# Supplementary Material for 'Learning Tree Structure in Multi-Task Learning'

## A. Basic Lemmas for Proving Theorem 3

Before presenting the proof for Theorem 3, we first prove some useful lemmas.

LEMMA 3. $\|\mathbf{C}\mathbf{W}_h^T\|_{1,2} \le (m-1)\sqrt{d}\|\mathbf{W}_h\|_F$.

**Proof:** For any matrix $\mathbf{A} \in \mathbb{R}^{r_1 \times r_2}$, we have $\|\mathbf{A}\|_{1,2} \le \sqrt{r}\|\mathbf{A}\|_F$, where $r \le \min(r_1, r_2)$ denotes the rank of $\mathbf{A}$ and the inequality holds due to the Cauchy-Schwarz inequality. Based on the definition of the matrix $\mathbf{C}$, we have

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

where the first inequality holds due to the triangular inequality for vector norms. So we complete the proof. ■

LEMMA 4. *For any matrix pair $\mathbf{A}, \hat{\mathbf{A}} \in \mathbb{R}^{d\times m}$, we have*

$$\|\mathbf{C}\mathbf{A}^T\|_{1,2} - \|\mathbf{C}\hat{\mathbf{A}}^T\|_{1,2} \le \left\|(\mathbf{C}\mathbf{A}^T - \mathbf{C}\hat{\mathbf{A}}^T)^{E(\mathbf{A})}\right\|_{1,2}.$$

The proof of Lemma 4 is similar to that of Lemma 1 in [1] and hence we omit it here.

LEMMA 5. *Assume that the training data is normalized to have zero mean and unit variance. For $h \in \mathbb{N}_H$, if the regularization parameter $\lambda_h$ satisfies Eq. (19), then with probability 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 $\mathbf{W} = \sum_{h=1}^H \mathbf{W}_h \in \mathbb{R}^{d\times m}$, where $\{\mathbf{W}_h\}_{h=1}^H$ satisfy the sequential constraints, we have*

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

**Proof**. Since $\hat{\mathbf{W}}$ is an optimal solution of problem (3), $\{\hat{\mathbf{W}}_h\}_{h=1}^H$ satisfy the sequential constraints, and for any $\mathbf{W} = \sum_{h=1}^H \mathbf{W}_h$ satisfying the constraints too, we have

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

By substituting $y_{ji} = (\mathbf{x}_j^{(i)})^T\mathbf{w}_i^* + \epsilon_{ji}, i \in \mathbb{N}_m, j \in \mathbb{N}_n$ into the above inequality, we can obtain

$$\frac{1}{mn}\sum_{i=1}^m \|\mathbf{X}_i^T\sum_{h=1}^H \hat{\mathbf{w}}_{h,i} - \mathbf{f}_i^*\|_2^2 \le \frac{1}{mn}\sum_{i=1}^m \|\mathbf{X}_i^T\sum_{h=1}^H \mathbf{w}_{h,i} - \mathbf{f}_i^*\|_2^2 + \sum_{h=1}^H \lambda_h\left(\|\mathbf{C}\mathbf{W}_h^T\|_{1,2} - \|\mathbf{C}\hat{\mathbf{W}}_h^T\|_{1,2}\right) + \frac{2}{mn}\sum_{h=1}^H \left\langle \mathbf{Z}, \hat{\mathbf{W}}_h - \mathbf{W}_h\right\rangle, \quad (26)$$

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 entry computed as $z_{ji} = \sum_{k=1}^n x_{ji}^{(i)}\epsilon_{ki}$ and $x_{jk}^{(i)}$ denotes the $(j,i)$th entry in $\mathbf{X}_i$ for the $i$th task. Since $\mathbf{x}_j^{(i)}$ is normalized to have 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

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

which follows a chi-squared distribution with the degree of freedom as $md$. According to the Wallace inequality [2], for any $\delta > 0$ we have

$$\Pr(u \ge md + \delta) \le \exp\left(-\frac{1}{2}\left(\delta - md\log\left(1 + \frac{\delta}{md}\right)\right)\right).$$

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

$$\begin{aligned}&\Pr\left(\frac{2}{mn}\|\mathbf{Z}\|_F \le \frac{2\sigma}{mn}\sqrt{md+\delta}\right) \\ &= \Pr(u \le md+\delta) \\ &\ge 1 - \exp\left(-\frac{1}{2}\left(\delta - md\log\left(1+\frac{\delta}{md}\right)\right)\right).\end{aligned} \quad (27)$$

Based on Assumption 1 and Eq. (27), with probability at least $1 - \exp(-\frac{1}{2}(\delta - md\log(1+\frac{\delta}{md})))$ we have

$$\begin{aligned}&\frac{2}{mn}\sum_{h=1}^H \left\langle \mathbf{Z}, \hat{\mathbf{W}}_h - \mathbf{W}_h\right\rangle \\ &\le \frac{2}{mn}\|\mathbf{Z}\|_F\sum_{h=1}^H \|\hat{\mathbf{W}}_h - \mathbf{W}_h\|_F \\ &\le \frac{2\sigma}{mn}\sqrt{md+\delta}\sum_{h=1}^H \theta_h\left\|\left(\hat{\mathbf{W}}_h - \mathbf{W}_h\right)^{D(\mathbf{W}_h)}\right\|_F.\end{aligned} \quad (28)$$

Moreover, by using Lemma 3 and 4, we have

$$\begin{aligned}&\|\mathbf{C}\mathbf{W}_h^T\|_{1,2} - \|\mathbf{C}\hat{\mathbf{W}}_h^T\|_{1,2} \\ &\le \left\|\left(\mathbf{C}\mathbf{W}_h^T - \mathbf{C}\hat{\mathbf{W}}_h^T\right)^{E(\mathbf{W}_h)}\right\|_{1,2} \\ &\le (m-1)\sqrt{d}\left\|\left(\mathbf{W}_h - \hat{\mathbf{W}}_h\right)^{D(\mathbf{W}_h)}\right\|_F.\end{aligned} \quad (29)$$

By combing Eqs. (26), (28), and (29), with probability at least $1 - \exp(-\frac{1}{2}(\delta - md\log(1+\frac{\delta}{md})))$ we have

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

By plugging Eq. (19) into the above equation, we complete the proof. ■

## B. Proof of Theorem 3

**Proof**. By making $\mathbf{W}_h$ take value of $\mathbf{W}_h^*$ for $h \in \mathbb{N}_H$ in Eq. (25), we obtain

$$\frac{1}{mn}\|\mathbf{X}^T\mathrm{vec}(\boldsymbol{\Delta})\|_2^2 \le (m-1)\sqrt{d}\sum_{h=1}^H \lambda_h(\theta_h+1)\left\|\boldsymbol{\Delta}_h^{D(\mathbf{W}_h)}\right\|_F, \quad (30)$$

where $\mathbf{\Delta}_h = \hat{\mathbf{W}}_h - \mathbf{W}_h^*$ and $\mathbf{\Delta} = \sum_{h=1}^H \mathbf{\Delta}_h$. Under Assumption 1, we have

$$\left\|\mathbf{\Delta}_h^{D(\mathbf{W}_h)}\right\|_F \leq \frac{\left\|\mathbf{X}^T \text{vec}(\mathbf{\Delta})\right\|_2}{\beta_h\sqrt{mn}}. \tag{31}$$

By substituting Eq. (31) into Eq. (30), we obtain

$$\|\mathbf{X}^T \text{vec}(\mathbf{\Delta})\|_2 \leq (m-1)\sqrt{mnd}\mathcal{C}. \tag{32}$$

Therefore we can directly get Eq. (20) from Eq. (32). Since from Assumption 1, we have

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

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

By combing Eqs. (29), (31), and (20), we can easily prove Eqs. (21) and (22).

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

$$\forall (i,j) \in \hat{E}_h \Rightarrow (i,j) \in E(\mathbf{W}_h^*), \tag{33}$$

$$\forall (i,j) \in E(\mathbf{W}_h^*) \Rightarrow (i,j) \in \hat{E}_h. \tag{34}$$

We first prove Eq. (33) 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

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

which contradicts Eq. (22), hence we prove Eq. (33). Next we prove Eq. (34) 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. (24) we have

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

Furthermore, using the condition in Eq. (23), we have

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

which contradicts Eq. (22). So Eq. (34) is correct, which completes the proof. ■

# References


[1] P. Gong, J. Ye, and C. Zhang. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 895–903, 2012.

[2] D. L. Wallace. Bounds on normal approximations to student's and the chi-square distributions. *The Annals of Mathematical Statistics*, pages 1121–1130, 1959.