"Self-Paced Learning for Matrix Factorization": Supplementary Material

Qian Zhao[1], **Deyu Meng**[1,*], **Lu Jiang**[2], **Qi Xie**[1], **Zongben Xu**[1], **Alexander G. Hauptmann**[2]
[1]School of Mathematics and Statistics, Xi'an Jiaotong University
[2]School of Computer Science, Carnegie Mellon University
timmy.zhaoqian@gmail.com, dymeng@mail.xjtu.edu.cn, lujiang@cs.cmu.edu
xq.liwu@stu.xjtu.edu.cn, zbxu@mail.xjtu.edu.cn, alex@cs.cmu.edu
*Corresponding author

Abstract

In this supplementary material, we give the proof of Theorem 1 in the maintext.

A Lemmas

We first give some useful lemmas before proving the main theorem.

Lemma A.1 (Boucheron, Lugosi, and Bousquet 2004)**.** *Let X be a random variable with $\mathbb{E}[X] = 0$ and $a \leq X \leq b$ with $b > a$. Then for any $s > 0$, the following inequality holds:*

$$\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{s^2(b-a)^2}{8}\right). \tag{1}$$

Lemma A.2. *Let $C = \{c_1, \dots, c_N\}$ be a finite set, $X_1, \dots, X_n$ denote a random sample without replacement from C and $Y_1, \dots, Y_n$ denote a random sample with replacement from C. Then for any $\mathbf{w} = (w_1, \dots, w_n)$ with $w_i > 0$, if the function $f(x)$ is continuous and convex, then the following inequality holds:*

$$\mathbb{E}[f(\sum_{i=1}^n w_i X_i)] \leq \mathbb{E}[f(\sum_{i=1}^n w_i Y_i)]. \tag{2}$$

Proof. Let $g(x_1, \dots, x_n) = f(w_1 x_1 + \cdots + w_n x_n)$. As mentioned in (Hoeffding 1963), we can find a function g^*, which is not uniquely determined, such that

$$\mathbb{E}[g(Y_1, \dots, Y_n)] = \mathbb{E}[g^*(X_1, \dots, X_n)]. \tag{3}$$

Specifically, we can find one of the g^*s, denoted as $\bar{g}$, with the following form:

$$\begin{aligned}
&\bar{g}(x_1, \dots, x_n) \\
&\quad= \sum_{i_1, i_2, \dots, i_n} p_{i_1 i_2 \dots i_n} f(w_1 x_{i_1} + w_2 x_{i_2} + \cdots + w_n x_{i_n}) \\
&\quad= \sum_{i_1, i_2, \dots, i_n} p_{i_1 i_2 \dots i_n} f\left(\sum_{i=1}^n \left(\sum_{k=1}^n \mathbb{I}(i_k = l) w_k\right) x_l\right),
\end{aligned} \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function (equals 1 if the equation within the brackets holds, and 0 otherwise), and the outside sum is taken over $i_k = 1, \dots, n$ for $k = 1, \dots, n$. The coefficients $p_{i_1 i_2 \dots i_n}$s are positive and do not depend on the function f. Let $f(x) = 1$, by (3) and (4), we have

$$\sum_{i_1, i_2, \dots, i_n} p_{i_1 i_2 \dots i_n} = 1. \tag{5}$$

We also have

$$\begin{aligned}
&\mathbb{E}[g(Y_1, \dots, Y_n)] = \mathbb{E}[\bar{g}(X_1, \dots, X_n)] \\
&= \mathbb{E}\left[\sum_{i_1, i_2, \dots, i_n} p_{i_1 i_2 \dots i_n} f\left(\sum_{i=1}^n \left(\sum_{k=1}^n \mathbb{I}(i_k = l) w_k\right) x_l\right)\right] \\
&= \sum_{i_1, i_2, \dots, i_n} p_{i_1 i_2 \dots i_n} p_{i_1 i_2 \dots i_n} \mathbb{E}\left[f\left(\sum_{i=1}^n \left(\sum_{k=1}^n \mathbb{I}(i_k = l) w_k\right) x_l\right)\right].
\end{aligned} \tag{6}$$

Since (5) holds, it suffices to prove (2) by showing that

$$\mathbb{E}\left[f\left(\sum_{i=1}^n w_i X_i\right)\right] \leq \mathbb{E}\left[f\left(\sum_{i=1}^n \left(\sum_{k=1}^n \mathbb{I}(i_k = l) w_k\right) x_l\right)\right] \tag{7}$$

holds for any $k, r_1, \dots, r_k, i_1, \dots, i_k$ satisfying the same condition as in (4).

If $i_k, i_2, \dots, i_n$ are taken pairwise different values from $\{1, 2, \dots, n\}$, then (7) holds by equality. Otherwise, it suffices to show

$$\begin{aligned}
\mathbb{E}\left[f\left(\sum_{i=1}^n w_i X_i\right)\right] &\leq \mathbb{E}\left[f\left((w_1 + w_2) X_1 + \sum_{i=3}^n w_i X_i\right)\right] \\
&= \mathbb{E}\left[f\left((w_1 + w_2) X_2 + \sum_{i=3}^n w_i X_i\right)\right],
\end{aligned} \tag{8}$$

since other cases of (7) can be induced by it. Now we prove

(8). We have

$$
\begin{aligned}
\mathbb{E}\left[f\left(\sum_{i=1}^{n} w_i X_i\right)\right] &= \mathbb{E}\left[f\left(w_1 X_1 + w_2 X_2 \sum_{i=3}^{n} w_i X_i\right)\right] \\
&= \mathbb{E}\Bigg[f\Bigg(\frac{w_1}{w_1+w_2}\bigg((w_1+w_2)X_1 + \sum_{i=3}^{n} w_i X_i\bigg) \\
&\qquad + \frac{w_2}{w_1+w_2}\bigg((w_1+w_2)X_2 + \sum_{i=3}^{n} w_i X_i\bigg)\Bigg)\Bigg] \\
&\le \frac{w_1}{w_1+w_2}\mathbb{E}\bigg[f\bigg((w_1+w_2)X_1 + \sum_{i=3}^{n} w_i X_i\bigg)\bigg] \\
&\quad + \frac{w_2}{w_1+w_2}\mathbb{E}\bigg[f\bigg((w_1+w_2)X_2 + \sum_{i=3}^{n} w_i X_i\bigg)\bigg],
\end{aligned} \tag{9}
$$

where the inequality holds by convexity of f. By symmetry, we have

$$
\begin{aligned}
&\mathbb{E}\left[f\left((w_1+w_2)X_1 + \sum_{i=3}^{n} w_i X_i\right)\right] \\
&\quad = \mathbb{E}\left[f\left((w_1+w_2)X_2 + \sum_{i=3}^{n} w_i X_i\right)\right].
\end{aligned} \tag{10}
$$

Then (8) holds by taking (10) back to (9), which completes the proof.

□

Lemma A.3. *Let $C = \{c_1, \dots, c_N\}$ be a finite set with mean $\mu = \frac{1}{N}\sum_{i=1}^{N} c_i$, $X_1, \dots, X_n$ denote a random sample without replacement from C, $a \triangleq \min_i c_i$, $b \triangleq \max_i c_i$ and $\mathbf{w} = (w_1, \dots, w_n)$ satisfying $\sum_{i=1}^{n} w_i = n$ and $w_i > 0$ for $i = 1, \dots, n$. Then we have:*

$$
\Pr(|\frac{1}{n}\sum_{i=1}^{n} w_i X_i - \mu| \ge t) \le 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n} w_i^2 (b-a)^2}\right) \tag{11}
$$

Proof. We first introduce $Y_1, \dots, Y_n$ as a random sample with replacement from C. It is obvious that Y_is are independent with $\mathbb{E}[Y_i] = \mu$ for $i = 1, \dots, n$. For any $s > 0$, by Markov's inequality, we have

$$
\begin{aligned}
&\Pr(\frac{1}{n}\sum_{i=1}^{n} w_i X_i - \mu \ge t) \\
&\quad = \Pr\left(\exp\left(s(\frac{1}{n}\sum_{i=1}^{n} w_i X_i - \mu)\right) \ge \exp(st)\right) \\
&\quad \le \exp(-st)\mathbb{E}\left[\exp\left(s(\frac{1}{n}\sum_{i=1}^{n} w_i X_i - \mu)\right)\right].
\end{aligned} \tag{12}
$$

Applying Lemma A.2 to $\exp\left(s(\frac{1}{n}\sum_{i=1}^{n} w_i X_i - \mu)\right)$ and $\exp\left(s(\frac{1}{n}\sum_{i=1}^{n} w_i Y_i - \mu)\right)$, we get

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(s(\frac{1}{n}\sum_{i=1}^{n} w_i X_i - \mu)\right)\right] \\
&\quad \le \mathbb{E}\left[\exp\left(s(\frac{1}{n}\sum_{i=1}^{n} w_i Y_i - \mu)\right)\right] \\
&\quad = \mathbb{E}\left[\exp\left(\frac{s}{n}(\sum_{i=1}^{n} w_i (Y_i - \mu))\right)\right] \\
&\quad = \prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\frac{sw_i}{n}(Y_i - \mu)\right)\right] \\
&\quad \le \prod_{i=1}^{n}\exp\left(\frac{s^2 w_i^2 (b-a)^2}{8n^2}\right) \\
&\quad = \exp\left(\frac{s^2\sum_{i=1}^{n} w_i^2 (b-a)^2}{8n^2}\right),
\end{aligned}
$$

where the second equality holds by the independence of Y_is and the second inequality holds by Lemma A.1. Substitute this result to (12), and then we obtain

$$
\begin{aligned}
&\Pr(\frac{1}{n}\sum_{i=1}^{n} w_i X_i - \mu \ge t) \\
&\quad \le \exp(-st)\exp\left(\frac{s^2\sum_{i=1}^{n} w_i^2 (b-a)^2}{8n^2}\right) \\
&\quad \le \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n} w_i^2 (b-a)^2}\right),
\end{aligned}
$$

where the last equality holds by taking $s = \frac{4n^2 t}{\sum_{i=1}^{n} w_i^2 (b-a)^2}$ to minimize the upper bound. Similarly, we can prove

$$
\Pr(\frac{1}{n}\sum_{i=1}^{n} w_i X_i - \mu \le -t) \le \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n} w_i^2 (b-a)^2}\right).
$$

Thus we can conclude

$$
\Pr(|\frac{1}{n}\sum_{i=1}^{n} w_i X_i - \mu| \ge t) \le 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n} w_i^2 (b-a)^2}\right). \tag{13}
$$

□

Lemma A.4 (Wang and Xu 2012)**.** *Let $S_r = \{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2} : \operatorname{rank}(\mathbf{X}) \le r, \|\mathbf{X}\|_F \le K\}$. Then there exists an ϵ-net $\bar{S}_r$ for Frobenius norm obeying*

$$
|\bar{S}_r| \le (9K/\epsilon)^{(n_1+n_2+1)r}.
$$

B Proof of Theorem 1

To prove Theorem 1, we need the following result:

Theorem B.1. *Let $\hat{\mathcal{L}}(\mathbf{X}) = \frac{1}{\sqrt{|\Omega|}}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \hat{\mathbf{Y}})\|_F$ and $\mathcal{L}(\mathbf{X}) = \frac{1}{\sqrt{mn}}\|\mathbf{X} - \hat{\mathbf{Y}}\|_F$. Furthermore, assume $\max_{(i,j)} |x_{ij}| \le b$. Then given matrix $\mathbf{W}$ satisfying*

$$
w_{ij}\begin{cases} > 0, & (i,j) \in \Omega \\ = 0, & \text{otherwise} \end{cases},
$$

$\sum_{(i,j)\in\Omega} w_{ij} = |\Omega|$, and $\sum_{(i,j)\in\Omega} w_{ij}^2 \le 2|\Omega|$, for all rank-r matrices $\mathbf{X}$, with probability greater than $1 - 2\exp(-n)$, there exists a fixed constant C such that

$$\sup_{\mathbf{X}\in S_r} |\hat{\mathcal{L}}(\mathbf{X}) - \mathcal{L}(\mathbf{X})| \le Ck \left(\frac{nr\log(n)}{|\Omega|}\right)^{\frac{1}{4}}.$$

Here, we assume $m \le n$.

Proof. This proof follows the similar way as the proof of Theorem 2 in (Wang and Xu 2012). Fix $\mathbf{X} \in S_r$. Define

$$\hat{u}(\mathbf{X}) = \frac{1}{|\Omega|}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \hat{\mathbf{Y}})\|_F^2 = (\hat{\mathcal{L}}(\mathbf{X}))^2,$$
$$u(\mathbf{X}) = \frac{1}{mn}\|\mathbf{X} - \hat{\mathbf{Y}}\|_F^2 = (\mathcal{L}(\mathbf{X}))^2.$$

Then by Lemma A.3, we have

$$\Pr(|\hat{u}(\mathbf{X}) - u(\mathbf{X})| \ge t) \le 2\exp\left(-\frac{2|\Omega|^2 t^2}{\sum_{(i,j)\in\Omega} w_{ij}^2 M^2}\right) \tag{14}$$

where $M \triangleq \max_{(i,j)}(x_{ij} - \hat{y}_{ij})^2 \le 4b^2$. Applying union bound over all $\mathbf{X} \in \bar{S}_r(\epsilon)$, we have

$$\Pr\left(\sup_{\bar{\mathbf{X}}\in\bar{S}_r(\epsilon)} |\hat{u}(\bar{\mathbf{X}}) - u(\bar{\mathbf{X}})| \ge t\right) \le 2|\bar{S}_r(\epsilon)|\exp\left(-\frac{2|\Omega|^2 t^2}{\sum_{(i,j)\in\Omega} w_{ij}^2 M^2}\right).$$

Equivalently, with probability at least $1-2\exp(-n)$, it holds that

$$\sup_{\bar{\mathbf{X}}\in\bar{S}_r(\epsilon)} |\hat{u}(\bar{\mathbf{X}}) - u(\bar{\mathbf{X}})| \le \left[\frac{M^2}{2}\left(\log|\bar{S}_r(\epsilon)| + n\right)\frac{\sum_{(i,j)\in\Omega} w_{ij}^2}{|\Omega|^2}\right]^{\frac{1}{2}}.$$

Since $\|\bar{\mathbf{X}}\|_F \le \sqrt{mn}b$, by Lemma A.4, we obtain

$$\sup_{\bar{\mathbf{X}}\in\bar{S}_r(\epsilon)} |\hat{u}(\bar{\mathbf{X}}) - u(\bar{\mathbf{X}})| \le \left[\frac{M^2}{2}\left((m+n+1)r\log(9b\sqrt{mn}/\epsilon) + n\right)\frac{\sum_{(i,j)\in\Omega} w_{ij}^2}{|\Omega|^2}\right]^{\frac{1}{2}} := \xi(\Omega, \mathbf{W}).$$

Notice that $\hat{u}(\bar{\mathbf{X}}) = (\hat{\mathcal{L}}(\bar{\mathbf{X}}))^2$ and $u(\bar{\mathbf{X}}) = (\mathcal{L}(\bar{\mathbf{X}}))^2$, and thus we have

$$\sup_{\bar{\mathbf{X}}\in\bar{S}_r(\epsilon)} |\hat{\mathcal{L}}(\mathbf{X}) - \mathcal{L}(\mathbf{X})| \le \sqrt{\xi(\Omega, \mathbf{W})}.$$

For any $\mathbf{X} \in S_r$, there exists $c(\mathbf{X}) \in S_r(\epsilon)$ such that

$$\|\mathbf{X} - c(\mathbf{X})\|_F \le \epsilon, \quad \|\sqrt{\mathbf{W}} \odot P_\Omega(\mathbf{X} - c(\mathbf{X}))\|_F \le (2|\Omega|)^{\frac{1}{4}}\epsilon,$$

where the second inequality holds due to the assumption $\sum_{(i,j)\in\Omega} w_{ij}^2 \le 2|\Omega|$. These two inequalities imply

$$|\mathcal{L}(\mathbf{X}) - \mathcal{L}(c(\mathbf{X}))| = \frac{1}{\sqrt{mn}}\left|\|\mathbf{X} - \bar{\mathbf{Y}}\|_F - \|c(\mathbf{X}) - \bar{\mathbf{Y}}\|_F\right| \le \frac{\epsilon}{\sqrt{mn}},$$

$$\begin{aligned}
&|\hat{\mathcal{L}}(\mathbf{X}) - \hat{\mathcal{L}}(c(\mathbf{X}))| \\
&= \frac{1}{\sqrt{|\Omega|}}\left|\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \bar{\mathbf{Y}})\|_F - \|\sqrt{\mathbf{W}} \odot (c(\mathbf{X}) - \bar{\mathbf{Y}})\|_F\right| \\
&\le \left(\frac{2}{|\Omega|}\right)^{\frac{1}{4}}\epsilon.
\end{aligned}$$

Thus we have

$$\begin{aligned}
&\sup_{\mathbf{X}\in S_r} |\hat{\mathcal{L}}(\mathbf{X}) - \mathcal{L}(\mathbf{X})| \\
&\le \sup_{\mathbf{X}\in S_r}\Big\{|\hat{\mathcal{L}}(\mathbf{X}) - \hat{\mathcal{L}}(c(\mathbf{X}))| + |\mathcal{L}(c(\mathbf{X})) - \mathcal{L}(\mathbf{X})| \\
&\qquad + |\hat{\mathcal{L}}(c(\mathbf{X})) - \mathcal{L}(c(\mathbf{X}))|\Big\} \\
&\le \left(\frac{2}{|\Omega|}\right)^{\frac{1}{4}}\epsilon + \frac{\epsilon}{\sqrt{mn}} + \sup_{\mathbf{X}\in S_r} |\hat{\mathcal{L}}(c(\mathbf{X})) - \mathcal{L}(c(\mathbf{X}))| \\
&\le \left(\frac{2}{|\Omega|}\right)^{\frac{1}{4}}\epsilon + \frac{\epsilon}{\sqrt{mn}} + \sup_{\bar{\mathbf{X}}\in S_r} |\hat{\mathcal{L}}(\bar{\mathbf{X}}) - \mathcal{L}(\bar{\mathbf{X}})| \\
&\le \left(\frac{2}{|\Omega|}\right)^{\frac{1}{4}}\epsilon + \frac{\epsilon}{\sqrt{mn}} + \sqrt{\xi(\Omega, \mathbf{W})}.
\end{aligned}$$

Substitute the expression of $\sqrt{\xi(\Omega, \mathbf{W})}$ into the above inequality and take $\epsilon = 9b$, and then we have

$$\begin{aligned}
&\sup_{\mathbf{X}\in S_r} |\hat{\mathcal{L}}(\mathbf{X}) - \mathcal{L}(\mathbf{X})| \\
&\le 2\left(\frac{2}{|\Omega|}\right)^{\frac{1}{4}}\epsilon + \left(\frac{M^2}{2}\frac{3nr\log(n)\sum_{(i.j)\in\Omega} w_{ij}^2}{|\Omega|^2}\right)^{\frac{1}{4}} \\
&\le 18b\left(\frac{2}{|\Omega|}\right)^{\frac{1}{4}} + 2\sqrt[4]{3}\left(\frac{nr\log(n)}{|\Omega|}\right)^{\frac{1}{4}} \\
&\le Ck\left(\frac{nr\log(n)}{|\Omega|}\right)^{\frac{1}{4}},
\end{aligned}$$

for a constant C. □

Now we can prove Theorem 1 in the maintext.

Theorem B.2 (Theorem 1 in the maintext)**.** *For a given matrix $\mathbf{W}$ which satisfies $w_{ij} \begin{cases} > 0, & (i,j) \in \Omega \\ = 0, & \text{otherwise} \end{cases}$, with $\sum_{(i,j)\in\Omega} w_{ij} = |\Omega|$, and $\sum_{(i,j)\in\Omega} w_{ij}^2 \le 2|\Omega|$, there exists an constant C, such that with probability at least $1 - 2\exp(-n)$,*

$$\text{RMSE} \le \frac{1}{\sqrt{|\Omega|}}\left\|\sqrt{\mathbf{W}} \odot \mathbf{E}\right\|_F + \frac{1}{\sqrt{mn}}\|\mathbf{E}\|_F + Ck\left(\frac{nr\log(n)}{|\Omega|}\right)^{\frac{1}{4}}. \tag{15}$$

Here, we assume $m \le n$ without loss of generality.

Proof.

$$
\begin{aligned}
\mathrm{RMSE} &= \frac{1}{\sqrt{mn}}\|\mathbf{Y}^* - \mathbf{Y}\|_F = \frac{1}{\sqrt{mn}}\|\mathbf{Y}^* - \hat{\mathbf{Y}} + \mathbf{E}\|_F \\
&\leq \frac{1}{\sqrt{mn}}\|\mathbf{Y}^* - \hat{\mathbf{Y}}\|_F + \frac{1}{\sqrt{mn}}\|\mathbf{E}\|_F \\
&\leq \frac{1}{\sqrt{|\Omega|}}\|\sqrt{\mathbf{W}} \odot (\mathbf{Y}^* - \hat{\mathbf{Y}})\|_F + \frac{1}{\sqrt{mn}}\|\mathbf{E}\|_F \\
&\quad + \left| \frac{1}{\sqrt{|\Omega|}}\|\sqrt{\mathbf{W}} \odot (\mathbf{Y}^* - \hat{\mathbf{Y}})\|_F - \frac{1}{\sqrt{mn}}\|\mathbf{Y}^* - \hat{\mathbf{Y}}\|_F \right| \\
&\leq \frac{1}{\sqrt{|\Omega|}}\|\sqrt{\mathbf{W}} \odot (\mathbf{Y} - \hat{\mathbf{Y}})\|_F + \frac{1}{\sqrt{mn}}\|\mathbf{E}\|_F \\
&\quad + \left| \frac{1}{\sqrt{|\Omega|}}\|\sqrt{\mathbf{W}} \odot (\mathbf{Y}^* - \hat{\mathbf{Y}})\|_F - \frac{1}{\sqrt{mn}}\|\mathbf{Y}^* - \hat{\mathbf{Y}}\|_F \right| \\
&\leq \frac{1}{\sqrt{|\Omega|}}\|\sqrt{\mathbf{W}} \odot \mathbf{E}\|_F + \frac{1}{\sqrt{mn}}\|\mathbf{E}\|_F \\
&\quad + \left| \frac{1}{\sqrt{|\Omega|}}\|\sqrt{\mathbf{W}} \odot (\mathbf{Y} - \hat{\mathbf{Y}})\|_F - \frac{1}{\sqrt{mn}}\|\mathbf{Y}^* - \hat{\mathbf{Y}}\|_F \right|.
\end{aligned}
$$

Here, the third inequality holds because $\mathbf{Y}^*$ is the optimal solution of optimization (9) in maintext. Since $\mathbf{Y}^* \in S_r$, applying Theorem B.1 completes the proof. □

References

Boucheron, S.; Lugosi, G.; and Bousquet, O. 2004. Concentration inequalities. In *Advanced Lectures on Machine Learning*. Springer. 208–240.

Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association* 58(301):13–30.

Wang, Y., and Xu, H. 2012. Stability of matrix factorization for collaborative filtering. In *ICML*.