Supplementary Material for Submission "Jointly Learning Heterogeneous Features for RGB-D Activity Recognition"

Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang

Abstract—In this supplementary document, we provide a more detailed proof of the convergence of our optimization algorithm for the proposed **j**oint heter**o**geneous feat**u**res **le**arning (JOULE) model in our main submission, which is excluded from the main submission due to space limitation.

1 Optimization of JOULE

In our main submission, we propose a joint learning model to explore the shared and feature-specific structures for RGB-D activity recognition as an instance of heterogeneous multi-task learning. Specially, our JOULE model is given by

$$\begin{aligned}\min_{\mathbf{W}_0,\{\mathbf{W}_i\}\atop\{\mathbf{\Theta}_i\}} &\sum_{i=1,...,S}(\|(\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}_i\|_F^2\\ &-\gamma\|\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2)+\alpha\|\mathbf{W}_0\|_F^2+\beta\sum_{i=1,...,S}\|\mathbf{W}_i\|_F^2\\ s.t.\ &\mathbf{\Theta}_i^T\mathbf{\Theta}_i=\mathbf{I}, i=1,2,...S\end{aligned} \tag{1}$$

As mentioned in the main submission, we develop an efficient optimization algorithm for the proposed model (1) by iterating the following three steps.

STEP 1. Minimizing the objective function with respect to $\mathbf{W}_0$ for fixed coefficients $\mathbf{W}_i$ and $\mathbf{\Theta}_i$:

$$\min_{\mathbf{W}_0}\sum_{i=1}^{S}\|(\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}_i\|_F^2+\alpha\|\mathbf{W}_0\|_F^2 \tag{2}$$

STEP 2. Minimizing the function with respect to $\mathbf{W}_i$ for fixed the coefficients $\mathbf{W}_0$ and $\mathbf{\Theta}_i$:

$$\min_{\{\mathbf{W}_i\}}\sum_{i=1}^{S}\|(\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}_i\|_F^2+\beta\|\mathbf{W}_i\|_F^2 \tag{3}$$

- *J. Hu is with the School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, China. E-mail: hujianf@mail2.sysu.edu.cn*
- *W.S. Zheng and J. Lai are with the School of Information Science and Technology, Sun Yat-Sen University, Guangzhou, China. E-mail:wszheng@ieee.org and stsljh@mail.sysu.edu.cn*
- *J. Zhang is with the School of Science and Engineering (Computing), University of Dundee United Kingdom. E-mail:j.n.zhang@dundee.ac.uk*

We turn to optimizing the following S subproblems

$$\min_{\mathbf{W}_i}\|(\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}_i\|_F^2+\beta\|\mathbf{W}_i\|_F^2 \tag{4}$$

STEP 3. Finally, optimizing $\mathbf{\Theta}_i$ for fixed $\mathbf{W}_0, \mathbf{W}_i$:

$$\begin{aligned}\min_{\mathbf{\Theta}_i}&\sum_{i=1}^{S}(\|(\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}_i\|_F^2-\gamma\|\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2)\\ &s.t.\ \mathbf{\Theta}_i^T\mathbf{\Theta}_i=\mathbf{I}, i=1,2,...S\end{aligned} \tag{5}$$

Equivalently, we can solve the following S subproblems

$$\begin{aligned}&\min_{\mathbf{\Theta}_i}\|(\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}_i\|_F^2-\gamma\|\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2\\ &s.t.\ \mathbf{\Theta}_i^T\mathbf{\Theta}_i=\mathbf{I}\end{aligned} \tag{6}$$

Each subproblem can be solved by a gradient based updating scheme $\mathbf{\Theta}_i(t+1)=(\mathbf{I}+\frac{\tau}{2}\nabla)^{-1}(\mathbf{I}-\frac{\tau}{2}\nabla)\mathbf{\Theta}_i(t)$, where ∇ is defined by the gradient of the objective function $\mathbf{G}$ as $\nabla=\mathbf{G}\mathbf{\Theta}_i(t)^T-\mathbf{\Theta}_i(t)\mathbf{G}^T$. The gradient $\mathbf{G}$ is given by $\mathbf{G}=\mathbf{X}_i((\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i(t)^T\mathbf{X}_i-\mathbf{Y}_i)^T(\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T-2\gamma\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i(t)$.

2 Convergence of the Optimization

In the following, we will theoretically illustrate that our proposed three-step iterative optimization algorithm is guaranteed to converge to a locally optimal solution. To accomplish the proof, we firstly prove that all of the three steps would monotonously decrease the objective function value in Section 2.1, 2.2, and 2.3, and then show that the objective function is lower bounded and finally conclude its convergence in Section 2.4.

2.1 Optimization of STEP 1

Lemma 1. *The solution of problem (2) is given by* $\mathbf{W}_0^*=(\sum_i\lambda^2\mathbf{\Theta}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i+\alpha\mathbf{I})^{-1}\sum_i(\mathbf{\Theta}_i^T\mathbf{X}_i(\mathbf{Y}_i^T-(1-\lambda)\mathbf{X}_i^T\mathbf{\Theta}_i\mathbf{W}_i))$.

Proof. Let $J_1(\mathbf{W}_0)=\sum_{i=1}^{S}\|(\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}_i\|_F^2+\alpha\|\mathbf{W}_0\|_F^2$. Here, we aim to find an optimal $\mathbf{W}_0$ that minimizes $J_1(\mathbf{W}_0)$. The partial derivative of $J_1(\mathbf{W}_0)$ with respect to $\mathbf{W}_0$ is given by

$$\frac{\partial J_1}{\partial\mathbf{W}_0}=2\mathbf{\Phi}\mathbf{W}_0+2\lambda\sum_{i=1}^{S}\mathbf{\Theta}_i^T\mathbf{X}_i((1-\lambda)\mathbf{X}_i^T\mathbf{\Theta}_i\mathbf{W}_i-\mathbf{Y}_i^T)$$

$$\mathbf{\Phi}=\lambda^2\sum_{i=1}^{S}\mathbf{\Theta}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i+\alpha\mathbf{I}$$

By setting $\frac{\partial J_1}{\partial \mathbf{W}_0}$ to 0, we can obtain an analytic solution of the problem (2) as

$$\mathbf{W}_0^* = \lambda \mathbf{\Phi}^{-1} \sum_i \mathbf{\Theta}_i^T \mathbf{X}_i (\mathbf{Y}_i^T - (1-\lambda)\mathbf{X}_i^T \mathbf{\Theta}_i \mathbf{W}_i)$$

□

We also note that

$$\frac{\partial^2 J_1}{\partial \mathbf{W}_0^2} = 2(\lambda^2 \sum_{i=1}^{S} \mathbf{\Theta}_i^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{\Theta}_i + \alpha \mathbf{I}) \succeq 0$$

where $\succeq 0$ indicates *positive semi-definite*. Hence, The objective function in problem (2) is convex with respect to $\mathbf{W}_0$, and replacing $\mathbf{W}_0$ with $\mathbf{W}_0^*$ would obviously decrease the value of our objective function.

2.2 Optimization of STEP 2

Lemma 2. *The solution of the i^{th} subproblem in Step 2 (i.e. Formula 3) can be determined by* $\mathbf{W}_i^* = (1-\lambda)((1-\lambda)^2 \mathbf{\Theta}_i^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{\Theta}_i + \beta \mathbf{I})^{-1} \mathbf{\Theta}_i^T \mathbf{X}_i (\mathbf{Y}_i^T - \lambda \mathbf{X}_i^T \mathbf{\Theta}_i \mathbf{W}_0)$.

Proof. We denote the objective function of the i^{th} subproblem (4) as J_{2i}. Then the partial derivative of J_{2i} with respect to $\mathbf{W}_i$ is indicated by

$$\frac{\partial J_{2i}}{\partial \mathbf{W}_i} = 2\beta_i \mathbf{W}_i + 2(1-\lambda)\mathbf{\Theta}_i^T \mathbf{X}_i (\Delta Y_i)^T$$
$$\Delta Y_i = (\lambda \mathbf{W}_0 + (1-\lambda)\mathbf{W}_i)^T \mathbf{\Theta}_i^T \mathbf{X}_i - \mathbf{Y}_i$$

By setting the derivative equal to zero, we can obtain the optimal solution W_i^*. □

Similar to that in STEP 1, we can easily derive the second order derivative as

$$\frac{\partial^2 J_{2i}}{\partial \mathbf{W}_i^2} = 2((1-\lambda)^2 \mathbf{\Theta}_i^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{\Theta}_i + \beta \mathbf{I}) \succeq 0$$

Therefore, problem (3) is convex. This lemma indicates that the updating scheme in STEP 2 would decrease the value of our objective function in (1).

2.3 Optimization of STEP 3

Finally, we prove that the optimization of STEP 3 continues decreasing the objective function. Here we provide the proof based on some tricks in [1], which provides a feasible optimization technique with orthogonality constraints.

Theorem 1. *Let J_3 be the objective function of the i^{th} subproblem in (6). The problem (5) in STEP 3 would converge to a minimum using the proposed updating scheme.*

In order to prove the above theorem, we first need the following two lemmas.

Lemma 3. *1) The orthogonality constraints would be preserved in our updating scheme, i.e. $\mathbf{\Theta}_i(t+1)$ satisfying the orthogonality constraint, if $\mathbf{\Theta}_i(t)^T \mathbf{\Theta}_i(t) = \mathbf{I}$.*
2) Define $\mathbf{U}(\tau) = (\mathbf{I} + \frac{\tau}{2}\nabla)^{-1}(\mathbf{I} - \frac{\tau}{2}\nabla)\mathbf{\Theta}_i$, then $\frac{\partial J_3(\mathbf{U}(\tau))}{\partial \tau}|_{\tau=0} = -\frac{1}{2} \parallel \nabla \parallel_F^2$.

Proof. Part 1): Provided that ∇ is a skew-symmetric matrix, i.e. $\nabla^T = -\nabla$, we then have

$$\begin{aligned}
&\mathbf{\Theta}_i(t+1)^T \mathbf{\Theta}_i(t+1) \\
=&\mathbf{\Theta}_i(t)^T (\mathbf{I} + \frac{\tau}{2}\nabla)(\mathbf{I} - \frac{\tau}{2}\nabla)^{-1}(\mathbf{I} + \frac{\tau}{2}\nabla)^{-1}(\mathbf{I} - \frac{\tau}{2}\nabla)\mathbf{\Theta}_i(t) \\
=&\mathbf{\Theta}_i(t)^T \underline{(\mathbf{I} - \frac{\tau}{2}\nabla)^{-1}(\mathbf{I} - \frac{\tau}{2}\nabla)(\mathbf{I} + \frac{\tau}{2}\nabla)(\mathbf{I} - \frac{\tau}{2}\nabla)^{-1}} \\
&\underline{(\mathbf{I} + \frac{\tau}{2}\nabla)^{-1}(\mathbf{I} - \frac{\tau}{2}\nabla)}\mathbf{\Theta}_i(t) \\
=&\underline{\mathbf{\Theta}_i(t)^T \mathbf{\Theta}_i(t)} \\
=&\mathbf{I}
\end{aligned}$$

For the derivation of the equations marked as underlined in the above, we use the fact that

$$(\mathbf{I} - \frac{\tau}{2}\nabla)(\mathbf{I} + \frac{\tau}{2}\nabla) = (\mathbf{I} + \frac{\tau}{2}\nabla)(\mathbf{I} - \frac{\tau}{2}\nabla)$$

Part 2): We define $\mathbf{U}(\tau) = (\mathbf{I} + \frac{\tau}{2}\nabla)^{-1}(\mathbf{I} - \frac{\tau}{2}\nabla)\mathbf{\Theta}_i$. Indeed, this is the newly updated point with a step size τ. Then we have

$$(\mathbf{I} + \frac{\tau}{2}\nabla)\mathbf{U}(\tau) = (\mathbf{I} - \frac{\tau}{2}\nabla)\mathbf{\Theta}_i$$

By taking the derivative of the above equation with respect to τ, we can obtain

$$(\mathbf{I} + \frac{\tau}{2}\nabla)\frac{\partial \mathbf{U}(\tau)}{\partial \tau} + \frac{\nabla}{2}\mathbf{U}(\tau) = -\frac{1}{2}\nabla \mathbf{\Theta}_i$$

By solving the above equation and setting $\tau = 0$, we can obtain

$$\frac{\partial \mathbf{U}}{\partial \tau}|_{\tau=0} = -(\mathbf{I} + \frac{\tau\nabla}{2})^{-1}\frac{\nabla}{2}(\mathbf{\Theta}_i + \mathbf{U}(\tau))|_{\tau=0} = -\nabla\mathbf{\Theta}_i \quad (7)$$

Then

$$\begin{aligned}
\frac{\partial J_3(\mathbf{U}(\tau))}{\partial \tau}|_{\tau=0} =& tr(\frac{\partial J_3(\mathbf{U})}{\partial \mathbf{U}}^T \frac{\partial \mathbf{U}}{\partial \tau})|_{\tau=0} \\
=& -tarce(\mathbf{G}^T \nabla \mathbf{\Theta}_i) \\
=& -tr(\mathbf{G}^T(\mathbf{G}\mathbf{\Theta}_i^T - \mathbf{\Theta}_i\mathbf{G}^T)\mathbf{\Theta}_i) \\
=& -tr(\mathbf{G}\mathbf{G}^T - \mathbf{G}^T\mathbf{\Theta}_i\mathbf{G}^T\mathbf{\Theta}_i^T) \\
=& -\frac{1}{2}tr(\mathbf{G}\mathbf{\Theta}_i^T\mathbf{\Theta}_i\mathbf{G}^T + \mathbf{\Theta}_i\mathbf{G}^T\mathbf{G}\mathbf{\Theta}_i^T \\
& - \mathbf{G}\mathbf{\Theta}_i^T\mathbf{G}\mathbf{\Theta}_i^T - \mathbf{\Theta}_i\mathbf{G}^T\mathbf{\Theta}_i\mathbf{G}^T) \\
=& -\frac{1}{2}tr(\mathbf{G}\mathbf{\Theta}_i^T - \mathbf{\Theta}_i\mathbf{G}^T)(\mathbf{G}\mathbf{\Theta}_i^T - \mathbf{\Theta}_i\mathbf{G}^T)^T \\
=& -\frac{1}{2}tr(\nabla\nabla^T) \\
=& -\frac{1}{2} \parallel \nabla \parallel_F^2
\end{aligned}$$

In the aforementioned derivations, we use the fact that $tr(\mathbf{G}\mathbf{G}^T) = tr(\mathbf{\Theta}_i\mathbf{G}^T\mathbf{G}\mathbf{\Theta}_i^T)$ and $tr(\mathbf{G}\mathbf{\Theta}_i^T\mathbf{G}\mathbf{\Theta}_i^T) = tr(\mathbf{\Theta}_i\mathbf{G}^T\mathbf{\Theta}_i\mathbf{G}^T)$ as the following equation, $tr(\mathbf{A}\mathbf{A}') = tr(\mathbf{A}'\mathbf{A})$, holds for any matrix $\mathbf{A}$. □

Lemma 4. *With proper step size τ^*, updating $\mathbf{\Theta}_i(t)$ with $\mathbf{\Theta}_i(t+1) = (\mathbf{I} + \frac{\tau^*}{2}\nabla)^{-1}(\mathbf{I} - \frac{\tau^*}{2}\nabla)\mathbf{\Theta}_i(t)$ would decrease the value of the objective function in the i^{th} subproblem (6).*

Proof. According to Lemma 3, we can select a proper τ^* such that $\frac{\partial J_3(\mathbf{\Theta}_i(t+1))}{\partial \tau}|_{\tau^*} \leq 0$. Then the selected τ^* would decrease the objective function value. □

Proof of Theorem 1. Now, it is straightforward to prove the theorem 1. For each step in our algorithm, according to Lemma 4 and part 1 of Lemma 3, the proposed updating scheme would decrease the objective function and simultaneously satisfy orthogonality constraint in each iteration. Since the objective function has a lower bound, the updating scheme of STEP 3 will converge to a minimum. □

2.4 Convergence of the algorithm

As stated in section 2.1, 2.2, and 2.3, all the three steps would decrease the objective function in our JOULE model (1). Note that

$$-\|\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2 = -\|\mathbf{X}_i\|_F^2 + \|\mathbf{X}_i - \mathbf{\Theta}_i\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2 \geq -\|\mathbf{X}_i\|_F^2$$

Hence, the objective function in (1) is lower bounded when $\alpha, \beta, \gamma \geq 0$

Therefore, the proposed optimization algorithm is guaranteed to converge to a minimum.

REFERENCES

[1] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.