1. SUPPLEMENTARY MATERIAL

1.1. Group Sparsemax

The original optimization problem is:

$$\min_{p\in\Delta^d} \frac{1}{2}(y-p)^2 + \lambda\sum_{i=1}^{n}||p_i||_2, \tag{1}$$

where $p = \{p_1, p_2, p_3, ..., p_n\}$ is the generated probability distribution, $p_i = \{\theta_i^1, \theta_i^2, \theta_i^3, ..., \theta_i^m\}$ is $i-th$ group, θ_i^j is $j-th$ probability unit of the $i-th$ probability group, $\Delta^d = \{P \in R^d | 1^T P = 1, P \geq 0\}$ is d-dimension probability simplex.

We convert (1) into the format of Group Lasso, which is as follows:

$$\begin{aligned}
\min \quad & \frac{1}{2}(y - \sum_{i=1}^{n} H_i p_i)^2 + \lambda\sum_{i=1}^{n}||p_i||_2 \\
=& \frac{1}{2}(y - \sum_{i=1}^{n} H_i p_i)^T(y - \sum_{i=1}^{n} H_i p_i) + \lambda\sum_{i=1}^{n}||p_i||_2 \\
=& \frac{1}{2}(y^T - \sum_{i=1}^{n} p_i^T H_i^T)(y - \sum_{i=1}^{n} H_i p_i) + \lambda\sum_{i=1}^{n}||p_i||_2 \\
=& \frac{1}{2}(y^T y - y^T\sum_{i=1}^{n} H_i p_i - \sum_{i=1}^{n} p_i^T H_i^T y + \\
& \sum_{i=1}^{n}\sum_{j=1}^{n} p_i^T H_i^T H_j p_j) + \lambda\sum_{i=1}^{n}||p_i||_2 \\
=& \frac{1}{2}(y^T y - 2y^T\sum_{i=1}^{n} p_i + \sum_{i=1}^{n} p_i^T p_i) + \lambda\sum_{i=1}^{n}||p_i||_2 \\
s.t \quad & 1^T p = 1, \\
& p \geq 0,
\end{aligned} \tag{2}$$

where $H_i = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$.

The Lagrangian of the optimization problem in (2) is:

$$\begin{aligned}
\mathcal{L}(y, \mu, \tau) =& \frac{1}{2}(y^T y - 2y^T\sum_{i=1}^{n} p_i + \sum_{i=1}^{n} p_i^T p_i) + \\
& \lambda\sum_{i=1}^{n}||p_i||_2 - \mu p + \tau(1^T p - 1).
\end{aligned} \tag{3}$$

The optimal $\{p^*, \mu^*, \tau^*\}$ must satisfy the following Karush-Kuhn-Tucker conditions:

1) When $p_i > 0$,

$$p^* \geq 0, \mu^* \geq 0, \mu^* p = 0, 1^T p^* = 1 \tag{4}$$

$$\begin{aligned}
\frac{\partial\mathcal{L}}{p_i^*} &= -y_i^T + p_i^* + \lambda\frac{p_i^*}{||p_i^*||_2} + \tau^* 1^T - \mu_i^* = 0 \\
&\Rightarrow (1 + \frac{\lambda}{||p_i^*||_2})p_i^* = y_i - \tau^* + \mu_i^* \\
&\Rightarrow p_i^* = (1 + \frac{\lambda}{||p_i^*||_2})^{-1}(y_i - \tau^* + \mu_i^*)_+
\end{aligned} \tag{5}$$

Because $||p_i^*||_2$ exists in the above formula, so we can apply the l_2 norm on both sides at the same time, and we can get $||p_i^*||_2$, let $s_i = (y_i - \tau^* + \mu_i^*)_+$, where $s_{ij} = (\theta_i^j - \tau^* + \mu_{ij}^*)_+$.

$$\begin{aligned}
||p_i^*||_2 &= ||(1 + \frac{\lambda}{||p_i^*||_2})^{-1} s_i||_2 \\
||p_i^*||_2 &= \sqrt{(1 + \frac{\lambda}{||p_i^*||_2})^{-2}\sum_{j=1}^{m} s_{ij}^2} \\
||p_i^*||_2 &= (1 + \frac{\lambda}{||p_i^*||_2})^{-1}||s_i||_2 \\
||p_i^*||_2(1 + \frac{\lambda}{||p_i^*||_2}) &= ||s_i||_2 \\
||p_i^*||_2 + \lambda &= ||s_i||_2 \\
||p_i^*||_2 &= ||s_i||_2 - \lambda
\end{aligned} \tag{6}$$

Take the resulting $||p_i^*||_2$ into (5):

$$\begin{aligned}
p_i^* &= (1 + \frac{\lambda}{||s_i||_2})^{-1} s_i \\
p_i^* &= (1 - \frac{\lambda}{||s_i||_2}) s_i
\end{aligned} \tag{7}$$

2) When $p_i = 0$, because the derivative does not exist, we try to solve it by sub gradient: let $f(p_i) = ||p_i||^2$,when $p_i = 0$,

$$\begin{aligned}
\frac{\partial f(p_i)}{\partial p_i} &= \{v \in R^d | f(p_i^{'}) \leq f(p_i) + v^T(p_i^{'} - p_i), \forall p_i^{'} \in R^d\} \\
&= \{v \in R^d | ||p_i^{'}||_2 \leq v^T p_i^{'}, \forall p_i^{'} \in R^d\}
\end{aligned} \tag{8}$$

Thus, the sub gradient v of p_i have to satisfy $||v|| \leq 1$ when $p_i = 0$. And it has to satisfy $0 \in -y_i + \lambda v + \tau - \mu_i$ according the KKT conditions, we can get the follows:

$$\begin{aligned} -y_i + \lambda v + \tau - \mu_i &= 0 \\ \lambda v &= y_i - \tau + \mu_i \\ v &= \frac{1}{\lambda}(y_i - \tau + \mu_i) \\ v &= \frac{1}{\lambda} s_i \end{aligned} \tag{9}$$

Since $||v||_2 \leq 1$, thus:

$$\begin{aligned} ||\frac{1}{\lambda} s_i||_2 &\leq 1 \\ ||s_i||_2 &\leq \lambda \end{aligned} \tag{10}$$

Thus, we can get $p_i = 0$ when $||s_i||_2 \leq \lambda$.

According to the above analysis, we can get that the solution of p_i is as follows:

$$\begin{aligned} p_i &= (1 - \frac{\lambda}{||s_i||_2})_+ s_i \\ &= (1 - \frac{\lambda}{||(y_i - \tau + \mu_i)_+||_2})_+ (y_i - \tau + \mu_i)_+ \end{aligned} \tag{11}$$

Furthermore. according to (4), the $\mu_i = 0$ when $p_i > 0$. Therefore, it can be reduced to:

$$p_i = (1 - \frac{\lambda}{||(y_i - \tau)_+||_2})_+ (y_i - \tau(y))_+ \tag{12}$$