“Self-Paced Learning for Matrix Factorization”:
Supplementary Material

Qian Zhao1, Deyu Meng1,*, Lu Jiang2, Qi Xie1, Zongben Xu1, Alexander G. Hauptmann2
1School of Mathematics and Statistics, Xi’an Jiaotong University
2School of Computer Science, Carnegie Mellon University
timmy.zhaoqian@gmail.com, dymeng@mail.xjtu.edu.cn, lujiang@cs.cmu.edu
xq.liwu@stjtu.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 $E[X] = 0$ and $a \leq X \leq b$ with $b > a$. Then for any $s > 0$, the following inequality holds:

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

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

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

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

$$E[g(Y_1, \ldots, Y_n)] = E[g^*(X_1, \ldots, X_n)].$$

(3)

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

$$\bar{g}(x_1, \ldots, x_n) = \sum_{i_1, i_2, \ldots, i_n} p_{i_1 i_2 \cdots i_n} f(w_1 x_{i_1} + w_2 x_{i_2} + \cdots + w_n x_{i_n})$$

$$= \sum_{i_1, i_2, \ldots, i_n} p_{i_1 i_2 \cdots i_n} \bar{f}(\sum_{i=1}^n (\sum_{k=1}^n I(i_k = l) w_k) x_i),$$

(4)

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

$$\sum_{i_1, i_2, \ldots, i_n} p_{i_1 i_2 \cdots i_n} = 1.$$

(5)

We also have

$$E[g(Y_1, \ldots, Y_n)] = E[\bar{g}(X_1, \ldots, X_n)]$$

$$= E\left[\sum_{i_1, i_2, \ldots, i_n} p_{i_1 i_2 \cdots i_n} f\left(\sum_{i=1}^n \left(\sum_{k=1}^n I(i_k = l) w_k\right) x_i\right)\right]$$

$$= \sum_{i_1, i_2, \ldots, i_n} p_{i_1 i_2 \cdots i_n} \bar{g}\left(\sum_{i=1}^n \left(\sum_{k=1}^n I(i_k = l) w_k\right) x_i\right).$$

(6)

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

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

(7)

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

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

$$E\left[f\left(\sum_{i=1}^n w_i X_i\right)\right] \leq E\left[f\left(w_1 + w_2\right) X_1 + \sum_{i=3}^n w_i X_i\right]$$

$$= E\left[f\left(w_1 + w_2\right) X_2 + \sum_{i=3}^n w_i X_i\right],$$

(8)

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

\[
\mathbb{E} \left[ f \left( \sum_{i=1}^{n} w_i X_i \right) \right] = \mathbb{E} \left[ f \left( w_1 X_1 + w_2 X_2 + \sum_{i=3}^{n} w_i X_i \right) \right]
\]
\[
= \mathbb{E} \left[ f \left( \frac{w_1}{w_1 + w_2} (w_1 + w_2) X_1 + \sum_{i=3}^{n} w_i X_i \right) \right]
\]
\[
+ \frac{w_2}{w_1 + w_2} \mathbb{E} \left[ f \left( (w_1 + w_2) X_2 + \sum_{i=3}^{n} w_i X_i \right) \right]
\]
\[
\leq \frac{w_1}{w_1 + w_2} \mathbb{E} \left[ f \left( (w_1 + w_2) X_1 + \sum_{i=3}^{n} w_i X_i \right) \right]
\]
\[
+ \frac{w_2}{w_1 + w_2} \mathbb{E} \left[ f \left( (w_1 + w_2) X_2 + \sum_{i=3}^{n} w_i X_i \right) \right],
\]

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

\[
\mathbb{E} \left[ f \left( (w_1 + w_2) X_1 + \sum_{i=3}^{n} w_i X_i \right) \right]
\]
\[
= \mathbb{E} \left[ f \left( (w_1 + w_2) X_2 + \sum_{i=3}^{n} w_i X_i \right) \right].
\]

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

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

\[
\Pr \left( \frac{1}{n} \sum_{i=1}^{n} w_i X_i - \mu \geq t \right) \leq 2 \exp \left( -\frac{2n^2 t^2}{\sum_{i=1}^{n} w_i^2 (b-a)^2} \right)
\]
(11)

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

\[
\Pr \left( \frac{1}{n} \sum_{i=1}^{n} w_i X_i - \mu \geq t \right)
\]
\[
= \Pr \left( \exp \left( s \left( \frac{1}{n} \sum_{i=1}^{n} w_i X_i - \mu \right) \right) \geq \exp(st) \right)
\]
\[
\leq \exp(-st) \mathbb{E} \left[ \exp \left( s \left( \frac{1}{n} \sum_{i=1}^{n} w_i X_i - \mu \right) \right) \right].
\]

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

\[
\mathbb{E} \left[ \exp \left( s \left( \frac{1}{n} \sum_{i=1}^{n} w_i X_i - \mu \right) \right) \right]
\]
\[
\leq \mathbb{E} \left[ \exp \left( s \left( \frac{1}{n} \sum_{i=1}^{n} w_i Y_i - \mu \right) \right) \right]
\]
\[
= \mathbb{E} \left[ \exp \left( s \left( \sum_{i=1}^{n} w_i (Y_i - \mu) \right) \right) \right]
\]
\[
= \prod_{i=1}^{n} \mathbb{E} \left[ \exp \left( s w_i (Y_i - \mu) \right) \right]
\]
\[
\leq \prod_{i=1}^{n} \exp \left( \frac{s^2 w_i^2 (b-a)^2}{8n^2} \right)
\]

\[
= \exp \left( \frac{s^2 \sum_{i=1}^{n} w_i^2 (b-a)^2}{8n^2} \right),
\]

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

\[
\Pr \left( \frac{1}{n} \sum_{i=1}^{n} w_i X_i - \mu \geq t \right)
\]
\[
\leq \exp(-st) \exp \left( \frac{s^2 \sum_{i=1}^{n} w_i^2 (b-a)^2}{8n^2} \right)
\]
\[
\leq \exp \left( -\frac{2n^2 t^2}{\sum_{i=1}^{n} w_i^2 (b-a)^2} \right),
\]

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

\[
\Pr \left( \frac{1}{n} \sum_{i=1}^{n} w_i X_i - \mu \leq -t \right) \leq \exp \left( -\frac{2n^2 t^2}{\sum_{i=1}^{n} w_i^2 (b-a)^2} \right),
\]

Thus we can conclude

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

(13)

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

\[
||\hat{S}_r|| \leq (9K/\epsilon)^{(n_1+n_2+1)r}.
\]

**B Proof of Theorem 1**

To prove Theorem 1, we need the following result:

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

\[
\begin{align*}
   w_{ij} &> 0, \quad (i,j) \in \Omega \\
   w_{ij} &= 0, \quad \text{otherwise},
\end{align*}
\]
\[ \sum_{(i,j) \in \Omega} w_{ij} = |\Omega|, \text{ and } \sum_{(i,j) \in \Omega} w_{ij}^2 \leq 2|\Omega|, \text{ for all rank-} r \text{ matrices } X, \text{ with probability greater than } 1 - 2\exp(-n), \text{ there exists a fixed constant } C \text{ such that } \]
\[ \sup_{X \in S_r} \| \hat{L}(X) - L(X) \| \leq Ck \left( \frac{nr \log(n)}{|\Omega|} \right)^{\frac{1}{2}}. \]

Here, we assume \( m \leq n \).

**Proof.** This proof follows the similar way as the proof of Theorem 2 in (Wang and Xu 2012). Fix \( X \in S_r \). Define
\[ \hat{u}(X) = \frac{1}{|\Omega|} \sqrt{\| W \odot (X - \bar{Y}) \|_F^2} = (\hat{L}(X))^2, \]
\[ u(X) = \frac{1}{mn} \| X - \bar{Y} \|_F^2 = (L(X))^2. \]

Then by Lemma A.3, we have
\[ \Pr\left( \sup_{X \in S_r} | \hat{u}(X) - u(X) | \geq t \right) \leq 2 \exp \left( - \frac{2|\Omega|^2 t^2}{\sum_{(i,j) \in \Omega} w_{ij}^2 |\Omega|^2 M^2} \right) \]
where \( M \triangleq \max_{(i,j)} (x_{ij} - \hat{y}_{ij})^2 \leq 4b^2 \). Applying union bound over all \( X \in S_r(\epsilon) \), we have
\[ \Pr\left( \sup_{X \in S_r(\epsilon)} | \hat{u}(X) - u(X) | \geq t \right) \leq 2|S_r(\epsilon)| \exp \left( - \frac{2|\Omega|^2 t^2}{\sum_{(i,j) \in \Omega} w_{ij}^2 |\Omega|^2 M^2} \right). \]
Equivalently, with probability at least \( 1 - 2 \exp(-n) \), it holds that
\[ \sup_{X \in S_r(\epsilon)} | \hat{u}(X) - u(X) | \leq \left[ \frac{M^2}{2} \left( \log |S_r(\epsilon)| + n \right) \sum_{(i,j) \in \Omega} w_{ij}^2 \right]^{\frac{1}{2}}. \]
Since \( \| \bar{X} \|_F \leq \sqrt{mn}b \), by Lemma A.4, we obtain
\[ \sup_{X \in S_r(\epsilon)} | \hat{u}(X) - u(X) | \leq \left[ \frac{M^2}{2} \left( (m + n + 1)r \log(9b\sqrt{mn}/\epsilon) + n \right) \sum_{(i,j) \in \Omega} w_{ij}^2 \right]^{\frac{1}{2}} \]
\[ := \xi(\Omega, W). \]
Notice that \( \hat{u}(X) = (\hat{L}(X))^2 \) and \( u(X) = (L(X))^2 \), and thus we have
\[ \sup_{X \in S_r(\epsilon)} | \hat{L}(X) - L(X) | \leq \sqrt{\xi(\Omega, W)}. \]
For any \( X \in S_r \), there exists \( C(X) \in S_r(\epsilon) \) such that
\[ \| X - c(X) \|_F \leq \epsilon, \quad \| \sqrt{W} \odot P_\Omega (X - c(X)) \|_F \leq (2|\Omega|)^{\frac{1}{2}} \epsilon, \]
where the second inequality holds due to the assumption \( \sum_{(i,j) \in \Omega} w_{ij}^2 \leq 2|\Omega| \). These two inequalities imply
\[ |L(X) - L(c(X))| \leq \frac{1}{\sqrt{mn}} \| X - \bar{Y} \|_F + \| c(X) - \bar{Y} \|_F \]
\[ \leq \frac{\epsilon}{\sqrt{mn}}. \]

\[ |\hat{L}(X) - \hat{L}(c(X))| \]
\[ \leq \frac{1}{\sqrt{|\Omega|}} \left( \| \sqrt{W} \odot (X - \bar{Y}) \|_F - \| \sqrt{W} \odot (c(X) - \bar{Y}) \|_F \right) \]
\[ \leq \left( \frac{2}{|\Omega|} \right)^{\frac{1}{2}} \epsilon. \]
Thus we have
\[ \sup_{X \in S_r} | \hat{L}(X) - L(X) | \]
\[ \leq \sup_{X \in S_r} \left\{ | \hat{L}(X) - \hat{L}(c(X)) | + | \hat{L}(c(X)) - L(c(X)) | \right\} \]
\[ \leq \left( \frac{2}{|\Omega|} \right)^{\frac{1}{2}} \epsilon + \frac{\epsilon}{\sqrt{mn}} + \sup_{X \in S_r} | \hat{L}(c(X)) - L(c(X)) | \]
\[ \leq \left( \frac{2}{|\Omega|} \right)^{\frac{1}{2}} \epsilon + \frac{\epsilon}{\sqrt{mn}} + \sup_{X \in S_r} | \hat{L}(X) - L(X) | \]
\[ \leq \left( \frac{2}{|\Omega|} \right)^{\frac{1}{2}} \epsilon + \epsilon \sqrt{m} \epsilon + \sqrt{\xi(\Omega, W)}. \]
Substitute the expression of \( \sqrt{\xi(\Omega, W)} \) into the above inequality and take \( \epsilon = 9b \), and then we have
\[ \sup_{X \in S_r} | \hat{L}(X) - L(X) | \]
\[ \leq 2 \left( \frac{2}{|\Omega|} \right)^{\frac{1}{2}} \epsilon + \left( \frac{M^2 3nr \log(n) \sum_{(i,j) \in \Omega} w_{ij}^2}{|\Omega|^2} \right)^{\frac{1}{2}} \]
\[ \leq 18b \left( \frac{2}{|\Omega|} \right)^{\frac{1}{2}} + 2 \sqrt{3} \left( \frac{nr \log(n)}{|\Omega|} \right)^{\frac{1}{4}} \]
\[ \leq Ck \left( \frac{nr \log(n)}{|\Omega|} \right)^{\frac{1}{4}}, \]
for a constant \( C \).

\[ \Box \]

Now we can prove Theorem 2 in the maintext.

**Theorem B.2** (Theorem 1 in the maintext). For a given matrix \( W \) which satisfies \( w_{ij} > 0, \ (i, j) \in \Omega \), \( \sum_{(i,j) \in \Omega} w_{ij} = |\Omega| \), \( \sum_{(i,j) \in \Omega} w_{ij}^2 \leq 2|\Omega| \), there exists an constant \( C \), such that with probability at least \( 1 - 2\exp(-n) \),
\[ \text{RMSE} \leq \frac{1}{\sqrt{|\Omega|}} \left( \| \sqrt{W} \odot E \|_F + \frac{1}{\sqrt{mn}} \| E \|_F + Ck \left( \frac{nr \log(n)}{|\Omega|} \right)^{\frac{1}{4}} \right). \]
Here, we assume \( m \leq n \) without loss of generality.
Proof.

\[
\text{RMSE} = \frac{1}{\sqrt{mn}} \| \mathbf{Y}^\ast - \mathbf{Y} \|_F = \frac{1}{\sqrt{mn}} \| \mathbf{Y}^\ast - \hat{\mathbf{Y}} + \mathbf{E} \|_F \\
\leq \frac{1}{\sqrt{mn}} \| \mathbf{Y}^\ast - \hat{\mathbf{Y}} \|_F + \frac{1}{\sqrt{mn}} \| \mathbf{E} \|_F \\
\leq \frac{1}{\sqrt{|\Omega|}} \| \sqrt{\mathbf{W}} \circ (\mathbf{Y}^\ast - \hat{\mathbf{Y}}) \|_F + \frac{1}{\sqrt{mn}} \| \mathbf{E} \|_F \\
+ \left| \frac{1}{\sqrt{|\Omega|}} \| \sqrt{\mathbf{W}} \circ (\mathbf{Y}^\ast - \hat{\mathbf{Y}}) \|_F - \frac{1}{\sqrt{mn}} \| \mathbf{Y}^\ast - \hat{\mathbf{Y}} \|_F \right| \\
\leq \frac{1}{\sqrt{|\Omega|}} \| \sqrt{\mathbf{W}} \circ \mathbf{E} \|_F + \frac{1}{\sqrt{mn}} \| \mathbf{E} \|_F \\
+ \left| \frac{1}{\sqrt{|\Omega|}} \| \sqrt{\mathbf{W}} \circ (\mathbf{Y} - \hat{\mathbf{Y}}) \|_F - \frac{1}{\sqrt{mn}} \| \mathbf{Y}^\ast - \hat{\mathbf{Y}} \|_F \right|. 
\]

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

References

