Supplementary Materials for
"RaFM: Rank-Aware Factorization Machines"

1. Proof of Theorem 6

Proof. Recall the estimated gradient:

$$\widehat{grad} = \frac{\partial L(\mathcal{B}_{1,m}, y)}{\partial\, \boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}} + \frac{1}{N}\sum_{\boldsymbol{x'}} \frac{\partial L(\mathcal{B}'_{1,m}, y)}{\partial\, \boldsymbol{v}^{(p-1)}\big|_{\mathcal{F}_p}} \boldsymbol{G}^{-1} \frac{\partial^2 L(\mathcal{B}_{1,p-1}, \mathcal{B}_{1,p})}{\partial\, \boldsymbol{v}^{(p-1)}\big|_{\mathcal{F}_p}\, \partial\, \boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}}$$

According to the chain rule and the fact that $\mathcal{B}_{1,p-1}$ does not contain $\boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}$, we have

$$\begin{aligned}
&\frac{\partial^2 L(\mathcal{B}_{1,p-1}, \mathcal{B}_{1,p})}{\partial\, \boldsymbol{v}^{(p-1)}\big|_{\mathcal{F}_p}\, \partial\, \boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}} \\
=&L_{12}(\mathcal{B}_{1,p-1}, \mathcal{B}_{1,p}) \left(\frac{\partial \mathcal{B}_{1,p-1}}{\partial\, \boldsymbol{v}^{(p-1)}\big|_{\mathcal{F}_p}}\right)^{\top} \frac{\partial \mathcal{B}_{1,p}}{\partial\, \boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}} \\
&+ L_{22}(\mathcal{B}_{1,p-1}, \mathcal{B}_{1,p}) \left(\frac{\partial \mathcal{B}_{1,p}}{\partial\, \boldsymbol{v}^{(p-1)}\big|_{\mathcal{F}_p}}\right)^{\top} \frac{\partial \mathcal{B}_{1,p}}{\partial\, \boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}} \\
&+ L_{2}(\mathcal{B}_{1,p-1}, \mathcal{B}_{1,p}) \frac{\partial^2 \mathcal{B}_{1,p}}{\partial\, \boldsymbol{v}^{(p-1)}\big|_{\mathcal{F}_p}\, \partial\, \boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}}
\end{aligned} \tag{S1}$$

where L_2 means the partial derivative of L with regard to its first value, while L_{12} and L_{22} are the second-order partial derivatives of L. Note that the last line of (S1) equals 0 because that each pairwise interaction in the RaFM will not contain vectors from different FMs. Therefore,

$$\frac{\partial^2 L(\mathcal{B}_{1,p-1}, \mathcal{B}_{1,p})}{\partial\, \boldsymbol{v}^{(p-1)}\big|_{\mathcal{F}_p}\, \partial\, \boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}} = \boldsymbol{H} \frac{\partial \mathcal{B}_{1,p}}{\boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}}$$

where $\boldsymbol{H}$ is a $(|\mathcal{F}_p|\, D_{p-1}) \times 1$ matrix:

$$\boldsymbol{H} = L_{12}(\mathcal{B}_{1,p-1}, \mathcal{B}_{1,p}) \left(\frac{\partial \mathcal{B}_{1,p-1}}{\boldsymbol{v}^{(p-1)}\big|_{\mathcal{F}_p}}\right)^{\top} + L_{22}(\mathcal{B}_{1,p-1}, \mathcal{B}_{1,p}) \left(\frac{\partial \mathcal{B}_{1,p}}{\boldsymbol{v}^{(p-1)}\big|_{\mathcal{F}_p}}\right)^{\top}$$

Then we have

$$\begin{aligned}
\widehat{grad} =&L_1(\mathcal{B}_{1,m}, y) \frac{\partial \mathcal{B}_{1,m}}{\partial\, \boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}} \\
&+ \frac{1}{N}\sum_{\boldsymbol{x'}} L_1(\mathcal{B}'_{1,m}, y) \frac{\partial \mathcal{B}'_{1,m}}{\partial\, \boldsymbol{v}^{(p-1)}\big|_{\mathcal{F}_p}} \boldsymbol{G}^{-1} \boldsymbol{H} \frac{\partial \mathcal{B}_{1,p}}{\boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}} \\
=&L_1(\mathcal{B}_{1,m}, y) \frac{\partial \mathcal{B}_{1,m}}{\partial\, \boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}} + \lambda \frac{\partial \mathcal{B}_{1,p}}{\boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}}
\end{aligned}$$

where

$$\lambda = \frac{1}{N}\sum_{\boldsymbol{x'}} L_1(\mathcal{B}'_{1,m}, y) \frac{\partial \mathcal{B}'_{1,m}}{\partial\, \boldsymbol{v}^{(p-1)}\big|_{\mathcal{F}_p}} \boldsymbol{G}^{-1} \boldsymbol{H}$$

Note that λ is the multiplication of a $1 \times |\mathcal{F}_p|\, D_{p-1}$ matrix, a $|\mathcal{F}_p|\, D_{p-1} \times |\mathcal{F}_p|\, D_{p-1}$ matrix, and a $|\mathcal{F}_p|\, D_{p-1} \times 1$ matrix, and thus is a scalar. Moreover, the derivative of $\mathcal{B}_{1,m}$ and $\mathcal{B}_{1,p}$ with respect to $\boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}$ is the same. Therefore, the direction of $\widehat{grad}$ is parallel to that of $L_1(\mathcal{B}_{1,m}, y) \left(\partial \mathcal{B}_{1,m} / \partial\, \boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}\right)$, i.e. $\partial L(\mathcal{B}_{1,m}, y) / \partial\, \boldsymbol{v}^{(p)}\big|_{\mathcal{F}_p - \mathcal{F}_{p+1}}$. $\square$

2. Performance Bound of the Learning Algorithm

Theorem S1. *Assume there exist two nonnegative functions $d(\cdot)$ and $\Delta(\cdot,\cdot)$ such that d is monotonically increasing, and for all $f_1(\cdot), f_2(\cdot)$ we have*

$$d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(f_1(\boldsymbol{x}), y)\right) \leq d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(f_2(\boldsymbol{x}), y)\right) + d\left(\frac{1}{N}\sum_{\boldsymbol{x}} \Delta(f_1(\boldsymbol{x}), f_2(\boldsymbol{x}))\right) \tag{S2}$$

then regarding the training error of the k-th FM model, i.e. $\mathcal{B}_{k,k}$, we have

$$d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{k,k}, y)\right) \leq d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{1,m}, y)\right) + \sum_{p=k}^{m-1} d\left(\frac{1}{N}\sum_{\boldsymbol{x}} \Delta(\mathcal{B}^*_{1,p}, \mathcal{B}^*_{1,p+1})\right) \tag{S3}$$

where $\mathcal{B}^*_{1,m}$ is the optimal $\mathcal{B}_{1,m}$, and $\mathcal{B}^*_{1,p}$ is defined in the same way.

Remark: In practice, Δ represents the error of expressing f_2 by f_1. Eq. (S2) is an extension to the triangle inequality, and can be applied to both regression tasks and classification tasks. In regression tasks, L is the square loss, then we can set d as the square root function and $\Delta = L$. In classification tasks, L is the logarithm loss, then we can let d be an identity function, and define Δ as

$$\Delta(f_1(\boldsymbol{x}), f_2(\boldsymbol{x})) = C_{\theta,\delta}\mathcal{D}_{KL}\left[f_1(\boldsymbol{x})\|f_2(\boldsymbol{x})\right] + \log\delta \tag{S4}$$

where $\delta > 1$, $C_{\theta,\delta} = \frac{\log\delta}{\theta\log\delta + (1-\theta)\log\frac{1-\theta}{1-\theta/\delta}}$, and $\theta = \min_{\boldsymbol{x}}\left[yf_2(\boldsymbol{x}) + (1-y)(1-f_2(\boldsymbol{x}))\right]$. $\mathcal{D}_{KL}$ is the KL divergence of two bimonial variables. The readers can refer to Section 3 in the supplementary material for the proof of Eq. (S2) for logarithm loss.

In order to prove Theorem S1, we first provide the following lemma:

Lemma S2. *The following inequalities hold*

$$d\left(\frac{1}{N}\sum_{\boldsymbol{x}} l(\mathcal{B}^*_{l,k}, y)\right) \leq d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{l-1,k}, y)\right) \tag{S5}$$

Proof. Note that we have $\mathcal{B}_{l,k} = \mathcal{B}_{l-1,k}$ provided that

$$\boldsymbol{v}_i^{(l)} = \begin{bmatrix} \mathbf{0}_{D_l - D_{l-1}} \\ \boldsymbol{v}_i^{(l-1)} \end{bmatrix}, \forall i \in \mathcal{F}_l$$

And such solution also satisfies the constraint (12) in the main body of the paper. Therefore $\mathcal{B}_{l-1,k}$ is a submodel of $\mathcal{B}_{l,k}$, and thus the optimal training error of $\mathcal{B}_{l,k}$ is smaller than $\mathcal{B}_{l-1,k}$, and (S5) follows. □

Proof of Theorem S1. According to (S5) we have

$$\begin{aligned} d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{k,k}, y)\right) &\leq d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{k-1,k}, y)\right) \leq d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{k-2,k}, y)\right) \\ &\cdots \\ &\leq d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{1,k}, y)\right) \end{aligned}$$

Moreover, according to (S2) we have

$$
\begin{aligned}
d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{1,k}, y)\right) &\le d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{1,k+1}, y)\right) + d\left(\frac{1}{N}\sum_{\boldsymbol{x}} \Delta(\mathcal{B}^*_{1,k}, \mathcal{B}^*_{1,k+1})\right) \\
&\le d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{1,k+2}, y)\right) + \sum_{p=k}^{k+1} d\left(\frac{1}{N}\sum_{\boldsymbol{x}} \Delta(\mathcal{B}^*_{1,p}, \mathcal{B}^*_{1,p+1})\right) \\
&\dots \\
&\le d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{1,m}, y)\right) + \sum_{p=k}^{m-1} d\left(\frac{1}{N}\sum_{\boldsymbol{x}} \Delta(\mathcal{B}^*_{1,p}, \mathcal{B}^*_{1,p+1})\right)
\end{aligned}
$$

Therefore we have

$$
d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{k,k}, y)\right) \le d\left(\frac{1}{N}\sum_{\boldsymbol{x}} L(\mathcal{B}^*_{1,m}, y)\right) + \sum_{p=k}^{m-1} d\left(\frac{1}{N}\sum_{\boldsymbol{x}} \Delta(\mathcal{B}^*_{1,p}, \mathcal{B}^*_{1,p+1})\right)
$$

□

3. Quasi-Triangle Inequality for Logarithmic Loss

The following proposition is an extension of the triangle inequality for log loss.

Proposition S3. *Suppose $y \in \{0,1\}, 0 < \hat{y}_1, \hat{y}_2 < 1$, and define the log loss function $L(\hat{y}_i, y)$ and the KL divergence $\mathcal{D}_{KL}(\hat{y}_1 \| \hat{y}_2)$ as*

$$
\begin{aligned}
L(\hat{y}_i, y) &= -y \log \hat{y}_i - (1-y)\log(1-\hat{y}_i) \\
\mathcal{D}_{KL}(\hat{y}_1 \| \hat{y}_2) &= \hat{y}_1 \log \frac{\hat{y}_1}{\hat{y}_2} + (1-\hat{y}_1)\log\frac{1-\hat{y}_1}{1-\hat{y}_2}
\end{aligned}
$$

then $\forall \delta > 1, 0 < \theta \le y\hat{y}_1 + (1-y)(1-\hat{y}_1)$, we have

$$
L(\hat{y}_2, y) \le L(\hat{y}_1, y) + C_{\theta,\delta}\mathcal{D}_{KL}(\hat{y}_1 \| \hat{y}_2) + \log\delta \tag{S6}
$$

where

$$
C_{\theta,\delta} = \frac{\log\delta}{\theta\log\delta + (1-\theta)\log\frac{1-\theta}{1-\theta/\delta}} \tag{S7}
$$

Before proving Proposition S3, we first provide some lemmas.

Lemma S4. $\forall\theta > 0, \delta > 1$, *we have $C_{\theta,\delta} > 0$, and $C_{\theta,\delta}$ monotonically decreases when θ increases.*

Proof. Consider $g(\theta) = 1/C_{\theta,\delta} = \theta + \left[(1-\theta)\log\frac{1-\theta}{1-\theta/\delta} / \log\delta\right]$, then we have

$$
\begin{aligned}
g'(\theta)\log\delta &= \log\delta - \log\frac{1-\theta}{1-\theta/\delta} - 1 + \frac{1}{\delta}\frac{1-\theta}{1-\theta/\delta} \\
&= -\log\left(\frac{1}{\delta}\frac{1-\theta}{1-\theta/\delta}\right) + \left(\frac{1}{\delta}\frac{1-\theta}{1-\theta/\delta} - 1\right) > 0
\end{aligned}
$$

here we use the fact that $\log x < (x-1)$ unless $x = 1$. Due to that $\log\delta > 0$, we have $g'(\theta) \ge 0$, thus $g(\theta)$ is an increasing function. Moreover we have

$$
g(\theta) > g(0) = 0
$$

Therefore, $C_{\theta,\delta} = 1/g(\theta)$ is a decreasing function with respect to θ, and $C_{\theta,\delta} > 0$. □

Lemma S5. *For* $0 < \hat{y}_1, \hat{y}_2 < 1$*, we have*

$$\log \frac{\hat{y}_1}{\hat{y}_2} \leq C_{\hat{y}_1,\delta} \mathcal{D}_{KL}(\hat{y}_1 \| \hat{y}_2) + \log \delta \tag{S8}$$

Proof. When $\hat{y}_1 < \delta \hat{y}_2$, we have $\log(\hat{y}_1/\hat{y}_2) \leq \log \delta$, thus (S8) holds due to the nonnegativity of the KL divergence. Now we discuss the case when $\hat{y}_1 \geq \delta \hat{y}_2$. Consider the ratio between $\mathcal{D}_{KL}(\hat{y}_1 \| \hat{y}_2)$ and $\log(\hat{y}_1/\hat{y}_2)$:

$$\frac{\mathcal{D}_{KL}(\hat{y}_1 \| \hat{y}_2)}{\log(\hat{y}_1/\hat{y}_2)} = \hat{y}_1 + (1 - \hat{y}_1) \frac{\log(1 - \hat{y}_1) - \log(1 - \hat{y}_2)}{\log \hat{y}_1 - \log \hat{y}_2} = \hat{y}_1 + (1 - \hat{y}_2) h(\hat{y}_1, \hat{y}_2)$$

where $h(\hat{y}_1, \hat{y}_2) = \log \frac{1-\hat{y}_1}{1-\hat{y}_2} / \log \frac{\hat{y}_1}{\hat{y}_2}$. It is easy to show that

$$\frac{\partial h}{\partial \hat{y}_2} = -\frac{\mathcal{D}_{KL}(\hat{y}_2 \| \hat{y}_1)}{(\log \hat{y}_1 - \log \hat{y}_2)^2 \hat{y}_2 (1 - \hat{y}_2)} \leq 0$$

Therefore according to $\hat{y}_1 \geq \delta \hat{y}_2$, we have $h(\hat{y}_1, \hat{y}_2) \geq h(\hat{y}_1, \hat{y}_1/\delta)$, and

$$\frac{\mathcal{D}_{KL}(\hat{y}_1 \| \hat{y}_2)}{\log(\hat{y}_1/\hat{y}_2)} \geq \hat{y}_1 + (1 - \hat{y}_1) \frac{\log(1 - \hat{y}_1) - \log(1 - \hat{y}_1/\delta)}{\log \hat{y}_1 - \log(\hat{y}_1/\delta)} = \frac{1}{C_{\hat{y}_1,\delta}}$$

Therefore (S8) also holds when $\hat{y}_1 \geq \delta \hat{y}_2$. □

Proof of Proposition S3. We first prove the case when $y = 1$. In this case, we have $\theta \leq \hat{y}_1$, and

$$\begin{aligned} L(\hat{y}_2, y) - L(\hat{y}_1, y) &= \log \frac{\hat{y}_1}{\hat{y}_2} \\ &\leq C_{\hat{y}_1,\delta} \mathcal{D}_{KL}(\hat{y}_1 \| \hat{y}_2) + \log \delta \\ &\leq C_{\theta,\delta} \mathcal{D}_{KL}(\hat{y}_1 \| \hat{y}_2) + \log \delta \end{aligned}$$

where the first and second inequalities are according to Lemmas S5 and S4. For $y = 0$, we can let $y' = 1 - y, \hat{y}_1' = 1 - \hat{y}_1$, and $\hat{y}_2' = 1 - \hat{y}_2$, and use the same discussion for $y', \hat{y}_1'$ and $\hat{y}_2'$. □