Supplementary Material for 'Learning Multi-Level Task Groups in Multi-Task Learning'

A. Some Basic Lemmas

Before presenting the proofs in the main paper, we first provide some basic lemmas.

Lemma 1 $\|\mathbf{C}\mathbf{W}_h^T\|_{1,2} \leq (m-1)\sqrt{d}\|\mathbf{W}_h\|_F$.

Proof: Note that for any matrix $\mathbf{Q} \in \mathbb{R}^{r_1 \times r_2}$, $\|\mathbf{Q}\|_{1,2} \leq \sqrt{r}\|\mathbf{Q}\|_F$, where r is the rank of $\mathbf{Q}$ and we have $r \leq r_1, r \leq r_2$. Based on the definition of the matrix $\mathbf{C}$, we have

$$\begin{aligned}\|\mathbf{C}\mathbf{W}_h^T\|_{1,2} &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j\neq i}^{m}\|\mathbf{w}_{h,i}-\mathbf{w}_{h,j}\|_2 \\ &\leq \frac{1}{2}\sum_{i=1}^{m}\sum_{j\neq i}^{m}(\|\mathbf{w}_{h,i}\|_2+\|\mathbf{w}_{h,j}\|_2) \\ &= (m-1)\|\mathbf{W}_h^T\|_{1,2} \leq (m-1)\sqrt{d}\|\mathbf{W}_h\|_F,\end{aligned}$$

in which we complete the proof. □

Lemma 2 *For any matrix pair* $\mathbf{Q}, \hat{\mathbf{Q}} \in \mathbb{R}^{d\times m}$*, we have*

$$\|\mathbf{C}\mathbf{Q}^T\|_{1,2} - \|\mathbf{C}\hat{\mathbf{Q}}^T\|_{1,2} \leq \|(\mathbf{C}\mathbf{Q}^T - \mathbf{C}\hat{\mathbf{Q}}^T)^{E(\mathbf{Q})}\|_{1,2}.$$

The proof of Lemma 2 is the same with the proof of Lemma 1 in (Gong, Ye, and Zhang 2012) and is omitted here.

Lemma 3 *Assume that the training data is normalized with zero mean and unit variance. For* $h \in \mathbb{N}_H$*, if the regularization parameter* λ_h *satisfies Eq. (14), then with probability of at least* $1-\exp(-\frac{1}{2}(\delta - dm\log(1+\frac{\delta}{dm})))$*, for an optimal solution* $\hat{\mathbf{W}} = \sum_{h=1}^{H}\hat{\mathbf{W}}_h$ *of problem (3) and any* $\mathbf{W} = \sum_{h=1}^{H}\mathbf{W}_h \in \mathbb{R}^{d\times m}$*, we have*

$$\begin{aligned}&\frac{1}{mn}\|\mathbf{X}^T\mathrm{vec}(\hat{\mathbf{W}}) - \mathrm{vec}(\mathbf{F}^*)\|_2^2 \leq \frac{1}{mn}\|\mathbf{X}^T\mathrm{vec}(\mathbf{W}) - \mathrm{vec}(\mathbf{F}^*)\|_2^2 \\ &+ (m-1)\sqrt{d}\sum_{h=1}^{H}\lambda_h(\theta_h+1)\|\left(\hat{\mathbf{W}}_h - \mathbf{W}_h\right)^{D(\mathbf{W}_h)}\|_F. \quad (20)\end{aligned}$$

B. Long Proofs

Proof of Lemma 3: Since $\hat{\mathbf{W}}$ is an optimal solution of problem (3), for any $\mathbf{W}$ we have

$$\begin{aligned}\frac{1}{mn}\sum_{i=1}^{m}\|\mathbf{X}_i^T\sum_{h=1}^{H}\hat{\mathbf{w}}_{h,i} - \mathbf{y}_i\|_2^2 &\leq \frac{1}{mn}\sum_{i=1}^{m}\|\mathbf{X}_i^T\sum_{h=1}^{H}\mathbf{w}_{h,i} - \mathbf{y}_i\|_2^2 \\ &+ \sum_{h=1}^{H}\lambda_h\left(\|\mathbf{C}\mathbf{W}_h^T\|_{1,2} - \|\mathbf{C}\hat{\mathbf{W}}_h^T\|_{1,2}\right)\end{aligned}$$

Substituting Eq. (13) into this inequality, we can obtain

$$\begin{aligned}\frac{1}{mn}\sum_{i=1}^{m}\|\mathbf{X}_i^T\sum_{h=1}^{H}\hat{\mathbf{w}}_{h,i} - \mathbf{f}_i^*\|_2^2 &\leq \frac{1}{mn}\sum_{i=1}^{m}\|\mathbf{X}_i^T\sum_{h=1}^{H}\mathbf{w}_{h,i} - \mathbf{f}_i^*\|_2^2 \\ &+ \sum_{h=1}^{H}\lambda_h\left(\|\mathbf{C}\mathbf{W}_h^T\|_{1,2} - \|\mathbf{C}\hat{\mathbf{W}}_h^T\|_{1,2}\right) \\ &+ \frac{2}{mn}\sum_{h=1}^{H}\left\langle\mathbf{Z}, \hat{\mathbf{W}}_h - \mathbf{W}_h\right\rangle, \quad (21)\end{aligned}$$

where $\mathbf{Z} = [\mathbf{X}_1\boldsymbol{\epsilon}_1, \cdots, \mathbf{X}_m\boldsymbol{\epsilon}_m] \in \mathbb{R}^{d\times m}$ with its (j,i)th element computed as $z_{ji} = \sum_{k=1}^{n} x_{ji}^{(i)}\epsilon_{ki}$ and $x_{jk}^{(i)}$ denotes the (j,i)th element in $\mathbf{X}_i$ for the ith task. Since $\mathbf{x}_j^{(i)}$ is normalized with zero mean and unit variance and $\epsilon_{ji} \sim \mathcal{N}(0,\sigma^2)$, we have

$$z_{ji} \sim \mathcal{N}(0,\sigma^2).$$

By defining a variable $v_{ji} = \frac{1}{\sigma}z_{ji}$, we can get that $v_{ji} \sim \mathcal{N}(0,1)$. Thus we can get that a variable u with the definition as

$$u = \sum_{j=1}^{d}\sum_{i=1}^{m}v_{ji}^2 = \frac{1}{\sigma^2}\|\mathbf{Z}\|_F^2$$

follows a chi-squared distribution with the degree of freedom as dm. According to the Wallace inequality (Wallace 1959), for any $\delta > 0$ we have

$$\Pr(u \geq dm+\delta) \leq \exp\left(-\frac{1}{2}\left(\delta - dm\log\left(1+\frac{\delta}{dm}\right)\right)\right).$$

Since $u = \frac{1}{\sigma^2}\|\mathbf{Z}\|_F^2$, we obtain that

$$\begin{aligned}\Pr\left(\frac{2}{mn}\|\mathbf{Z}\|_F \leq \frac{2\sigma}{mn}\sqrt{dm+\delta}\right) &= \Pr\left(u \leq dm+\delta\right) \\ &\geq 1-\exp\left(-\frac{1}{2}\left(\delta - dm\log\left(1+\frac{\delta}{dm}\right)\right)\right).\end{aligned} \quad (22)$$

Based on Assumption 1 and Eq. (22), with probability of at least $1-\exp(-\frac{1}{2}(\delta - dm\log(1+\frac{\delta}{dm})))$ we have

$$\begin{aligned}\frac{2}{mn}\sum_{h=1}^{H}\left\langle\mathbf{Z}, \hat{\mathbf{W}}_h - \mathbf{W}_h\right\rangle &\leq \frac{2}{mn}\|\mathbf{Z}\|_F\sum_{h=1}^{H}\|\hat{\mathbf{W}}_h - \mathbf{W}_h\|_F \\ &\leq \frac{2\sigma}{mn}\sqrt{dm+\delta}\sum_{h=1}^{H}\theta_h\|\left(\hat{\mathbf{W}}_h - \mathbf{W}_h\right)^{D(\mathbf{W}_h)}\|_F.\end{aligned} \quad (23)$$

Moreover, by using Lemma 1 and 2, we have

$$\begin{aligned}\|\mathbf{C}\mathbf{W}_h^T\|_{1,2} - \|\mathbf{C}\hat{\mathbf{W}}_h^T\|_{1,2} &\leq \|\left(\mathbf{C}\mathbf{W}_h^T - \mathbf{C}\hat{\mathbf{W}}_h^T\right)^{E(\mathbf{W}_h)}\|_{1,2} \\ &\leq (m-1)\sqrt{d}\|\left(\mathbf{W}_h - \hat{\mathbf{W}}_h\right)^{D(\mathbf{W}_h)}\|_F. \quad (24)\end{aligned}$$

Combing Eqs. (21), (23), and (24), with probability of at least $1-\exp(-\frac{1}{2}(\delta - dm\log(1+\frac{\delta}{dm})))$ we have

$$\begin{aligned}&\frac{1}{mn}\|\mathbf{X}^T\mathrm{vec}(\hat{\mathbf{W}}) - \mathrm{vec}(\mathbf{F}^*)\|_2^2 \leq \frac{1}{mn}\|\mathbf{X}^T\mathrm{vec}(\mathbf{W}) - \mathrm{vec}(\mathbf{F}^*)\|_2^2 \\ &+ \sum_{h=1}^{H}\left(\frac{2\sigma}{mn}\sqrt{dm+\delta}\theta_h + (m-1)\sqrt{d}\lambda_h\right)\|\left(\hat{\mathbf{W}}_h - \mathbf{W}_h\right)^{D(\mathbf{W}_h)}\|_F.\end{aligned}$$

Plugging Eq. (14) into the above equation, we complete the proof. □

Proof of Theorem 1: Let $\mathbf{W}_h = \mathbf{W}_h^*$ for $h \in \mathbb{N}_H$ in Eq. (20) and so $\boldsymbol{\Delta}_h = \hat{\mathbf{W}}_h - \mathbf{W}_h^*$. Then we obtain

$$\frac{1}{mn}\|\mathbf{X}^T\mathrm{vec}(\boldsymbol{\Delta})\|_2^2 \leq (m-1)\sqrt{d}\sum_{h=1}^{H}\lambda_h(\theta_h+1)\|\boldsymbol{\Delta}_h^{D(\mathbf{W}_h)}\|_F. \quad (25)$$

Under Assumption 1, we have

$$\|\mathbf{\Delta}_h^{D(\mathbf{W}_h)}\|_F \leq \frac{\|\mathbf{X}^T \mathrm{vec}(\mathbf{\Delta})\|_2}{\beta_h \sqrt{mn}} \tag{26}$$

By substituting Eq. (26) into Eq. (25), we obtain

$$\|\mathbf{X}^T \mathrm{vec}(\mathbf{\Delta})\|_2 \leq (m-1)\sqrt{mnd}\mathcal{H}. \tag{27}$$

Therefore we can directly get Eq. (15) from Eq. (27). Since from Assumption 1, we have

$$\|\hat{\mathbf{W}}_h - \mathbf{W}_h^*\|_F = \theta_h \| \left(\hat{\mathbf{W}}_h - \mathbf{W}_h^*\right)^{D(\mathbf{W}_h)} \|_F,$$

$$\|\mathbf{C}\hat{\mathbf{W}}_h^T - \mathbf{C}(\mathbf{W}_h^*)^T\|_{1,2} = \gamma_h \| \left(\mathbf{C}\hat{\mathbf{W}}_h^T - \mathbf{C}(\mathbf{W}_h^*)^T\right)^{E(\mathbf{W}_h)} \|_{1,2}.$$

Combing Eqs. (24), (26) and (15), we can easily prove Eqs. (16) and (17).

To prove $\hat{E}_h = E(\mathbf{W}_h^*)$, we need to prove the following two statements:

$$\forall (i,j) \in \hat{E}_h \Rightarrow (i,j) \in E(\mathbf{W}_h^*), \tag{28}$$

$$\forall (i,j) \in E(\mathbf{W}_h^*) \Rightarrow (i,j) \in \hat{E}_h. \tag{29}$$

We first prove Eq. (28) by contradiction. Assume there exists a pair (i',j') such that $(i',j') \in \hat{E}_h$, but $(i',j') \notin E(\mathbf{W}_h^*)$. Then according to the definitions of $\hat{E}_h$ and $E(\mathbf{W}_h^*)$, we have

$$\begin{aligned} \| \left(\mathbf{C}\hat{\mathbf{W}}_h^T - \mathbf{C}(\mathbf{W}_h^*)^T\right)^{(i',j')} \|_2 &= \| \left(\mathbf{C}\hat{\mathbf{W}}_h^T\right)^{(i',j')} \|_2 \\ &> \frac{\gamma_h (m-1)^2 d\mathcal{H}}{\beta_h}, \end{aligned}$$

which contradicts with the proved Eq. (17), so we prove Eq. (28). Next we prove Eq. (29) by contradiction. Similarly, assume there exists $(i'',j'') \in E(\mathbf{W}_h^*)$, but $(i'',j'') \notin \hat{E}_h$. Since $(i'',j'') \notin \hat{E}_h$, based on the definition of $\hat{E}_h$ in Eq. (19) we have

$$\| \left(\mathbf{C}\hat{\mathbf{W}}_h^T\right)^{(i'',j'')} \|_2 \leq \frac{\gamma_h (m-1)^2 d\mathcal{H}}{\beta_h}.$$

Furthermore, using the condition in Eq. (18), we have

$$\begin{aligned} &\| \left(\mathbf{C}\hat{\mathbf{W}}_h^T - \mathbf{C}(\mathbf{W}_h^*)^T\right)^{(i'',j'')} \|_2 \geq \| \left(\mathbf{C}(\mathbf{W}_h^*)^T\right)^{(i'',j'')} \|_2 \\ &- \| \left(\mathbf{C}\hat{\mathbf{W}}_h^T\right)^{(i'',j'')} \|_2 > \frac{\gamma_h (m-1)^2 d\mathcal{H}}{\beta_h}. \end{aligned}$$

which contradicts with Eq. (17). So Eq. (29) is correct, which completes the proof. □

Details and Proof for Remark 2: In the robust multi-task learning (rMTL) (Gong, Ye, and Zhang 2012), the parameter matrix $\mathbf{W}$ is decomposed into two components $\mathbf{W}_1$ and $\mathbf{W}_2$. The performance bound of their model is

$$\|\mathbf{X}^T \mathrm{vec}\left(\hat{\mathbf{W}}^{(\mathrm{r})}\right) - \mathrm{vec}(\mathbf{F}^*)\|_2 \leq \sqrt{mn}\left(\frac{2\eta_1\sqrt{r}}{\kappa_1} + \frac{2\eta_2\sqrt{c}}{\kappa_2}\right), \tag{30}$$

where $\hat{\mathbf{W}}^{(\mathrm{r})}$ is the estimator from rMTL, and η_1 and η_2 are the regularization parameters that satisfy $\eta_1, \eta_2 \geq \frac{2\sigma}{mn}\sqrt{dm+\delta}$, $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ are the ground truth for $\mathbf{W}_1$ and $\mathbf{W}_2$ respectively, r is the number of non-zero rows in $\mathbf{W}_1^*$, c is the number of non-zero columns in $\mathbf{W}_2^*$, and κ_1 and κ_2 are also some parameters defined in the restrict eigenvalue assumption as

$$\kappa_1 = \min_{\mathbf{\Delta}_1 \neq \mathbf{0}} \frac{\|\mathbf{X}^T \mathrm{vec}(\mathbf{\Delta})\|_2}{\sqrt{mn}\|\mathbf{\Delta}_1^{\mathcal{J}(\mathbf{W}_1^*)}\|_F}, \kappa_2 = \min_{\mathbf{\Delta}_2 \neq \mathbf{0}} \frac{\|\mathbf{X}^T \mathrm{vec}(\mathbf{\Delta})\|_2}{\sqrt{mn}\|\mathbf{\Delta}_2^{\mathcal{J}(\mathbf{W}_2^{*T})}\|_F}.$$

By denoting by $\mathcal{J}(\mathbf{Q})$ the set of the indices of the non-zero rows in a matrix $\mathbf{Q}$, we have $r = |\mathcal{J}(\mathbf{W}_1^*)|$ and $c = |\mathcal{J}(\mathbf{W}_2^{*T})|$ where $|\cdot|$ denotes the cardinality of a set. By setting the number of task levels H to be 2 in our MeTaG model, we have the bound as

$$\begin{aligned} &\|\mathbf{X}^T \mathrm{vec}(\hat{\mathbf{W}}) - \mathrm{vec}(\mathbf{F}^*)\|_2 \\ &\leq (m-1)\sqrt{mnd}\left(\frac{\lambda_1(\theta_1+1)}{\beta_1} + \frac{\lambda_2(\theta_2+1)}{\beta_2}\right). \end{aligned} \tag{31}$$

A direct comparison between the two bounds in Eqs. (30) and (31) is difficult due to the use of different projection sets $\mathcal{J}(\cdot)$ and $D(\cdot)$. However, we can compare those two bounds in the worst case where $\mathcal{J}_c(\mathbf{W}_1^*) = \emptyset$, $\mathcal{J}_c(\mathbf{W}_2^{*T}) = \emptyset$, $D_c(\mathbf{W}_1^*) = \emptyset$, and $D_c(\mathbf{W}_2^*) = \emptyset$. In the worst case, we can get that $\kappa_1 = \beta_1$, $\kappa_2 = \beta_2$, $r = d$, $c = m$, $\theta_1 = 1$ and $\theta_2 = 1$. Then the bound in Eq. (30) can be rewritten as

$$\begin{aligned} &\|\mathbf{X}^T \mathrm{vec}\left(\hat{\mathbf{W}}^{(\mathrm{r})}\right) - \mathrm{vec}(\mathbf{F}^*)\|_2 \\ &= O\left(\frac{4\sigma\sqrt{dm+\delta}}{\sqrt{mn}}\left(\frac{\sqrt{d}}{\beta_1} + \frac{\sqrt{m}}{\beta_2}\right)\right), \end{aligned} \tag{32}$$

and the bound in Eq. (31) can be rewritten as

$$\|\mathbf{X}^T \mathrm{vec}(\hat{\mathbf{W}}) - \mathrm{vec}(\mathbf{F}^*)\|_2 = O\left(\frac{4\sigma\sqrt{dm+\delta}}{\sqrt{mn}}\left(\frac{1}{\beta_1} + \frac{1}{\beta_2}\right)\right). \tag{33}$$

By comparing those two bounds, we can observe that the bound of our MeTaG method is considerably better than that of the rMTL method in the worst case especially when the feature dimension d and number of tasks m are large. □