Cross-session Emotion Recognition by Joint Label-common and Label-specific EEG Features Exploration - Supplementary Material

Yong Peng, Honggang Liu, Junhua Li, Jun Huang, Bao-Liang Lu, and Wanzeng Kong*

APPENDIX: OPTIMIZATION TO JCSFE MODEL OBJECTIVE FUNCTION

Due to the page limit, the optimization to the JCSFE model objective function is provided in this supplementary material.

■ **The $\mathbf{Y}_u$ step**. Here we provide the detailed derivation to objective function defined on variable $\mathbf{y}^i|_{i=l+1}^n$. By denoting $\mathbf{m}^i \triangleq \mathbf{x}_i^T\mathbf{W}$, $i = l+1, l+2, \cdots, l+u$, we have

$$\min_{\mathbf{y}^i\geq\mathbf{0},\ \mathbf{y}^i\mathbf{1}_c=1} \|\mathbf{y}^i - \mathbf{m}^i\|_2^2. \tag{1}$$

The corresponding Lagrangian function is

$$\mathcal{L}(\mathbf{y}^i, \eta, \boldsymbol{\delta}) = \|\mathbf{y}^i - \mathbf{m}^i\|_2^2 - \eta(\mathbf{y}^i\mathbf{1}_c - 1) - \mathbf{y}^i\boldsymbol{\delta}^T, \tag{2}$$

where η and $\boldsymbol{\delta} \in \mathbb{R}^{1\times c}$ are two Lagrange multipliers respectively in scalar and vector forms. Below we show how both the Lagrange multipliers are determined. Suppose that the optimal solution to problem (1) is $\mathbf{y}^{i*}$, and the corresponding Lagrange multipliers are η^* and $\boldsymbol{\delta}^*$. Then, according to the Karush-Kuhn-Tucker (KKT) condition, we have the following equations and inequalities

$$\begin{cases} \forall j, \quad y_{ij}^* - m_{ij} - \eta^* - \delta_j^* = 0, & (3)\\ \forall j, \quad y_{ij}^* \geq 0, & (4)\\ \forall j, \quad \delta_j^* \geq 0, & (5)\\ \forall j, \quad y_{ij}^*\beta_j^* = 0, & (6) \end{cases}$$

where y_{ij}^* is the j-th element of vector $\mathbf{y}^{i*}$. The vector form of (3) is

$$\mathbf{y}^{i*} - \mathbf{m}^i - \eta^*\mathbf{1}_c^T - \boldsymbol{\delta}^* = \mathbf{0}. \tag{7}$$

Since we have the constraint $\mathbf{y}^i\mathbf{1}_c = 1$, the above equation can be reformulated into

$$\eta^* = \frac{1 - \mathbf{m}^i\mathbf{1}_c - \boldsymbol{\delta}^*\mathbf{1}_c}{c}. \tag{8}$$

By replacing η^* in (7) with (8), we have

$$\mathbf{y}^{i*} = \mathbf{m}^i - \frac{\mathbf{m}_i\mathbf{1}_c}{c}\mathbf{1}_c^T + \frac{1}{c}\mathbf{1}_c^T - \frac{\boldsymbol{\delta}^*\mathbf{1}_c}{c}\mathbf{1}_c^T + \boldsymbol{\delta}^*. \tag{9}$$

By denoting $\bar{\delta}^* = \frac{\boldsymbol{\delta}^*\mathbf{1}_c}{c}$ and $\mathbf{q} = \mathbf{m}^i - \frac{\mathbf{m}_i\mathbf{1}_c}{c}\mathbf{1}_c^T + \frac{1}{c}\mathbf{1}_c^T$, we can rewrite the above equation as

$$\mathbf{y}^{i*} = \mathbf{q} + \boldsymbol{\delta}^* - \bar{\delta}^*\mathbf{1}_c^T. \tag{10}$$

Accordingly, for each $j = 1, 2, \cdots, c$, we have

$$y_{ij}^* = q_j + \delta_j^* - \bar{\delta}^*. \tag{11}$$

Considering equations (4), (5), (6), and (11) together, we know that $q_j + \delta_j^* - \bar{\delta}^* = (q_j - \bar{\delta}^*)_+$, where $(f(\cdot))_+ = \max(f(\cdot), 0)$. Therefore, we have

$$y_{ij}^* = (q_j - \bar{\delta}^*)_+. \tag{12}$$

Till now, if $\bar{\delta}^*$ could be determined, $\mathbf{y}_i^*$ will be accordingly determined by (12). From (11), we have $\delta_j^* = y_{ij}^* + \bar{\delta}^* - q_j$ such that $\delta_j^* = (\bar{\delta}^* - q_j)_+$. Therefore, $\bar{\delta}^*$ can be calculated as

$$\bar{\delta}^* = \frac{1}{c}\sum_{j=1}^{c}(\bar{\delta}^* - q_j)_+. \tag{13}$$

According to the constraint $\mathbf{y}^i\mathbf{1}_c = 1$ and (12), we define the following function

$$f(\bar{\delta}) = \sum_{j=1}^{c}(q_j - \bar{\beta})_+ - 1, \tag{14}$$

and the optimal $\bar{\delta}^*$ should satisfy $f(\bar{\delta}^*) = 0$. When (14) equals to zero, the optimal $\bar{\delta}^*$ can be obtained via Newton method, namely,

$$\bar{\delta}^{(k+1)} = \bar{\delta}^{(k)} - \frac{f(\bar{\beta}^{(k)})}{f'(\bar{\beta}^{(k)})}. \tag{15}$$

It is obvious that $f(\bar{\delta})$ is a piecewise linear and monotonically increasing function. When $q_j \geq \bar{\delta}$, we have $f(\bar{\delta}) = \sum_{j=1}^{c} q_j - \bar{\delta} - 1$ and $f'(\bar{\delta}) = -1$. When $q_j \leq \bar{\delta}$, we have $f(\bar{\delta}) = -1$ and its derivative $f'(\bar{\delta}) = 0$. As a result, we obtain $f'(\bar{\delta})$ by counting the number of positive values in $(q_j - \bar{\delta})|_{j=1}^{c}$.

■ **The W step.** First, the convex optimization problem in the general APG method is defined as

$$\min_{\mathbf{W}\in\mathcal{H}} F(\mathbf{W}) = f(\mathbf{W}) + g(\mathbf{W}), \tag{16}$$

where $\mathcal{H}$ indicates the real Hilbert space. $f(\mathbf{W})$ is convex and smooth, and $g(\mathbf{W})$ is convex but typically non-smooth. $f(\mathbf{W})$ further satisfies the Lipschitz continuous condition; that is

$$\|\nabla f(\mathbf{W}_1) - \nabla f(\mathbf{W}_2)\|_2 \leq L_f\|\Delta\mathbf{W}\|_2, \tag{17}$$

where L_f is termed as the Lipschitz constant and $\Delta\mathbf{W} = \mathbf{W}_1 - \mathbf{W}_2$. Below we propose to minimize the separable quadratic approximation sequence of $f(\mathbf{W})$ by the proximal gradient algorithm rather than minimizing it directly, which is expressed as

$$\begin{aligned} Q(\mathbf{W}, \mathbf{W}^{(t)}) =& f(\mathbf{W}^{(t)}) + \langle\nabla f(\mathbf{W}^{(t)}), \mathbf{W} - \mathbf{W}^{(t)}\rangle \\ &+ \frac{L_f}{2}\|\mathbf{W} - \mathbf{W}^{(t)}\|_2^2 + g(\mathbf{W}). \end{aligned} \tag{18}$$

By denoting $\mathbf{G}^{(t)} = \mathbf{W}^{(t)} - \frac{1}{L_f}\nabla f(\mathbf{W}^{(t)})$, we rewrite the above expression as

$$Q(\mathbf{W}, \mathbf{W}^{(t)}) = g(\mathbf{W}) + \frac{L_f}{2}\|\mathbf{W} - \mathbf{G}^{(t)}\|_2^2. \tag{19}$$

According to the JCSFE objective function and equation (16), we have

$$f(\mathbf{W}) = \|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|_2^2 + \gamma \mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \beta \mathrm{Tr}(\mathbf{W}^T\mathbf{A}\mathbf{W}), \tag{20}$$

and

$$g(\mathbf{W}) = \alpha\|\mathbf{W}\|_1. \tag{21}$$

By combining equations (19), (20) and (21) together, we obtain the objective function in terms of variable $\mathbf{W}$ as

$$\mathbf{W} = \arg\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{W} - \mathbf{G}^{(t)}\|_2^2 + \frac{\alpha}{L_f}\|\mathbf{W}\|_1. \tag{22}$$

According to the existing studies [1], [2], we set $\mathbf{W}^{(t)} = \mathbf{W}^{(t)} + \frac{b^{(t-1)}-1}{b^{(t)}}(\mathbf{W}^{(t)} - \mathbf{W}^{(t-1)})$ and then the convergence speed of the proximal gradient method can be accelerated to $\mathcal{O}(t^{-2})$, where sequence $b^{(t)}$ satisfies $(b^t)^2 - b^t \leqslant (b^{(t-1)})^2$ and $\mathbf{W}^{(t)}$ is the updated result at t-th iteration. It is obvious that (22) is an ℓ_1-norm regularized problem which can be solved by the following soft-shrinkage operator

$$\mathcal{S}_\varepsilon[x] = \begin{cases} x - \varepsilon, & \text{if } x > \varepsilon, \\ x + \varepsilon, & \text{if } x < -\varepsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

The ε above is usually a small positive value. This operator can be extended to vectors and matrices by applying it element-wisely. Then, by setting $\varepsilon = \frac{\alpha}{L_f}$, we can obtain $\mathbf{W}^{(t+1)}$ by solving

$$\mathcal{S}_\varepsilon[\mathbf{G}^{(t)}] = \arg\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{W} - \mathbf{G}^{(t)}\|_2^2 + \varepsilon\|\mathbf{W}\|_1. \tag{24}$$

For $\nabla f(\mathbf{W})$, it can be obtained by taking the derivative of equation (20) with respect to $\mathbf{W}$. That is

$$\nabla f(\mathbf{W}) = \mathbf{X}\mathbf{X}^T\mathbf{W} - \mathbf{X}\mathbf{Y} + \gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W} + \beta\mathbf{A}\mathbf{W}. \tag{25}$$

When $\mathbf{W}_1$ and $\mathbf{W}_2$ are given, we have

$$\begin{aligned}
&\|\nabla f(\mathbf{W}_1) - \nabla f(\mathbf{W}_2)\|_2^2 \\
=&\|\mathbf{X}\mathbf{X}^T\Delta\mathbf{W} + \gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\Delta\mathbf{W} + \beta\mathbf{A}\Delta\mathbf{W}\|_2^2 \\
\leq&(\|\mathbf{X}\mathbf{X}^T\Delta\mathbf{W}\| + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\Delta\mathbf{W}\| + \|\beta\mathbf{A}\Delta\mathbf{W}\|)^2 \\
=&\|\mathbf{X}\mathbf{X}^T\Delta\mathbf{W}\|^2 + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\Delta\mathbf{W}\|^2 + \|\beta\mathbf{A}\Delta\mathbf{W}\|^2 \\
&+ 2\|\mathbf{X}\mathbf{X}^T\Delta\mathbf{W}\| \cdot \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\Delta\mathbf{W}\| \\
&+ 2\|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\Delta\mathbf{W}\| \cdot \|\beta\mathbf{A}\Delta\mathbf{W}\| \\
&+ 2\|\mathbf{X}\mathbf{X}^T\Delta\mathbf{W}\| \cdot \|\beta\mathbf{A}\Delta\mathbf{W}\| \\
\leq&3(\|\mathbf{X}\mathbf{X}^T\Delta\mathbf{W}\|^2 + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\Delta\mathbf{W}\|^2 + \|\beta\mathbf{A}\Delta\mathbf{W}\|^2) \\
\leq&3(\|\mathbf{X}\mathbf{X}^T\|^2\|\Delta\mathbf{W}\|^2 + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\|^2\|\Delta\mathbf{W}\|^2 + \|\beta\mathbf{A}\|^2\|\Delta\mathbf{W}\|^2) \\
=&3(\|\mathbf{X}\mathbf{X}^T\|^2 + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\|^2 + \|\beta\mathbf{A}\|^2)\|\Delta\mathbf{W}\|^2
\end{aligned} \tag{26}$$

By comparing inequalities (17) and (26), the Lipschitz constant L_f can be set as

$$L_f = \sqrt{3(\|\mathbf{X}\mathbf{X}^T\|_2^2 + \|\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\|_2^2 + \|\beta\mathbf{A}\|_2^2)}. \tag{27}$$

When $\mathbf{W}$ is given, then $\mathbf{A}$ is fixed and further L_f is a constant value.

REFERENCES

[1] H. Li and Z. Lin, "Accelerated proximal gradient methods for nonconvex programming," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 379–387.

[2] J. Li, P. Li, X. Hu, and K. Yu, "Learning common and label-specific features for multi-label classification with correlation information," *Pattern Recogn.*, vol. 121, no. 108259, pp. 1–15, 2022.