Supplemental Materials: Multiplicative Sparse Feature Decomposition for Efficient Multi-View Multi-Task Learning

For convenience, we first introduce the general formula of the proposed SPLIT:

$$\begin{aligned}&\min_{\mathbf{\Theta}} \sum_{t=1}^{T} L\left(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\boldsymbol{\theta}_t\right) + \lambda_1 \sum_{k=1}^{K}\sum_{v=1}^{V} \|\boldsymbol{\alpha}_k^v\|_p^p + \lambda_2 \sum_{k=1}^{K}\sum_{v=1}^{V} |\beta_k^v|^k + \eta \|\mathbf{H}\|_{\mathrm{F}}^2, \\ &\text{s.t. } \mathbf{\Theta} = (\mathbf{A} \circ \mathbf{B})\mathbf{H}, \quad \mathbf{B} \geq 0,\end{aligned} \tag{1}$$

In the subsequent sections, we present the proofs for Theorems 1 and Proposition 1 in the main paper.

1 Proof of Theorem 1 in the main paper

Theorem 1. *Let* $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ *be the optimal solution of the following optimization problem,*

$$\min_{\mathbf{\Theta}=\mathbf{WH}} \sum_{t=1}^{T} L(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\boldsymbol{\theta}_t) + \gamma \sum_{k=1}^{K}\sum_{v=1}^{V} \sqrt[2s]{\|\mathbf{w}_k^v\|_p^p} + \eta \|\mathbf{H}\|_F^2, \tag{2}$$

where $\mathbf{w}_t^v$ *is the* v*-th view sub-vector of the* k*-th column of* $\mathbf{W}$*. If* $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{H}})$ *is the optimal solution of (1), the following equation holds,*

$$\hat{\mathbf{W}} = \hat{\mathbf{A}} \circ \hat{\mathbf{B}} = \hat{\mathbf{A}} \circ \left(\begin{bmatrix} \mathbf{1}_{d_1} & & \\ & \ddots & \\ & & \mathbf{1}_{d_V} \end{bmatrix} \hat{\mathcal{B}} \right), \tag{3}$$

given $\gamma = 2\sqrt{\lambda_1^{2-\frac{p}{qs}} \lambda_2^{\frac{p}{qs}}}$ *and* $s = \frac{p+q}{2q}$*. In addition, the view-weighting solution* $\hat{\mathbf{B}}$ *is related with the feature sparse solution* $\hat{\mathbf{A}}$ *by the following formula*

$$\beta_k^v = \sqrt[q]{\lambda_1 \lambda_2^{-1} \|\boldsymbol{\alpha}_k^v\|_p^p}, \quad \forall v, k. \tag{4}$$

To prove the equivalence of (1) and (2), we need an intermediate optimization problem

$$\min_{\mathbf{\Theta}=\mathbf{WH}} \sum_{t=1}^{T} L(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\boldsymbol{\theta}_t) + \mu_1 \sum_{k=1}^{K}\sum_{v=1}^{V} (\rho_k^v)^{-1} \|\mathbf{w}_k^v\|_p^{p/s} + \mu_2 \sum_{k=1}^{K}\sum_{v=1}^{V} \rho_k^v + \eta \|\mathbf{H}\|_{\mathrm{F}}^2. \tag{5}$$

Similar with the proof in [Wang *et al.*, 2016], we first prove the equivalence of the optimal solutions of (2) and (5) by Theorem 2, then show the equivalence of the optimal solutions of (2) and (1) by Theorem 3, and finally reach our conclusion in Theorem 1.

Theorem 2. *Given arbitrary* $\mathbf{W}$ *and* $\mathbf{H}$*, the optimization problem (5) can be optimized w.r.t.* $\boldsymbol{\rho}$ *by the following setting,*

$$\rho_k^v = \sqrt{\mu_1 \mu_2^{-1} \|\mathbf{w}_k^v\|_p^{p/s}}, \quad \forall k, v, \tag{6}$$

with $\|\mathbf{w}_k^v\|_p^{p/s} = \sqrt[s]{\|\mathbf{w}_k^v\|_p^p}$*, in which case the objective function in (2) is the lower bound of the objective function in (5) with* $\gamma = 2\sqrt{\mu_1\mu_2}$*. Furthermore, the optimal solutions of (5) and (2) are identical with* $\gamma = 2\sqrt{\mu_1\mu_2}$*.*

Proof. According to the Cauchy-Schwarz inequality, the following inequality holds

$$\mu_1 \sum_{k=1}^{K}\sum_{v=1}^{V} (\rho_k^v)^{-1} \|\mathbf{w}_k^v\|_p^{p/s} + \mu_2 \sum_{k=1}^{K}\sum_{v=1}^{V} \rho_k^v \geq 2\sqrt{\mu_1\mu_2} \sum_{k=1}^{K}\sum_{v=1}^{V} \sqrt{\|\mathbf{w}_k^v\|_p^{p/s}}, \tag{7}$$

in which the equation is satisfied when $\rho_k^v = \sqrt{\mu_1 \mu_2^{-1} \|\mathbf{w}_k^v\|_p^{p/s}}$. Thus, substituting (7) into (5), we can see that the objective function of (2) is the lower bound of the objective function of (5) with the setting that $\gamma = 2\sqrt{\mu_1\mu_2}$. Let $J_1(\mathbf{W}, \mathbf{H})$ and $J_2(\mathbf{W}, \boldsymbol{\rho}, \mathbf{H})$ denote the objective functions of (2) and (5), respectively. Given arbitrary $\mathbf{W}$ and $\mathbf{H}$, we have $J_1(\mathbf{W}, \mathbf{H}) \leq J_2(\mathbf{W}, \boldsymbol{\rho}, \mathbf{H})$ and the equality holds if $\boldsymbol{\rho}$ is calculated by (6).

In order to prove the equivalence of the optimal solutions of (2) and (5), we first prove that once $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ optimizes (2), $(\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}})$ optimizes (5) with $\hat{\rho}_k^v = \sqrt{\mu_1 \mu_2^{-1} \|\hat{\mathbf{w}}_k^v\|_p^{p/s}}$, $\forall k, v$, leading to

$$(\hat{\mathbf{W}}, \hat{\mathbf{H}}) = \arg\min_{\mathbf{W},\mathbf{H}} J_1(\mathbf{W}, \mathbf{H}) \;\Rightarrow\; J_1(\hat{\mathbf{W}}, \hat{\mathbf{H}}) = J_2(\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}}). \tag{8}$$

Here, the proof is provided by the approach of contradiction. Let's assume that there exists an optimal solution $(\tilde{\mathbf{W}}, \tilde{\boldsymbol{\rho}}, \tilde{\mathbf{H}}) \neq (\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}})$ for problem (5), such that $J_2(\tilde{\mathbf{W}}, \tilde{\boldsymbol{\rho}}, \tilde{\mathbf{H}}) < J_2(\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}})$. According to (7) and $\tilde{\rho}_k^v = \sqrt{\mu_1 \mu_2^{-1} \left\|\tilde{\mathbf{w}}_k^v\right\|_p^{p/s}}$, $\forall k, v$, (2) can achieves lower objective value at $(\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$ than that of (5), since its objective function is a lower bound of (5), namely, $J_1(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}) \leq J_2(\tilde{\mathbf{W}}, \tilde{\boldsymbol{\rho}}, \tilde{\mathbf{H}})$. Thus, we have the following equation,

$$J_1(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}) \leq J_2(\tilde{\mathbf{W}}, \tilde{\boldsymbol{\rho}}, \tilde{\mathbf{H}}) < J_2(\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}}) = J_1(\hat{\mathbf{W}}, \hat{\mathbf{H}}), \tag{9}$$

which contradicts to the first formula in (8).

Then we prove that once $(\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}})$ optimizes (5) with $\hat{\rho}_k^v = \sqrt{\mu_1 \mu_2^{-1} \left\|\hat{\mathbf{w}}_k^v\right\|_p^{p/s}}$, $\forall k, v$, $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ also optimizes (2). In this case, what we need to prove is that

$$(\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}}) = \underset{\mathbf{W}, \boldsymbol{\rho}, \mathbf{H}}{\arg\min}\, J_2(\mathbf{W}, \boldsymbol{\rho}, \mathbf{H}) \;\Rightarrow\; J_1(\hat{\mathbf{W}}, \hat{\mathbf{H}}) = J_2(\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}}). \tag{10}$$

Similarly, we assume that there is an optimal solution $(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}) \neq (\hat{\mathbf{W}}, \hat{\mathbf{H}})$ for problem (2), such that $J_1(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}) < J_1(\hat{\mathbf{W}}, \hat{\mathbf{H}})$. Let $\tilde{\rho}_k^v = \sqrt{\mu_1 \mu_2^{-1} \left\|\tilde{\mathbf{w}}_k^v\right\|_p^{p/s}}$, it follows that $J_2(\tilde{\mathbf{W}}, \tilde{\boldsymbol{\rho}}, \tilde{\mathbf{H}}) = J_1(\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$. Hence, according to (10), we have $J_2(\tilde{\mathbf{W}}, \tilde{\boldsymbol{\rho}}, \tilde{\mathbf{H}}) < J_2(\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}})$, contradicting to the first formula in (10). □

Theorem 3. *Given the setting that $\boldsymbol{\alpha}_k^v = |\beta_k^v|^{-1}\mathbf{w}_k^v$ and $\rho_k^v = |\beta_k^v|^q$, the optimal solution $(\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}})$ of problem (5) is equivalent to the optimal solution $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{H}})$ of problem (1), when the following equations hold,*

$$\lambda_1 = \mu_1^{\frac{qs}{2qs-p}} \mu_2^{\frac{qs-p}{2qs-p}}, \quad \lambda_2 = \mu_2, \quad s = \frac{p+q}{2q}. \tag{11}$$

Proof. Based on $\boldsymbol{\alpha}_k^v = |\beta_k^v|^{-1}\mathbf{w}_k^v$, the objective function J_0 of (1) can be reformulated as follows

$$\begin{aligned} J_0 &= \sum_{t=1}^{T} L(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\boldsymbol{\theta}_t) + \lambda_1 \sum_{k=1}^{K}\sum_{v=1}^{V} (\beta_k^v)^{-p} \left\|\mathbf{w}_k^v\right\|_p^p + \lambda_2 \sum_{k=1}^{K}\sum_{v=1}^{V} |\beta_k^v|_q^q + \eta \left\|\mathbf{H}\right\|_{\mathrm{F}}^2, \\ &\geq \sum_{t=1}^{T} L(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\boldsymbol{\theta}_t) + 2\lambda_1^{\frac{q}{p+q}} \lambda_2^{\frac{p}{p+q}} \sum_{k=1}^{K}\sum_{v=1}^{V} (\left\|\mathbf{w}_k^v\right\|_p^p)^{\frac{q}{p+q}} + \eta \left\|\mathbf{H}\right\|_{\mathrm{F}}^2. \end{aligned} \tag{12}$$

The inequality in (12) holds due to the Cauchy-Schwarz inequality, in which the equation is satisfied if

$$\beta_k^v = \left(\lambda_1 \lambda_2^{-1} \left\|\mathbf{w}_k^v\right\|_p^p\right)^{\frac{1}{p+q}}. \tag{13}$$

In this sense, given arbitrary $\mathbf{W}$ and $\mathbf{H}$, the objective function of (1) is optimized when $\mathbf{B}$ follows (13), and its lower bound is achieved by (12).

First, let's prove the equivalence of optimal solutions $(\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}})$ from (5) to (1). Let's J_2 denote the objective function of (5), substituting $\hat{\mathbf{w}}_k^v = \hat{\beta}_k^v \hat{\boldsymbol{\alpha}}_k^v$ and $\hat{\rho}_k^v = |\hat{\beta}_k^v|^q$ into it, J_2 becomes

$$\begin{aligned} J_2 &= \sum_{t=1}^{T} L(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\hat{\boldsymbol{\theta}}_t) + \mu_1 \sum_{k=1}^{K}\sum_{v=1}^{V} |\hat{\beta}_k^v|^{-q} \left\|\hat{\beta}_k^v \hat{\boldsymbol{\alpha}}_k^v\right\|_p^{p/s} + \mu_2 \sum_{k=1}^{K}\sum_{v=1}^{V} |\hat{\beta}_k^v|^q + \eta \left\|\hat{\mathbf{H}}\right\|_{\mathrm{F}}^2 \\ &= \sum_{t=1}^{T} L(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\hat{\boldsymbol{\theta}}_t) + \mu_1 \sum_{k=1}^{K}\sum_{v=1}^{V} |\hat{\beta}_k^v|^{\frac{p}{s}-q} \left\|\hat{\boldsymbol{\alpha}}_k^v\right\|_p^{p/s} + \mu_2 \sum_{k=1}^{K}\sum_{v=1}^{V} |\hat{\beta}_k^v|^q + \eta \left\|\hat{\mathbf{H}}\right\|_{\mathrm{F}}^2. \end{aligned} \tag{14}$$

According to $\hat{\rho}_k^v = |\beta_k^v|^q$ and (6) in Theorem 2, we have

$$\hat{\beta}_k^v = \left(\mu_1 \mu_2^{-1} \left\|\hat{\boldsymbol{\alpha}}_k^v\right\|_p^{p/s}\right)^{\frac{s}{2qs-p}}. \tag{15}$$

Substituting (15) into (14) leads to

$$\begin{aligned} J_2 &= \sum_{t=1}^{T} L(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\hat{\boldsymbol{\theta}}_t) + \mu_1 \sum_{k=1}^{K}\sum_{v=1}^{V} \left(\mu_1 \mu_2^{-1} \left\|\hat{\boldsymbol{\alpha}}_k^v\right\|_p^{p/s}\right)^{\frac{p-qs}{2qs-p}} \left\|\hat{\boldsymbol{\alpha}}_k^v\right\|_p^{p/s} + \mu_2 \sum_{k=1}^{K}\sum_{v=1}^{V} |\hat{\beta}_k^v|^q + \eta \left\|\hat{\mathbf{H}}\right\|_{\mathrm{F}}^2 \\ &= \sum_{t=1}^{T} L(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\hat{\boldsymbol{\theta}}_t) + \mu_1^{\frac{qs}{2qs-p}} \mu_2^{\frac{qs-p}{2qs-p}} \sum_{k=1}^{K}\sum_{v=1}^{V} \left(\left\|\hat{\boldsymbol{\alpha}}_k^v\right\|_p^{p/s}\right)^{\frac{qs}{2qs-p}} + \mu_2 \sum_{k=1}^{K}\sum_{v=1}^{V} |\hat{\beta}_k^v|^q + \eta \left\|\hat{\mathbf{H}}\right\|_{\mathrm{F}}^2. \end{aligned} \tag{16}$$

Therefore, given the setting (11), once the solutions $(\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}})$ optimizes (5), then $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{H}})$ optimizes the problem (1) with $\hat{\boldsymbol{\alpha}}_k^v = |\hat{\beta}_k^v|^{-1}\hat{\mathbf{w}}_k^v$ and $\hat{\rho}_k^v = |\hat{\beta}_k^v|^q, \forall k, v$.

Then, we will prove the equivalence of optimal solutions from (1) to (5) in a similar way. Given the optimal solutions $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{H}})$ of (1), substituting $\hat{\boldsymbol{\alpha}}_k^v = |\hat{\beta}_k^v|^{-1}\hat{\mathbf{w}}_k^v$ and $\hat{\beta}_k^v = |\hat{\rho}_k^v|^{\frac{1}{q}}$ into (1), the objective J_0 of (1) becomes

$$J_0 = \sum_{t=1}^{T} L(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\hat{\boldsymbol{\theta}}_t) + \lambda_1 \sum_{k=1}^{K}\sum_{v=1}^{V}(\hat{\rho}_k^v)^{-\frac{p}{q}} \left\|\hat{\mathbf{w}}_k^v\right\|_p^p + \lambda_2 \sum_{k=1}^{K}\sum_{v=1}^{V}\hat{\rho}_k^v + \eta \left\|\hat{\mathbf{H}}\right\|_{\mathrm{F}}^2, \tag{17}$$

indicating that $\hat{\mathbf{W}}$ and $\hat{\boldsymbol{\rho}}$ optimize the problem (17). Based on equivalence condition of the Cauchy-Schwarz inequality, we can show that the optimal $\hat{\boldsymbol{\rho}}$ follows the following equation,

$$\hat{\rho}_k^v = \left(\lambda_1\lambda_2^{-1}\left\|\hat{\mathbf{w}}_k^v\right\|_p^p\right)^{\frac{q}{p+q}}. \tag{18}$$

Substituting (18) into (17), (17) is rewritten as

$$J_0 = \sum_{t=1}^{T} L(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\hat{\boldsymbol{\theta}}_t) + \lambda_1^{\frac{2q}{p+q}}\lambda_2^{\frac{p-q}{p+q}} \sum_{k=1}^{K}\sum_{v=1}^{V}(\hat{\rho}_k^v)^{-1}\left(\left\|\hat{\mathbf{w}}_k^v\right\|_p^p\right)^{\frac{2q}{p+q}} + \lambda_2 \sum_{k=1}^{K}\sum_{v=1}^{V}\hat{\rho}_k^v + \eta\left\|\hat{\mathbf{H}}\right\|_{\mathrm{F}}^2. \tag{19}$$

With the setting (11), (19) becomes the exact form of (5), and thus $(\hat{\mathbf{W}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{H}})$ optimizes the problem (5).

□

2 Proof of Proposition 1 in the main paper

Given the proposed optimization problem of SPLIT,

$$\begin{aligned} \min_{\boldsymbol{\Theta}} \sum_{t=1}^{T} L\left(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\boldsymbol{\theta}_t\right) + \lambda_1\left\|\mathbf{A}\right\|_{1,1} + \lambda_2\left\|\mathcal{B}\right\|_{\mathrm{F}}^2 + \eta\left\|\mathbf{H}\right\|_{\mathrm{F}}^2, \\ \text{s.t. } \boldsymbol{\Theta} = (\mathbf{A}\circ\mathbf{B})\mathbf{H}, \quad \mathbf{B} \geq 0, \end{aligned} \tag{20}$$

Proposition 1 is derived for the proposed optimization algorithm in the main paper.

Proposition 1. *The iterative optimization algorithm of Algorithm 1 does not increase the objective function of (20) in each iteration, indicating that*

$$J(\mathbf{A}^{(i+1)}, \mathbf{B}^{(i+1)}, \mathbf{H}^{(i+1)}) \leq J(\mathbf{A}^{(i)}, \mathbf{B}^{(i)}, \mathbf{H}^{(i)}), \tag{21}$$

in the $(i+1)$-th iteration, with $J(\mathbf{A}, \mathbf{B}, \mathbf{H})$ denoting the objective function of (20) w.r.t. $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{H}$.

Proof. Let $J(\mathbf{A}, \mathbf{B}, \mathbf{H})$ denotes the objective function in (1), we have

$$f(\mathbf{A}, \mathbf{B}, \mathbf{H}) = \sum_{t=1}^{T} L\left(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t(\mathbf{A}\circ\mathbf{B})\mathbf{h}_t\right) + \lambda_1\sum_{k=1}^{K}\left\|\boldsymbol{\alpha}_k\right\|_p^p + \lambda_2\sum_{k=1}^{K}\left\|\boldsymbol{\beta}_k\right\|_q^q + \eta\left\|\mathbf{H}\right\|_{\mathrm{F}}^2. \tag{22}$$

Note that here we aim to provide a proof for a general version of $J(\mathbf{A}, \mathbf{B}, \mathbf{H})$, where the objective function in (20) is a special case of (20) in setting of $p = 1$ and $q = 2$. According to the equation $\mathbf{W} = \mathbf{A}\circ\mathcal{B}$ and Theorem 1, (22) is equivalent to the following rewritten formula:

$$J(\mathbf{W}, \mathbf{B}, \mathbf{H}) = \sum_{t=1}^{T} L\left(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\mathbf{W}\mathbf{h}_t\right) + \lambda_1\sum_{k=1}^{K}\sum_{v=1}^{V}|\beta_k^v|^{-p}\left\|\mathbf{w}_k^v\right\|_p^p + \lambda_2\sum_{k=1}^{K}\sum_{v=1}^{V}|\beta_k^v|^q + \eta\left\|\mathbf{H}\right\|_{\mathrm{F}}^2. \tag{23}$$

Suppose that the optimal solution $\{\mathbf{A}^{(i)}, \mathbf{B}^{(i)}, \mathbf{H}^{(i)}\}$ at the i-th iteration is calculated by Algorithm 1, we first attempt to prove that, updating $\mathbf{b}$ based on fixed $\mathbf{A}$ and $\mathbf{H}$ and Table 1 in the main paper does not increase the objective (22), i.e.,

$$J(\mathbf{A}^{(i)}, \mathbf{B}^{(i+1)}, \mathbf{H}^{(i)}) \leq J(\mathbf{A}^{(i)}, \mathbf{B}^{(i)}, \mathbf{H}^{(i)}). \tag{24}$$

According Cauchy-Schwarz inequality, we have

$$J(\mathbf{W}, \mathbf{B}, \mathbf{H}) \geq \sum_{t=1}^{T} L\left(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\mathbf{W}\mathbf{h}_t\right) + \lambda_1^{\frac{q}{p+q}}\lambda_2^{\frac{p}{p+q}}\sum_{k=1}^{K}\sum_{v=1}^{V}(\left\|\mathbf{w}_k^v\right\|_p^p)^{\frac{q}{p+q}} + \eta\left\|\mathbf{H}\right\|_{\mathrm{F}}^2, \tag{25}$$

Algorithm 1 Optimization algorithm of SPLIT

Input: $\{\mathbf{X}_t^v\}_{t,v}$, $\{\mathbf{y}_t\}_t$, λ_1, λ_2, η, K.
Output: $\mathbf{\Theta} = \mathbf{WH} = (\mathbf{A} \circ \mathbf{B})\mathbf{H}$.

1: Initialized $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{H}$.
2: **repeat**
3: Update $\mathbf{A}$ by solving the problem: $\min_{\mathbf{A}} \sum_{t=1}^{T} \left\|\mathbf{y}_t - \frac{1}{V}\mathbf{X}_t(\mathbf{A} \circ \mathbf{B})\mathbf{h}_t\right\|^2 + \lambda_1 \|\mathbf{A}\|_{1,1}$.
4: Update $\mathbf{B}$ by computing $\beta_k^v = \sqrt{\lambda_1 \lambda_2^{-1} \|\boldsymbol{\alpha}_k^v\|_1}$, $\forall k, v$.
5: Update $\mathbf{H}$ by solving the problem: $\min_{\mathbf{H}} \sum_{t=1}^{T} \left\|\mathbf{y}_t - \frac{1}{V}\mathbf{X}_t\mathbf{W}\mathbf{h}_t\right\|^2 + \eta \|\mathbf{H}\|_{\mathrm{F}}^2$.
6: **until** *Convergence*

and the equation holds when $\mathbf{B}$ follows (4). Thus, updating $\mathbf{B}^{(i+1)}$ by (4) with fixed $\mathbf{A}^{(i)}$ and $\mathbf{H}^{(i)}$ will reach the lower bound of $f(\mathbf{W}^{(i)}, \mathbf{B}, \mathbf{H}^{(i)})$, leading to

$$J(\mathbf{W}^{(i)}, \mathbf{B}^{(i+1)}, \mathbf{H}^{(i)}) \leq f(\mathbf{W}^{(i)}, \mathbf{B}^{(i)}, \mathbf{H}^{(i)}). \tag{26}$$

Since $J(\mathbf{A}, \mathbf{B}, \mathbf{H}) = J(\mathbf{W}, \mathbf{B}, \mathbf{H})$ with $\mathbf{W} = \mathbf{A} \circ \mathcal{B}$, (26) can be reformulated as (24).

Next, when $\mathbf{B}$ and $\mathbf{H}$ are fixed, $\mathbf{A}$ is updated by solving the optimization problem:

$$\min_{\mathbf{A}} \sum_{t=1}^{T} L\left(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t(\mathbf{A} \circ \mathcal{B})\mathbf{h}_t\right) + \lambda_1 \sum_{k=1}^{K} \|\mathbf{a}_k\|_p^p. \tag{27}$$

Once a convex and smooth loss function, like squared loss, hinge loss and logistic loss, as well as a simple regularization norm, such as ℓ_2-norm and ℓ_1-norm, are employed in (27), efficient gradient descent algorithm [Robbins and Monro, 1951; Nesterov, 2013] can be applied to solve (27) with global convergence property w.r.t. $\mathbf{A}$. Therefore, updating $\mathbf{A}$ with fixed $\mathbf{B}^{(i+1)}$ and $\mathbf{H}^{(i)}$ leads to

$$J(\mathbf{A}^{(i+1)}, \mathbf{B}^{(i+1)}, \mathbf{H}^{(i)}) \leq J(\mathbf{A}^{(i)}, \mathbf{B}^{(i+1)}, \mathbf{H}^{(i)}). \tag{28}$$

Finally, once $\mathbf{A}$ and $\mathbf{B}$ are fixed, $\mathbf{H}$ is updated by soling the following problem:

$$\min_{\mathbf{H}} \sum_{t=1}^{T} L\left(\mathbf{y}_t, \frac{1}{V}\mathbf{X}_t\mathbf{W}\mathbf{h}_t\right) + \eta \|\mathbf{H}\|_{\mathrm{F}}^2. \tag{29}$$

Given the squared loss function, problem (31) has a closed-form solution, indicating that

$$J(\mathbf{A}^{(i+1)}, \mathbf{B}^{(i+1)}, \mathbf{H}^{(i+1)}) \leq J(\mathbf{A}^{(i+1)}, \mathbf{B}^{(i+1)}, \mathbf{H}^{(i)}). \tag{30}$$

Taking (24), (28) and (30) into consideration, we have

$$J(\mathbf{A}^{(i+1)}, \mathbf{B}^{(i+1)}, \mathbf{H}^{(i+1)}) \leq J(\mathbf{A}^{(i+1)}, \mathbf{B}^{(i+1)}, \mathbf{H}^{(i)}) \leq J(\mathbf{A}^{(i)}, \mathbf{B}^{(i+1)}, \mathbf{H}^{(i)}) \leq J(\mathbf{A}^{(i)}, \mathbf{B}^{(i)}, \mathbf{H}^{(i)}), \tag{31}$$

in accordance with the conclusion (21) in Proposition 1. □

3 Implementation and convergence analysis for Section: Optimization algorithm

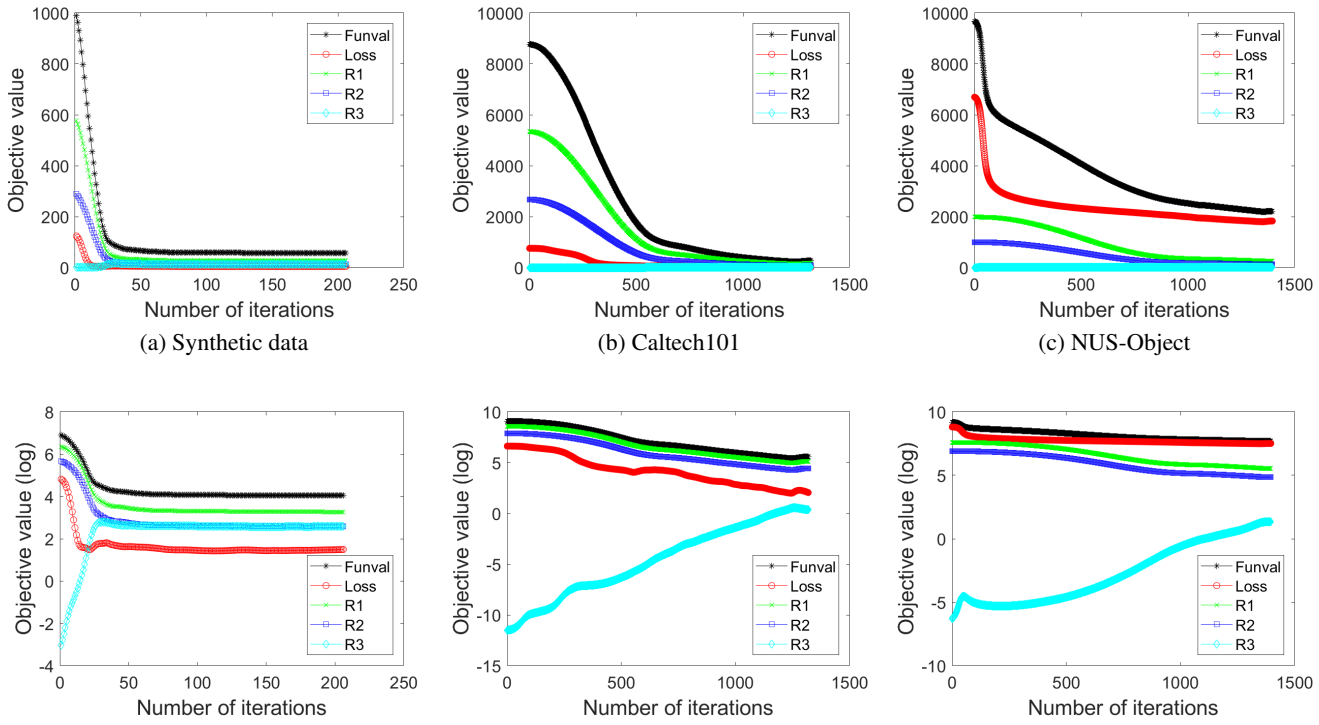
In Algorithm 1, we provide the optimization algorithm of SPLIT discussed in Sec.5 of the main paper. We apply an alternating algorithm to update $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{H}$ in an iterative manner. It is worth noting that, compared optimizing $\mathbf{W}$ directly, optimizing optimizing $\mathbf{A}$ and $\mathbf{B}$ separately brings in a very small number of extra variables, as the number $(V \times K)$ of effective parameters in $\mathbf{B}$ is typically very small.

To evaluate the convergence ability of Algorithm 1, we conduct experiment on one synthetic dataset, and two real-world datasets, Caltech101 and NUS-Object. In this experiment, we randomly select 30%, 20% and 20% of total samples as training set, validation set and testing set, respectively, and set the parameters of SPLIT as $\lambda_1 = \lambda_2 = \eta = 1$. We terminate Algorithm 1 once the relative change of its objective is below 10^{-5}. Figure 1 shows the convergence curves of the objective function value by Algorithm 1.

4 Data preparation for Section: Experimetns

The following four real-world datasets are used for evaluating comparing methods in the main paper.

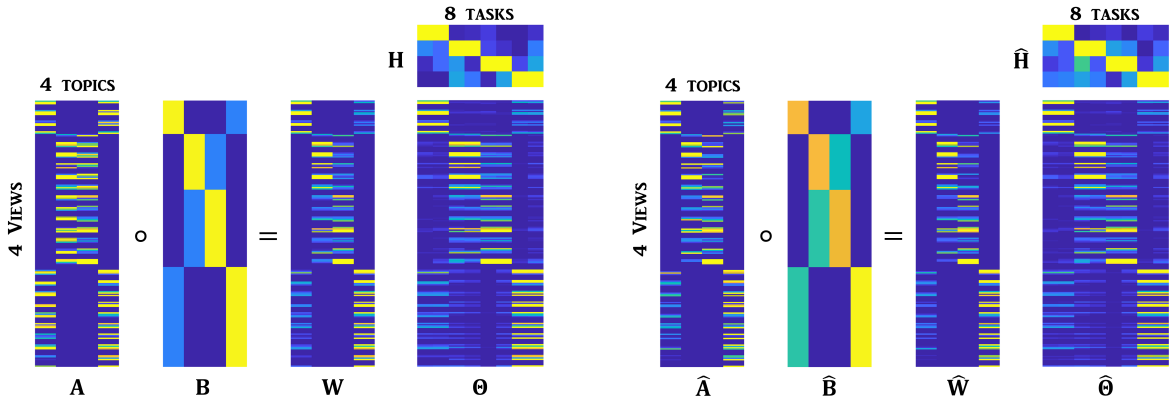
- **Mirflickr**: It collects 25,000 Flickr images, which has 15 relevant labels (*tasks*). Each image (*sample*) is represented by two types (*views*) of features: image edge histogram and image homogeneous texture.



(a) Synthetic data (b) Caltech101 (c) NUS-Object

(d) Synthetic data (e) Caltech101 (f) NUS-Object

Figure 1: Convergence analysis of Algorithm 1 ($\lambda_1 = \lambda_2 = \eta = 1$) on one synthetic dataset and two real-world datasets. The algorithm converged at the 206th, 1318th and 1395th iteration on the synthetic data, Caltech101, and NUS-Object, respectively. **The 1st row** shows the original objective value, while **the 2nd row** shows the objective value in the logarithmic scale. In each sub-figure, Funval and Loss denote the objective value and value of loss function, respectively, and R1, R2 and R3 denote the values of three regularization terms.



(a) Designed model $\boldsymbol{\Theta}$ (b) Learned model $\hat{\boldsymbol{\Theta}}$ by SPLIT$_1$

Figure 2: Illustration of multiplicative feature decomposition (**Right hand**) by SPLIT$_1$ on simulated data with designed model $\boldsymbol{\Theta}$ (**Left hand**). The model $\widehat{\boldsymbol{\Theta}} = (\widehat{\mathbf{A}} \circ \widehat{\mathbf{B}})\widehat{\mathbf{H}}$ is learned by SPLIT$_1$. Darker yellow (blue) color indicates larger (smaller) values in magnitude.

- **Caltech101**: It collects images (*samples*) of objects belonging to 101 categories (*tasks*), where each category (*task*) has 40-800 images (*samples*), and each image has 6 types (*views*) of low-level features.
- **NUS-Object**, **NUS-Scene**: They are extracted from the NUS-WIDE dataset for web image annotation. Images (*samples*) are annotated by a set of class labels (*tasks*), and each image has 5 types (*views*) of low-level features.

To preprocess the datasets, we filter out textual features with a low frequency in ℓ_2-normalized TFIDF, and discard tasks with a relatively small number of positive instances.

5 Case study on feature decomposition for Section: Experiments

We illustrate multiplicative feature decomposition of SPLIT on one designed synthetic dataset in Fig 2, where $\widehat{\boldsymbol{\Theta}} = (\widehat{\mathbf{A}} \circ \widehat{\mathbf{B}})\widehat{\mathbf{H}}$ is learned by SPLIT with the setting $\lambda_1 = 10^1$, $\lambda_2 = 10^3$ and $\eta = 10^4$. To quantitatively measure the closeness between the patterns of $\boldsymbol{\Theta}$ and $\hat{\boldsymbol{\Theta}}$ in Fig. 2, we introduce the following similarity metric for arbitrary matrices $\mathbf{C}$ and $\mathbf{D}$[1] in the same size,

$$sim(\mathbf{C}, \mathbf{D}) = 1 - \frac{\|\mathbf{C} - \mathbf{D}\|_F^2}{\|\mathbf{D}\|_F^2} \in [0, 1]. \tag{32}$$

The higher the value of $sim(\mathbf{C}, \mathbf{D})$, the more similar between $\mathbf{C}$ and $\mathbf{D}$. We calculate the similarity between the underlying model and the learned model, and list the results in Table 1. As shown in Fig. 2 and Table 1, SPLIT$_1$ successfully detects the

Table 1: Similarity between the underlying model and the learned model.

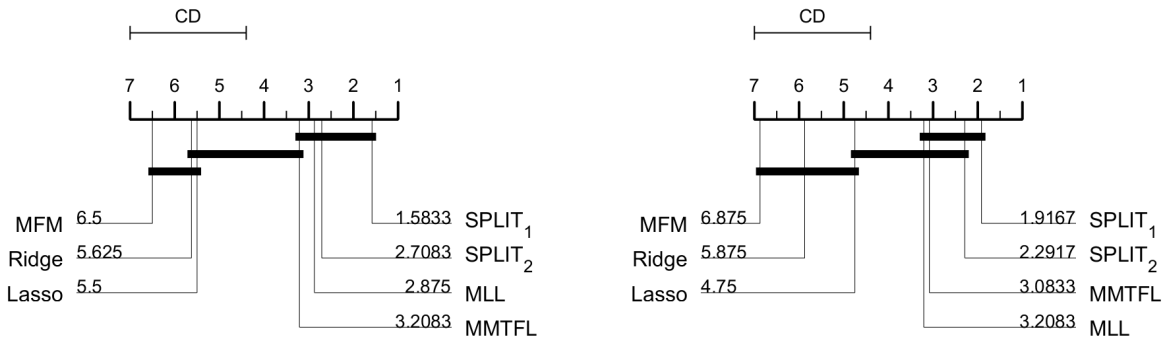
$sim(\mathbf{A}, \hat{\mathbf{A}})$	$sim(\mathbf{B}, \hat{\mathbf{B}})$	$sim(\mathbf{W}, \hat{\mathbf{W}})$	$sim(\mathbf{H}, \hat{\mathbf{H}})$	$sim(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}})$
0.8394	0.9341	0.9902	0.9944	0.9944

underlying models by selecting topic-specific features in $\mathbf{A}$, learning view-wise weights in $\mathbf{B}$, and saving task correlation in $\mathbf{W} = \mathbf{A} \circ \mathbf{B}$ and $\mathbf{H}$.

6 Statistical test for Section: Experiments

To perform statistical test on experimental results in Table 2 of the main paper, we apply Nemenyi test [Demšar, 2006], which allows to statistically evaluate the performance between every two methods. In Nemenyi test, the performance of two methods is regarded as significantly different if their average ranks differ by at least the critical difference (CD). Fig. 3 shows the CD diagrams for four evaluation metrics at 0.05 significance level. In each subfigure, the CD is given above the axis, where the averaged rank is marked. In Fig. 3, algorithms which are not significantly different are connected by a thick line. As shown in Fig. 3, SPLIT$_1$ and SPLIT$_2$ ranked 1st and 2nd, respectively, and statistically outperformed Lasso, Ridge and MFM in both evaluation metrics. Two multiplicative feature decomposition methods, MLL and MMTFL, achieved statistically comparable performance with SPLIT. The observation shows that decomposing weight matrix multiplicatively indeed incorporates better generalization ability with the learning model, leading to superior prediction accuracy than the baselines.

[1]Both matrices have been normalized such that any element in $\mathbf{C}$ and $\mathbf{D}$ lies in $[0, 1]$



(a) AUC (b) Accuracy

Figure 3: CD diagrams (0.05 significance level) of seven comparing methods in two evaluation metrics. The performance of two methods is regarded as significantly different if their average ranks differ by at least the Critical Difference (CD).

References

[Demšar, 2006] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[Nesterov, 2013] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[Robbins and Monro, 1951] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[Wang *et al.*, 2016] Xin Wang, Jinbo Bi, Shipeng Yu, Jiangwen Sun, and Minghu Song. Multiplicative multitask feature learning. *Journal of Machine Learning Research*, 17(80):1–33, 2016.