Supplementary Material: Multi-Task Personalized Learning with Sparse Network Lasso

Iterative algorithm for solving Eq. (8) in the main paper

To make the prediction on an unseen testing sample $\hat{\mathbf{x}}_t$, we can solve the following problem to learn its personalized model $\hat{\boldsymbol{\theta}}_t$:

$$\hat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{k} s_i^t \|\boldsymbol{\theta} - \boldsymbol{\theta}_{t,i}\|_2, \tag{1}$$

where k is the number of neighbors of $\hat{\mathbf{x}}_t$ in the training data, and s_i^t measures the similarity between $\hat{\mathbf{x}}_t$ and its neighbor $\mathbf{x}_{t,i}$.

The above problem lies in the general framework of Weber problem, which can be efficiently solved by an iterative algorithm [Kuhn, 1992]. At each step of the iterative algorithm, the model is moved closer to the optimal solution by setting $\boldsymbol{\theta}^{(j+1)}$ to be the solution of a weighted least squares problem:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{k} \frac{s_i^t}{\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}_{t,i}\|_2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{t,i}\|_2^2. \tag{2}$$

As the unique optimal solution to the above weighted least square problem, each successive is calculated by:

$$\boldsymbol{\theta}^{(j+1)} = \left(\sum_{i=1}^{k} \frac{s_i^t \boldsymbol{\theta}_{t,i}}{\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}_{t,i}\|_2} \right) \Bigg/ \left(\sum_{i=1}^{k} \frac{s_i^t}{\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}_{t,i}\|_2} \right). \tag{3}$$

Algorithm 1 Iterative algorithm for solving Eq. (8) in the main paper

---

Input: $\mathbf{s}^t \in \mathbb{R}^k, \boldsymbol{\Theta}_t = [\boldsymbol{\theta}_{t,1}, \boldsymbol{\theta}_{t,2}, ..., \boldsymbol{\theta}_{t,k}] \in \mathbb{R}^{d \times k}$
Output: $\hat{\boldsymbol{\theta}}_t \in \mathbb{R}^d$.
1: Set $j = 0$, initialize $\mathbf{f}_g \in \mathbb{R}^k$.
2: **repeat**
3: Compute $\boldsymbol{\theta}^{(j+1)} = \frac{1}{\mathbf{1}_k^T \mathbf{f}_g^{(j)}} \boldsymbol{\Theta}_t \mathbf{f}_g^{(j)}$.
4: Update $\mathbf{f}_g^{(j+1)}$, where $\left[\mathbf{f}_g^{(j+1)}\right]_i = \frac{[\mathbf{s}^t]_i}{\|\boldsymbol{\theta}^{(j+1)} - \boldsymbol{\theta}_{t,i}\|_2}, i = 1, 2, ..., k$.
5: $j = j + 1$.
6: **until** *Convergence*

Proposition and theorem used for deriving Eq. (11) in the main paper

Theorem 1 (Black-Rangarajan Duality [Black and Rangarajan, 1996]). *Given a robust cost function $\rho(x)$, $x \geq 0$, define $\phi(w) = \rho(\sqrt{w})$. If $\phi(w)$ satisfies $\lim_{w\to\infty} \phi'(w) = 0$ ($\lim_{w\to 0} \phi'(w) = 1, optional$), and $\phi''(w) < 0$, then $\rho(x)$ has a equivalent formulation w.r.t x, namely, the outlier process formulation:*

$$E(x, l) = lx^2 + \Psi(l), \tag{4}$$

where $l = \phi'(w) > 0$ is a slack variable, and the function $\Psi(l)$ is the penalty on l. The expression of $\Psi(l)$ depends on the choice of robust cost function $\rho(x)$, taking the form as:

$$\Psi(l) = \phi(w) - lw = \phi\big((\phi')^{-1}(l)\big) - l(\phi')^{-1}(l). \tag{5}$$

Proposition 1. *The optimization problem in the main paper:*

$$\min_{\mathbf{G}_t} \|\mathbf{y}_t - \mathcal{X}_t vec\left(\left(\mathbf{A} + \mathbf{B}_t\right) \mathbf{G}_t \mathbf{M}_t\right)\|_2^2 + \lambda_2 \sum_{i,j=1}^{n_t} s_{ij}^t \|\mathbf{g}_{t,i} - \mathbf{g}_{t,j}\|_2 + \lambda_3 \|\mathbf{G}_t\|_1, \tag{6}$$

has a equivalent formulation w.r.t $\mathbf{G}_t$, that is:

$$\min_{\mathbf{G}_t, \mathbf{L}} \|\mathbf{y}_t - \mathcal{X}_t vec\left(\left(\mathbf{A} + \mathbf{B}_t\right) \mathbf{G}_t \mathbf{M}_t\right)\|_2^2 + \lambda_2 \sum_{i,j=1}^{n_t} s_{ij}^t \big(l_{i,j}^t \|\mathbf{g}_{t,i} - \mathbf{g}_{t,j}\|_2^2 + \frac{1}{4}(l_{i,j}^t)^{-1}\big) + \lambda_3 \|\mathbf{G}_t\|_1, \tag{7}$$

where $l_{i,j}^t \geq 0$ is the auxiliary variable.

Proof. Since $\rho(x) = |x|$ is a robust cost function, and we have:

$$\phi(w) = \sqrt{w}, \tag{8}$$

$$\phi'(w) = \frac{1}{2\sqrt{w}}, \tag{9}$$

$$\phi''(w) = -\frac{1}{4w^{\frac{3}{2}}}, \tag{10}$$

which satisfy $\lim_{w\to\infty} \phi'(w) = 0$ and $\phi''(w) < 0$. According to the Black-Rangarajan Duality in Theorem 1, $\rho(x) = |x|$ has a equivalent formulation w.r.t x:

$$E(x, l) = lx^2 + \Psi(l), \tag{11}$$

where $l = \phi'(w)$ and $\Psi(l) = \phi(w) - lw$. Solving w gives rise to:

$$w = (\phi')^{-1}(l) = \frac{1}{4l^2}. \tag{12}$$

By substituting (12) into (11), we have:

$$E(x, l) = lx^2 + \frac{1}{4l}. \tag{13}$$

Thus, (7) can be obtained by setting $x = \|\mathbf{g}_{t,i} - \mathbf{g}_{t,j}\|_2$ in (13).

□

Proposition and theorem used for deriving Eq. (14) in the main paper

Proposition 2. *The gradient of*

$$\min_{\mathbf{G}_t} \|\mathbf{y}_t - \mathcal{X}_t vec\left((\mathbf{A} + \mathbf{B}_t)\,\mathbf{G}_t\mathbf{M}_t\right)\|_2^2 + 2\lambda_2 tr\Big((\mathbf{G}_t\mathbf{N}_t)\,(\mathbf{D}_t - \mathbf{W}_t)\,(\mathbf{G}_t\mathbf{N}_t)^T \Big), \tag{14}$$

w.r.t. $vec(\mathbf{G}_t)$ is

$$\nabla f(vec(\mathbf{G}_t)) = 2\mathbf{P}_t^T\mathbf{P}_t\, vec(\mathbf{G}_t) - 2\mathbf{P}_t^T\mathbf{y}_t + 4\lambda_2(\mathbf{N}_t^T \otimes \mathbf{I}_K)^T vec(\mathbf{G}_t\mathbf{N}_t(\mathbf{D}_t - \mathbf{W}_t)). \tag{15}$$

where $\mathbf{P}_t = \mathcal{X}_t(\mathbf{M}_t^T \otimes (\mathbf{A} + \mathbf{B}_t))$.

Proof. Based on the chain rule of derivative, because

$$\frac{\partial\|\mathbf{y}_t - \mathcal{X}_t\text{vec}\left((\mathbf{A} + \mathbf{B}_t)\,\mathbf{G}_t\mathbf{M}_t\right)\|_2^2}{\partial\text{vec}\left((\mathbf{A} + \mathbf{B}_t)\,\mathbf{G}_t\mathbf{M}_t\right)} = -2\mathcal{X}_t^T(\mathbf{y}_t - \mathcal{X}_t\text{vec}\left((\mathbf{A} + \mathbf{B}_t)\,\mathbf{G}_t\mathbf{M}_t\right)) \tag{16}$$

and

$$\frac{\partial\text{vec}\left((\mathbf{A} + \mathbf{B}_t)\,\mathbf{G}_t\mathbf{M}_t\right)}{\text{vec}(\mathbf{G}_t)} = \mathbf{M}^T \otimes (\mathbf{A} + \mathbf{B}_t)\,, \tag{17}$$

we have,

$$\frac{\partial\|\mathbf{y}_t - \mathcal{X}_t\text{vec}\left((\mathbf{A} + \mathbf{B}_t)\,\mathbf{G}_t\mathbf{M}_t\right)\|_2^2}{\text{vec}(\mathbf{G}_t)} = 2\mathbf{P}_t^T\mathbf{P}_t\ \text{vec}(\mathbf{G}_t) - 2\mathbf{P}_t^T\mathbf{y}_t, \tag{18}$$

where $\mathbf{P}_t = \mathcal{X}_t(\mathbf{M}_t^T \otimes (\mathbf{A} + \mathbf{B}_t))$.
Similarly, according to

$$\frac{\partial 2\lambda_2 tr\Big((\mathbf{G}_t\mathbf{N}_t)\,(\mathbf{D}_t - \mathbf{W}_t)\,(\mathbf{G}_t\mathbf{N}_t)^T \Big)}{\partial\text{vec}(\mathbf{G}_t\mathbf{N_t})} = 4\lambda_2(\mathbf{G}_t\mathbf{N_t})(\mathbf{D}_t - \mathbf{W}_t), \tag{19}$$

and

$$\frac{\partial\ \text{vec}(\mathbf{G}_t\mathbf{N_t})}{\partial\ \text{vec}(\mathbf{G}_t)} = \mathbf{N}_t^T \otimes \mathbf{I}_K, \tag{20}$$

we have,

$$\frac{\partial 2\lambda_2 tr\Big((\mathbf{G}_t\mathbf{N}_t)\,(\mathbf{D}_t - \mathbf{W}_t)\,(\mathbf{G}_t\mathbf{N}_t)^T \Big)}{\partial\text{vec}(\mathbf{G}_t)} = 4\lambda_2(\mathbf{N}_t^T \otimes \mathbf{I}_K)^T\text{vec}(\mathbf{G}_t\mathbf{N}_t(\mathbf{D}_t - \mathbf{W}_t)). \tag{21}$$

Therefore, we can reach the conclusion,

$$\nabla f(\text{vec}(\mathbf{G}_t)) = 2\mathbf{P}_t^T\mathbf{P}_t\ \text{vec}(\mathbf{G}_t) - 2\mathbf{P}_t^T\mathbf{y}_t + 4\lambda_2(\mathbf{N}_t^T \otimes \mathbf{I}_K)^T\text{vec}(\mathbf{G}_t\mathbf{N}_t(\mathbf{D}_t - \mathbf{W}_t)). \tag{22}$$

□

Implementation for Section: Optimization algorithm

In Algorithm 2, we provide the optimization algorithm of MTPL discussed in Sec. 4 of the main paper. Note that, we apply Accelerated Proximal Gradient (APG) method [Beck, 2017] in Algorithm 2 to accelerate the optimization algorithm.

Algorithm 2 MTPL: Optimization algorithm

Input: $\{\mathcal{X}_t\}_{t=1}^m$, $\{\mathbf{y}_t\}_{t=1}^m$, $\{\mathbf{S}_t\}_{t=1}^m$, λ_1, λ_2, λ_3.
Output: $\mathbf{\Theta}_t = (\mathbf{A} + \mathbf{B}_t)\mathbf{G}_t\mathbf{M}_t$.

```
Initialize A, {B_t}_{t=1}^m, {G_t}_{t=1}^m.
repeat
    repeat
        Update l^t_{i,j} = (2||g_{t,i} - g_{t,j}||_2)^{-1}, ∀t, i, j.
        Update G_t via APG based on Eq. (15) in the main paper.
    until Convergence
    Update {B_t}_{t=1}^m based on Eq. (17) in the main paper.
    Update A based on Eq. (19) in the main paper.
until Convergence
```

Real-world Datasets

To evaluate the proposed model, we conduct experiments on the following six real-world datasets:

- **School**: The School dataset contains examination records of 15,362 students from 139 schools. Each school is considered as a task and the aim is to predict the exam score of each student.
- **SARCOS**: The SARCOS dataset relates to an inverse dynamics problem, mapping from a 21-dimensional input space to 7 joint torques. We randomly select 1000 data points in SARCOS.
- **Sales**: The Sales dataset contains purchased quantities of 811 products over 52 weeks. We pre-process the dataset following the setup in [He *et al.*, 2019]. We treat each product's sales prediction as a task.
- **Parkinsons**: The Parkinsons dataset is to predict the disease symptom scores of Parkinson for 42 patients at different times by using 16 bio-medical features. The prediction of each patient is considered as a task.
- **Computer**: The Computer dataset is obtained from a survey of 190 students who rated their likelihood of purchasing 20 different computers. Here, students correspond to tasks. The aim is to predict students' purchase intention on different computers.
- **Isolet**: The Isolet dataset contains 150 subjects who spoke the name of each letter of the alphabet twice. The speakers are grouped into 5 subsets of 30 similar speakers, resulting in 5 tasks. We reduce the feature dimension to 100 by Principal Component Analysis (PCA) [Wold *et al.*, 1987].

References

[Beck, 2017] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[Black and Rangarajan, 1996] Michael J Black and Anand Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International journal of computer vision*, 19(1):57–91, 1996.

[He *et al.*, 2019] Xiao He, Francesco Alesiani, and Ammar Shaker. Efficient and scalable multi-task regression on massive number of tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3763–3770, 2019.

[Kuhn, 1992] Harold W. Kuhn. *An Efficient Algorithm for the Numerical Solution of the Generalized Weber Problem in Spatial Economics*, pages 223–240. Palgrave Macmillan UK, London, 1992.

[Wold *et al.*, 1987] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.