = \lambda$. Thus, all conditions of Lemma 2 are satisfied, by noticing the relationship between Eq. (5) and Eq. (10). Based on the definition of $\mathcal{G}$ ($\mathcal{G}_{(\ell)}$), we easily obtain $\forall j \in \mathbb{N}_d$:

$$(j,i) \in \mathcal{G}_i, \forall i \in \mathbb{N}_m \text{ or } (j,i) \notin \mathcal{G}_i, \forall i \in \mathbb{N}_m. \tag{19}$$

and hence $k_\ell = |\mathcal{G}_1^c| = \cdots = |\mathcal{G}_m^c|$ (k_ℓ is some integer). Now, we assume that at stage $\ell \geq 1$:

$$k_\ell = |\mathcal{G}_1^c| = \cdots = |\mathcal{G}_m^c| \leq 2\bar{r}. \tag{20}$$

We will show in the second part of this proof that Eq. (20) holds for all ℓ. Based on Lemma 6 and Eq. (4), we have:

$$\pi_i(2\bar{r}+s, s) \leq \frac{s^{1/2}}{2} \sqrt{\rho_i^+(s)/\rho_i^-(2\bar{r}+2s) - 1} \leq \frac{s^{1/2}}{2} \sqrt{1 + s/(2\bar{r}) - 1} = 0.5s(2\bar{r})^{-1/2},$$

which indicates that

$$0.5 \leq t_i = 1 - \pi_i(2\bar{r}+s, s)(2\bar{r})^{1/2} s^{-1} \leq 1.$$

For all $t_i \in [0.5, 1]$, under the conditions of Eq. (5) and Eq. (10), we have

$$\frac{2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \lambda}{\lambda - 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty} \leq \frac{2\|\bar{\Upsilon}\|_{\infty,\infty} + \lambda}{\lambda - 2\|\bar{\Upsilon}\|_{\infty,\infty}} \leq \frac{7}{5} \leq \frac{4 - t_i}{4 - 3t_i} \leq 3.$$

Following Lemma 2, we have

$$\|\hat{W}_{\mathcal{G}}\|_{1,1} \leq 3\|\Delta\hat{W}_{\mathcal{G}^c}\|_{1,1} = 3\|\Delta\hat{W} - \Delta\hat{W}_{\mathcal{G}}\|_{1,1} = 3\|\Delta\hat{W} - \hat{W}_{\mathcal{G}}\|_{1,1}.$$

Therefore

$$\begin{aligned}
&\|\Delta\hat{W} - \Delta\hat{W}_{\mathcal{I}}\|_{\infty,1} = \|\Delta\hat{W}_{\mathcal{G}} - \Delta\hat{W}_{\mathcal{J}}\|_{\infty,1} \\
&\leq \|\Delta\hat{W}_{\mathcal{J}}\|_{1,1}/s = (\|\Delta\hat{W}_{\mathcal{G}}\|_{1,1} - \|\Delta\hat{W} - \Delta\hat{W}_{\mathcal{I}}\|_{1,1})/s \\
&\leq s^{-1}(3\|\Delta\hat{W} - \hat{W}_{\mathcal{G}}\|_{1,1} - \|\Delta\hat{W} - \Delta\hat{W}_{\mathcal{I}}\|_{1,1}),
\end{aligned}$$

which implies that

$$\begin{aligned}
&\|\Delta\hat{W}\|_{2,1} - \|\Delta\hat{W}_{\mathcal{I}}\|_{2,1} \leq \|\Delta\hat{W} - \Delta\hat{W}_{\mathcal{I}}\|_{2,1} \\
&\leq (\|\Delta\hat{W} - \Delta\hat{W}_{\mathcal{I}}\|_{1,1}\|\Delta\hat{W} - \Delta\hat{W}_{\mathcal{I}}\|_{\infty,1})^{1/2} \\
&\leq \left(\|\Delta\hat{W} - \Delta\hat{W}_{\mathcal{I}}\|_{1,1}\right)^{1/2} \left(s^{-1}(3\|\Delta\hat{W} - \hat{W}_{\mathcal{G}}\|_{1,1} - \|\Delta\hat{W} - \Delta\hat{W}_{\mathcal{I}}\|_{1,1})\right)^{1/2} \\
&\leq \left(\left(3\|\Delta\hat{W} - \hat{W}_{\mathcal{G}}\|_{1,1}/2\right)^2\right)^{1/2} s^{-1/2} \\
&\leq (3/2)s^{-1/2}(2\bar{r})^{1/2}\|\Delta\hat{W} - \hat{W}_{\mathcal{G}}\|_{2,1} \\
&\leq (3/2)(2\bar{r}/s)^{1/2}\|\Delta\hat{W}_{\mathcal{I}}\|_{2,1}.
\end{aligned}$$

In the above derivation, the third inequality is due to $a(3b - a) \leq (3b/2)^2$, and the fourth inequality follows from Eq. (20) and $\bar{\mathcal{F}} \cap \mathcal{G} = \emptyset \Rightarrow \Delta\hat{W}_{\mathcal{G}} = \hat{W}_{\mathcal{G}}$. Rearranging the above inequality, we obtain at stage ℓ:

$$\|\Delta\hat{W}\|_{2,1} \leq \left(1 + 1.5\sqrt{\frac{2\bar{r}}{s}}\right)\|\Delta\hat{W}_{\mathcal{I}}\|_{2,1}. \tag{21}$$

From Lemma 7, we have:

$$\begin{aligned}
&\max(0, \Delta\hat{\mathbf{w}}_{\mathcal{I}_i}^T A_i \Delta\hat{\mathbf{w}}_i) \\
&\geq \rho_i^-(k_\ell + s)(\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\| - \pi_i(k_\ell + s, s)\|\hat{\mathbf{w}}_{\mathcal{G}_i}\|_1/s)\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\| \\
&\geq \rho_i^-(k_\ell + s)[1 - (1 - t_i)(4 - t_i)/(4 - 3t_i)]\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\|^2 \\
&\geq 0.5t_i\rho_i^-(k_\ell + s)\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\|^2 \\
&\geq 0.25\rho_i^-(2\bar{r} + s)\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\|^2 \\
&\geq 0.25\rho_{min}^-(2\bar{r} + s)\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\|^2,
\end{aligned}$$

where the second inequality is due to Eq. (18), that is

$$\begin{aligned}
\|\hat{\mathbf{w}}_{\mathcal{G}_i}\|_1 &\leq \frac{2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \hat{\lambda}_{0i}}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty}\|\Delta\hat{\mathbf{w}}_{\mathcal{G}_i^c}\|_1 \leq \frac{(2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \hat{\lambda}_{0i})\sqrt{k_\ell}}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty}\|\Delta\hat{\mathbf{w}}_{\mathcal{G}_i^c}\| \\
&\leq \frac{(2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \hat{\lambda}_{0i})\sqrt{k_\ell}}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty}\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\| \leq \frac{(4 - t_i)\sqrt{k_\ell}}{4 - 3t_i}\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\|;
\end{aligned}$$

the third inequality follows from $1 - (1 - t_i)(4 - t_i)/(4 - 3t_i) \geq 0.5t_i$ for $t_i \in [0.5, 1]$ and the fourth inequality follows from the assumption in Eq. (20) and $t_i \geq 0.5$.

If $\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}^T A_i \Delta\hat{\mathbf{w}}_i \leq 0$, then $\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\| = 0$. If $\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}^T A_i \Delta\hat{\mathbf{w}}_i > 0$, then we have

$$\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}^T A_i \Delta\hat{\mathbf{w}}_i \geq 0.25\rho_{min}^-(2\bar{r} + s)\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\|^2. \tag{22}$$

By letting $\mathbf{v} = \Delta\hat{\mathbf{w}}_{\mathcal{I}_i}$, we obtain the following from Eq. (17):

$$
\begin{aligned}
&2\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}^T A_i \Delta\hat{\mathbf{w}}_i = -2\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}^T \bar{\boldsymbol{\epsilon}}_i - \sum_{(j,i)\in\mathcal{I}_i} \hat{\lambda}_{ji}\Delta\hat{w}_{ji}\text{sign}(\hat{w}_{ji}) \\
&= -2\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}^T \bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i^c} - 2\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}^T \bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i} - \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}\Delta\hat{w}_{ji}\text{sign}(\hat{w}_{ji}) - \sum_{(j,i)\in\mathcal{J}_i} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| - \sum_{(j,i)\in\bar{\mathcal{F}}_i^c\cap\mathcal{G}_i^c} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| \\
&= -2\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}^T \bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i^c} - 2\Delta\hat{\mathbf{w}}_{\mathcal{J}_i}^T \bar{\boldsymbol{\epsilon}}_{\mathcal{J}_i} - \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}\Delta\hat{w}_{ji}\text{sign}(\hat{w}_{ji}) - \sum_{(j,i)\in\mathcal{J}_i} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| - \sum_{(j,i)\in\bar{\mathcal{F}}_i^c\cap\mathcal{G}_i^c} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| \\
&\leq 2\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\|\|\bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i^c}\| + 2\|\bar{\boldsymbol{\epsilon}}_{\mathcal{J}_i}\|_\infty \sum_{(j,i)\in\mathcal{J}_i} |\Delta\hat{w}_{ji}| + \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| - \sum_{(j,i)\in\mathcal{J}_i} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| \\
&\leq 2\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\|\|\bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i^c}\| + \left(\sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}^2\right)^{1/2} \|\Delta\hat{\mathbf{w}}_{\bar{\mathcal{F}}_i}\| \\
&\leq 2\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\|\|\bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i^c}\| + \left(\sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}^2\right)^{1/2} \|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\|. \qquad (23)
\end{aligned}
$$

In the above derivation, the second equality is due to $\mathcal{I}_i = \mathcal{J}_i \cup \bar{\mathcal{F}}_i \cup (\bar{\mathcal{F}}_i^c \cap \mathcal{G}_i^c)$; the third equality is due to $\mathcal{I}_i \cap \mathcal{G}_i = \mathcal{J}_i$; the second inequality follows from $\forall (j,i) \in \mathcal{J}_i, \hat{\lambda}_{ji} = \lambda \geq 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty \geq 2\|\bar{\boldsymbol{\epsilon}}_{\mathcal{J}_i}\|_\infty$ and the last inequality follows from $\bar{\mathcal{F}}_i \subseteq \mathcal{G}_i^c \subseteq \mathcal{I}_i$. Combining Eq. (22) and Eq. (23), we have

$$
\|\Delta\hat{\mathbf{w}}_{\mathcal{I}_i}\| \leq \frac{2}{\rho_{min}^-(2\bar{r}+s)} \left[2\|\bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i^c}\| + \left(\sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}^2\right)^{1/2}\right].
$$

Notice that

$$
\|\mathbf{x}_i\| \leq a(\|\mathbf{y}_i\| + \|\mathbf{z}_i\|) \Rightarrow \|X\|_{2,1}^2 \leq m\|X\|_F^2 = m\sum_i \|\mathbf{x}_i\|^2 \leq 2ma^2(\|Y\|_F^2 + \|Z\|_F^2).
$$

Thus, we have

$$
\|\Delta\hat{W}_{\mathcal{I}}\|_{2,1} \leq \frac{\sqrt{8m\left(4\|\bar{\Upsilon}_{\mathcal{G}_{(\ell)}^c}\|_F^2 + \sum_{(j,i)\in\bar{\mathcal{F}}} (\hat{\lambda}_{ji}^{(\ell-1)})^2\right)}}{\rho_{min}^-(2\bar{r}+s)}. \qquad (24)
$$

Therefore, at stage ℓ, Eq. (12) in Lemma 3 directly follows from Eq. (21) and Eq. (24). Following Eq. (12), we have:

$$
\begin{aligned}
&\|\hat{W}^{(\ell)} - \bar{W}\|_{2,1} = \|\Delta\hat{W}^{(\ell)}\|_{2,1} \\
&\leq \frac{\left(1 + 1.5\sqrt{\frac{2\bar{r}}{s}}\right)\sqrt{8m\left(4\|\bar{\Upsilon}_{\mathcal{G}_{(\ell)}^c}\|_F^2 + \sum_{(j,i)\in\bar{\mathcal{F}}} (\hat{\lambda}_{ji}^{(\ell-1)})^2\right)}}{\rho_{min}^-(2\bar{r}+s)} \\
&\leq \frac{8.83\sqrt{m}\sqrt{4\|\Upsilon\|_{\infty,\infty}^2|\mathcal{G}_{(\ell)}^c| + \bar{r}m\lambda^2}}{\rho_{min}^-(2\bar{r}+s)} \\
&\leq \frac{8.83\sqrt{m}\lambda\sqrt{\frac{8}{144}\bar{r}m + \bar{r}m}}{\rho_{min}^-(2\bar{r}+s)} \leq \frac{9.1m\lambda\sqrt{\bar{r}}}{\rho_{min}^-(2\bar{r}+s)},
\end{aligned}
$$

where the first inequality is due to Eq. (24); the second inequality is due to $s \geq \bar{r}$ (assumption in Theorem 1), $\hat{\lambda}_{ji} \leq \lambda$, $\bar{r}m = |\bar{\mathcal{H}}| \geq |\bar{\mathcal{F}}|$ and the third inequality follows from Eq. (20) and $\|\bar{\Upsilon}\|_{\infty,\infty}^2 \leq (1/144)\lambda^2$. Therefore, Eq. (13) in Lemma 3 holds at stage ℓ.

Notice that we obtain Lemma 3 at stage ℓ, by assuming that Eq. (20) is satisfied. To prove that Lemma 3 holds for all stages, we next need to prove by induction that Eq. (20) holds at all stages.

When $\ell = 1$, we have $\mathcal{G}^c_{(1)} = \bar{\mathcal{H}}$, which implies that Eq. (20) holds. Now, we assume that Eq. (20) holds at stage ℓ. Thus, by hypothesis induction, we have:

$$\sqrt{|\mathcal{G}^c_{(\ell+1)} \setminus \bar{\mathcal{H}}|} \leq \sqrt{m\theta^{-2}\|\hat{W}^{(\ell)}_{\mathcal{G}^c_{(\ell+1)}\setminus\bar{\mathcal{H}}} - \bar{W}_{\mathcal{G}^c_{(\ell+1)}\setminus\bar{\mathcal{H}}}\|^2_{2,1}}$$
$$\leq \sqrt{m}\theta^{-1}\left\|\hat{W}^{(\ell)} - \bar{W}\right\|_{2,1} \leq \frac{9.1m^{3/2}\lambda\sqrt{\bar{r}}\theta^{-1}}{\rho^-_{min}(2\bar{r}+s)} \leq \sqrt{\bar{r}m},$$

where θ is the thresholding parameter in Eq. (1); the first inequality above follows from the definition of $\mathcal{G}_{(\ell)}$ in Lemma 3:

$$\forall (j,i) \in \mathcal{G}^c_{(\ell+1)} \setminus \bar{\mathcal{H}}, \|(\hat{\mathbf{w}}^{(\ell)})^j\|^2_1/\theta^2 = \|(\hat{\mathbf{w}}^{(\ell)})^j - \bar{\mathbf{w}}^j\|^2_1/\theta^2 \geq 1$$
$$\Rightarrow |\mathcal{G}^c_{(\ell+1)} \setminus \bar{\mathcal{H}}| \leq m\theta^{-2}\|\hat{W}^{(\ell)}_{\mathcal{G}^c_{(\ell+1)}\setminus\bar{\mathcal{H}}} - \bar{W}_{\mathcal{G}^c_{(\ell+1)}\setminus\bar{\mathcal{H}}}\|^2_{2,1};$$

the last inequality is due to Eq. (6). Thus, we have:

$$|\mathcal{G}^c_{(\ell+1)} \setminus \bar{\mathcal{H}}| \leq \bar{r}m \Rightarrow |\mathcal{G}^c_{(\ell+1)}| \leq 2\bar{r}m \Rightarrow k_{\ell+1} \leq 2\bar{r}.$$

Therefore, Eq. (20) holds at all stages. Thus the two inequalities in Lemma 3 hold at all stages. This completes the proof of the lemma. □

A.4 Proof of Lemma 4

Proof The first inequality directly follows from $\bar{\mathcal{H}} \supseteq \bar{\mathcal{F}}$. Next, we focus on the second inequality. For each $(j,i) \in \bar{\mathcal{F}}$ $(\bar{\mathcal{H}})$, if $\|\hat{\mathbf{w}}^j\|_1 < \theta$, by considering Eq. (3), we have

$$\|\bar{\mathbf{w}}^j - \hat{\mathbf{w}}^j\|_1 \geq \|\bar{\mathbf{w}}^j\|_1 - \|\hat{\mathbf{w}}^j\|_1 \geq 2\theta - \theta = \theta.$$

Therefore, we have for each $(j,i) \in \bar{\mathcal{F}}$ $(\bar{\mathcal{H}})$:

$$I\left(\|\hat{\mathbf{w}}^j\|_1 < \theta\right) \leq \|\bar{\mathbf{w}}^j - \hat{\mathbf{w}}^j\|_1/\theta.$$

Thus, the second inequality of Lemma 4 directly follows from the above inequality. □

B. Lemmas from Zhang (2010) [26]

Lemma 5 *Let $\mathbf{a} \in \mathbb{R}^n$ be a fixed vector and $\mathbf{x} \in \mathbb{R}^n$ be a random vector which is composed of independent sub-Gaussian components with parameter σ. Then we have:*

$$\Pr(|\mathbf{a}^T\mathbf{x}| \geq t) \leq 2\exp\left(-t^2/(2\sigma^2\|\mathbf{a}\|^2)\right), \forall t > 0.$$

Lemma 6 $\pi_i(k_i, s_i) \leq \frac{s_i^{1/2}}{2}\sqrt{\rho^+_i(s_i)/\rho^-_i(k_i+s_i) - 1}$.

Lemma 7 *Let $\mathcal{G}_i \subseteq \mathbb{N}_d \times \{i\}$ such that $|\mathcal{G}^c_i| = k_i$, and let $\mathcal{J}_i$ be indices of the s_i largest components (in absolute values) of $\mathbf{w}_{\mathcal{G}_i}$ and $\mathcal{I}_i = \mathcal{G}^c_i \cup \mathcal{J}_i$. Then for any $\mathbf{w}_i \in \mathbb{R}^d$, we have*

$$\max(0, \mathbf{w}^T_{\mathcal{I}_i} A_i \mathbf{w}_i) \geq \rho^-_i(k_i+s_i)(\|\mathbf{w}_{\mathcal{I}_i}\| - \pi_i(k_i+s_i, s_i)\|\mathbf{w}_{\mathcal{G}_i}\|_1/s_i)\|\mathbf{w}_{\mathcal{I}_i}\|.$$

Lemma 8 *Let $\bar{\boldsymbol{\epsilon}}_i = [\bar{\epsilon}_{1i}, \cdots, \bar{\epsilon}_{di}] = \frac{1}{n}X_i^T(X_i\bar{\mathbf{w}}_i - \mathbf{y}_i)$ $(i \in \mathbb{N}_m)$, and $\bar{\mathcal{H}}_i \subseteq \mathbb{N}_d \times \{i\}$. Under the conditions of Assumption 1, the followings hold with probability larger than $1-\eta$:*

$$\|\bar{\boldsymbol{\epsilon}}_{\bar{\mathcal{H}}_i}\|^2 \leq \sigma^2\rho^+_i(|\bar{\mathcal{H}}_i|)(7.4|\bar{\mathcal{H}}_i| + 2.7\ln(2/\eta))/n.$$