Expectile Matrix Factorization for Skewed Data Analysis

Rui Zhu[1], Di Niu[1], Linglong Kong[2], and Zongpeng Li[3]

[1] Department of Electrical and Computer Engineering, University of Alberta, {rzhu3, dniu}@ualberta.ca
[2] Department of Mathematical and Statistical Sciences, University of Alberta, lkong@ualberta.ca
[3] Department of Computer Science, University of Calgary, zongpeng@ucalgary.ca

Abstract

Matrix factorization is a popular approach to solving matrix estimation problems based on partial observations. Existing matrix factorization is based on least squares and aims to yield a low-rank matrix to interpret the conditional sample means given the observations. However, in many real applications with skewed and extreme data, least squares cannot explain their central tendency or tail distributions, yielding undesired estimates. In this paper, we propose *expectile matrix factorization* by introducing asymmetric least squares, a key concept in expectile regression analysis, into the matrix factorization framework. We propose an efficient algorithm to solve the new problem based on alternating minimization and quadratic programming. We prove that our algorithm converges to a global optimum and exactly recovers the true underlying low-rank matrices when noise is zero. For synthetic data with skewed noise and a real-world dataset containing web service response times, the proposed scheme achieves lower recovery errors than the existing matrix factorization method based on least squares in a wide range of settings.

Introduction

Matrix estimation has wide applications in many fields such as recommendation systems (Koren, Bell, and Volinsky 2009), network latency estimation (Liao et al. 2013), computer vision (Chen and Suter 2004), system identification (Liu and Vandenberghe 2009), etc. In these problems, a low-rank matrix $M^* \in \mathbb{R}^{m\times n}$ or a linear mapping $\mathcal{A}(M^*)$ from the low-rank matrix M^* is assumed to underlie some possibly noisy observations, where $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$. The objective is to recover the underlying low-rank matrix based on partial observations b_i, $i = 1, \dots, p$. For example, a movie recommendation system aims to recover all user-movie preferences based on the ratings between some user-movie pairs (Koren, Bell, and Volinsky 2009; Su and Khoshgoftaar 2009), or based on implicit feedback, e.g., watching times/frequencies, that are logged for some users on some movies (Hu, Koren, and Volinsky 2008; Rendle et al. 2009). In network or web service latency estimation (Liao et al. 2013; Liu et al. 2015; Zheng, Zhang, and Lyu 2014), given partially collected latency measurements between some nodes that are possibly contaminated by noise, the goal is to recover the underlying low-rank latency matrix, which is present due to network path and function correlations.

Matrix factorization is a popular approach for low-rank matrix estimation, in which the underlying matrix $M^* \in \mathbb{R}^{m\times n}$ is assumed to be $M^* = XY^\mathsf{T}$, with $X \in \mathbb{R}^{m\times k}$ and $Y \in \mathbb{R}^{n\times k}$, such that the rank of M^* is enforced to k. The goal is to find $\hat{M}$ that minimizes the aggregate loss of the estimation $\mathcal{A}(\hat{M})$ on all observed samples b_i, $i = 1, \dots, p$. Matrix factorization problems, although being nonconvex, can be solved efficiently at a large scale by several standard optimization methods such as alternating minimization and stochastic gradient descent. As a result, matrix factorization has gained enormous success in real-world recommender systems, e.g., Netflix Prize competition (Koren, Bell, and Volinsky 2009), and large-scale network latency estimation, e.g., DMFSGD (Liao et al. 2013), due to its scalability, low computation cost per iteration, and the ease of distributed implementation. In contrast, another approach to matrix estimation and completion, namely nuclear-norm minimization (Candès and Tao 2010; Candes and Plan 2010) based on SVT (Cai, Candès, and Shen 2010) or proximal gradient methods (Ma, Goldfarb, and Chen 2011), is relatively less scalable to problems of huge sizes due to high computational cost per iteration (Sun and Luo 2015). Recently, a few studies (Sun and Luo 2015; Jain, Netrapalli, and Sanghavi 2013; Zhao, Wang, and Liu 2015) have also theoretically shown that many optimization algorithms converge to the global optimality of the matrix factorization formulation, and can recover the underlying true low-rank matrix under certain conditions.

Nevertheless, a common limitation of almost all existing studies on matrix estimation is that they have ignored the fact that observations in practice could be highly skewed and do not follow symmetric normal distributions in many applications. For example, latencies to web services over the Internet are highly skewed, in that most measurements are within hundreds of milliseconds while a small portion of outliers could be over several seconds due to network congestion or temporary service unavailability (Zheng, Zhang, and Lyu 2014; Liu et al. 2015). In a video recommender system based on implicit feedback (e.g., user viewing history), the watching time is also highly skewed, in the sense that a user may watch most videos for a short period of time and only finish a few videos that he or she truly likes (Hu, Koren, and Volinsky 2008).

In other words, the majority of existing matrix factorization methods are based on least squares and attempt to produce a low-rank matrix $\hat{M}$ such that $\mathcal{A}(\hat{M})$ estimates the conditional means of observations. However, in the presence of extreme and skewed data, this may incur large biases and may not fulfill practical requirements. For example, in web service latency estimation, we want to find the *most probable* latency between each client-service pair instead of its conditional mean that is biased towards large outliers. Alternatively, one may be interested in finding the tail latencies and exclude the services with long latency tails from being recommended to a client. Similarly, in recommender systems based on implicit feedback, predicting the conditional mean watching time of each user on a video is meaningless due to the skewness of watching times. Instead, we may want to find out the most likely time length that the user might spend on the video, and based the recommendation on that. For asymmetric, skewed and heavy-tailed data that are prevalent in the real world, new matrix factorization techniques need to be developed beyond symmetric least squares, in order to achieve robustness to outliers and to better interpret the central tendency or dispersion of observations.

In this paper, we propose the concept of *expectile matrix factorization (EMF)* by replacing the symmetric least squares loss function in conventional matrix factorization with a loss function similar to those used in expectile regression (Newey and Powell 1987). Our scheme is different from weighted matrix factorization (Singh and Gordon 2008), in that we not only assign different weights to different residuals, but assign each weight *conditioned on whether the residual is positive or negative*. Intuitively speaking, our expectile matrix factorization problem aims to produce a low-rank matrix $\hat{M}$ such that $\mathcal{A}(\hat{M})$ can estimate any ωth conditional expectiles of the observations, not only enhancing the robustness to outliers, but also offering more sophisticated statistical understanding of observations from a matrix beyond mean statistics.

We make multiple contributions in this paper. *First*, we propose an efficient algorithm based on alternating minimization and quadratic programming to solve expectile matrix factorization, which has low complexity similar to that of alternating least squares in conventional matrix factorization. *Second*, we theoretically prove that under certain conditions, expectile matrix factorization retains the desirable properties that without noise, it achieves the global optimality and exactly recovers the true underlying low-rank matrices. This result generalizes the prior result (Zhao, Wang, and Liu 2015) regarding the optimality of alternating minimization for matrix estimation under the symmetric least squares loss (corresponding to $\omega = 0.5$ in EMF) to a general class of "asymmetric least squares" loss functions for any $\omega \in (0, 1)$. The results are obtained by adapting a powerful tool we have developed on the theoretical properties of weighted matrix factorization involving varying weights across iterations. *Third*, for data generated from a low-rank matrix contaminated by skewed noise, we show that our schemes can achieve better approximation to the original low-rank matrix than conventional matrix factorization based on least squares. *Finally*, we also performed extensive evaluation based on a real-world dataset containing web service response times between 339 clients and 5825 web services distributed worldwide. We show that the proposed EMF saliently outperforms the state-of-the-art matrix factorization scheme based on least squares in terms of web service latency recovery from only 5-10% of samples.

Notation: Without specification, any vector $v = (v_1, \ldots, v_p)^\intercal \in \mathbb{R}^p$ is a column vector. We denote its l_p norm as $\|v\|_p = \left(\sum_j v_j^p\right)^{1/p}$. For a matrix $A \in \mathbb{R}^{m\times n}$, we denote A_{ij} as its (i, j)-entry. We denote the singular values of A as $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_k(A)$, where $k = \operatorname{rank}(A)$. Sometimes we also denote $\sigma_{\max}(A)$ as its maximum singular value and $\sigma_{\min}(A)$ as its minimum singular value. We denote $\|A\|_F = \sqrt{\sum_j \sigma_j^2}$ as its Frobenius norm and $\|A\|_2 = \sigma_{\max}(A)$ as its spectral norm. For any two matrices $A, B \in \mathbb{R}^{m\times n}$, we denote their inner product $\langle A, B\rangle = \operatorname{tr}(A^\intercal B) = \sum_{i,j} A_{ij}B_{ij}$. For a bivariate function $f(x, y)$, we denote the partial gradient w.r.t. x as $\nabla_x f(x, y)$ and that w.r.t. y as $\nabla_y f(x, y)$.

Expectile Matrix Factorization

Given a linear mapping $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$, we can get p observations of an $m\times n$ matrix $M^* \in \mathbb{R}^{m\times n}$. In particular, we can decompose the linear mapping $\mathcal{A}$ into p inner products, i.e., $\langle A_i, M^*\rangle$ for $i = 1, \ldots, p$, with $A_i \in \mathbb{R}^{m\times n}$. Denote the p observations by a column vector $b = (b_1, \ldots, b_p)^\intercal \in \mathbb{R}^p$, where b_i is the observation of $\langle A_i, M^*\rangle$ and may contain independent random noise. The matrix estimation problem is to recover the underlying true matrix M^* from observations b, assuming that M^* has a low rank.

Matrix factorization assumes that the matrix M^* has a rank no more than k, and can be factorized into two tall matrices $X \in \mathbb{R}^{m\times k}$ and $Y \in \mathbb{R}^{n\times k}$ with $k \ll \{m, n, p\}$. Specifically, it estimates M^* by solving the following nonconvex optimization problem:

$$\min_{X\in\mathbb{R}^{m\times k}, Y\in\mathbb{R}^{n\times k}} \sum_{i=1}^{p} \mathcal{L}(b_i, \langle A_i, M\rangle) \quad \text{s.t.} \quad M = XY^\intercal,$$

where $\mathcal{L}(\cdot, \cdot)$ is a loss function. We denote the optimal solution to the problem above by $\hat{M}$.

The most common loss function used in matrix factorization is the squared loss $(b_i - \langle A_i, XY^\intercal\rangle)^2$, with which the problem is to minimize the mean squared error (MSE):

$$\min_{X\in\mathbb{R}^{m\times k}, Y\in\mathbb{R}^{n\times k}} \sum_{i=1}^{p} \frac{1}{2}(b_i - \langle A_i, XY^\intercal\rangle)^2. \tag{1}$$

Just like linear regression based on least squares, (1) actually aims to produce an $\hat{M}$ which estimates the conditional mean of M^* given partial observations. For symmetric Gaussian noise, the conditional mean is the most efficient estimator. However, for skewed or heavy-tailed noise, the conditional mean can be far away from the central area where elements of the true M^* are distributed. In these cases, we need to develop new techniques to better characterize the central tendency, dispersion and tail behavior of observations, beyond mean statistics.

Quantile regression (Koenker and Bassett Jr 1978) is a type of regression analysis originated in statistics and econometrics and is more robust against outliers especially in heavy-tailed response measurements. However, in quantile regression, we need to minimize a non-smooth check loss function which is more computationally involving.

Similar to quantile regression, expectile regression (Newey and Powell 1987) is also a regression technique that achieves robustness against outliers, while in the meantime is more computationally efficient than quantile regression by adopting a smooth loss function. In particular, suppose samples $\{(x_i, y_i), i = 1, \ldots, n\}$ are generated from a linear model $y_i = x_i^\mathsf{T}\beta^* + \varepsilon_i$, where $x_i = (1, x_{i1}, \ldots, x_{ip})^\mathsf{T} \in \mathbb{R}^{p+1}$ are predictors and $y_i \in \mathbb{R}$ is the response variable. The expectile regression estimates β^* by solving

$$\underset{\beta}{\text{minimize}} \quad \sum_{i=1}^{n} \rho_\omega^{[2]}(y_i - x_i^\mathsf{T}\beta),$$

where for a chosen constant $\omega \in (0, 1)$, $\rho_\omega^{[2]}(\cdot)$ is the "asymmetric least squares" loss function given by

$$\rho_\omega^{[2]}(t) := t^2 \cdot |\omega - \mathbb{1}(t < 0)|,$$

where $\mathbb{1}(t < 0)$ is the indicator function such that it equals to 1 if $t < 0$ and 0 otherwise.

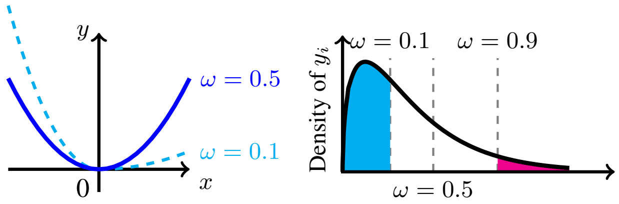
Fig. 1(a) illustrates the shape of $\rho_\omega^{[2]}(\cdot)$. When $\omega < 0.5$, we can see that the cost of a positive residual is lower than that of a negative residual, thus encouraging a smaller estimate $\hat{y}_i$ for the response variable, and vice versa when $\omega > 0.5$. This fact implies that when the response variable y_i are not Gaussian but highly skewed, we can choose an ω to push $\hat{y}_i$ to its most probable area (i.e., the mode or median) while being robust to outliers, as shown in Fig. 1(b).

We now extend expectile regression to the case of matrix estimation. Formally, define $r_i := b_i - \langle A_i, XY^\mathsf{T}\rangle$ as the residual for b_i. Then, in loss minimization, we weight each squared residual r_i^2 by either ω or $1 - \omega$, conditioned on whether it is positive or negative. Therefore, we formulate *expectile matrix factorization* (EMF) as the following problem:

$$\min_{X \in \mathbb{R}^{m \times k}, Y \in \mathbb{R}^{n \times k}} F(X, Y) := \sum_{i=1}^{p} \rho_\omega^{[2]}(b_i - \langle A_i, XY^\mathsf{T}\rangle). \quad (2)$$

Apparently, the MSE-based approach (1) is a special case of problem (2) by setting $\omega = 0.5$, which places equal weights on both positive and negative residuals.

Note that expectile matrix factorization proposed above is different from weighted MSE (Singh and Gordon 2008), where a different yet fixed (predefined) weight is assigned to different residuals. In expectile matrix factorization, each weight is either ω or $1 - \omega$, depending on whether the residual of the estimate is positive or negative, i.e., we do not know the assignments of weights before solving the optimization problem. In other words, problem (2) estimates an $\hat{M}$ such that each $\langle A_i, \hat{M}\rangle$ estimates the ωth *conditional expectile* (Newey and Powell 1987) of b_i. In the meantime, expectiles are based on second-order moments and thus it is feasible to solve EMF efficiently, which we show in the next section.



(a) $\rho_\omega^{[2]}(\cdot)$ (b) Expectile Regression

Figure 1: (a) The asymmetric least squares loss function, placing different weights on positive residuals and negative residuals. (b) For a skewed χ_3^2 distribution, expectile regression with $\omega = 0.1$ generates an estimate closer to the mode than the conditional mean ($\omega = 0.5$) does due to the long tail.

Just like expectile regression, the main attraction of expectile matrix factorization goes beyond robustness to outliers. Being able to estimate any ωth expectile of observations, EMF can characterize different measures of central tendency and statistical dispersion, and is useful to obtain a more comprehensive understanding of data distribution. For example, if we are interested in the tail behavior, we could set $\omega = 0.9$ and if we are interested in the conditional median in a highly skewed dataset, we could set $\omega < 0.5$.

Algorithm and Theoretical Results

We propose an efficient algorithm to solve expectile matrix factorization via a combined use of alternating minimization and quadratic programming, as shown in Algorithm 1, with complexity similar to that of alternating least squares in conventional matrix factorization. To better approach potential optimal solutions, we first sum up all measurement matrices A_i weighted by b_i, and perform Singular Value Decomposition (SVD) to get top k singular values.

Algorithm 1 Alternating minimization for expectile matrix factorization. In this algorithm, we use $\bar{X}$ to highlight that $\bar{X}$ is orthonormal.

```
Input: observations $b = (b_1, \ldots, b_p)^\mathsf{T} \in \mathbb{R}^p$, measurement matrices $A_i \in \mathbb{R}^{m \times n}$, $i = 1, \ldots, p$.
Parameter: Maximum number of iterations $T$
$(\bar{X}^{(0)}, D^{(0)}, \bar{Y}^{(0)}) = \mathtt{SVD}_k(\sum_{i=1}^{p} b_i A_i)$    ▷ Singular Value Decomposition to get top $k$ singular values
for $t = 0$ to $T - 1$ do
    $Y^{(t+0.5)} \leftarrow \arg\min_Y F(\bar{X}^{(t)}, Y)$
    $\bar{Y}^{(t+1)} \leftarrow \mathtt{QR}(Y^{(t+0.5)})$    ▷ QR decomposition
    $X^{(t+0.5)} \leftarrow \arg\min_X F(X, \bar{Y}^{(t+1)})$
    $\bar{X}^{(t+1)} \leftarrow \mathtt{QR}(X^{(t+0.5)})$
end for
Output: $M^{(T)} \leftarrow X^{(T-0.5)} \bar{Y}^{(T)\mathsf{T}}$
```

The QR decompositions in Step 6 and Step 8 are *not necessary* and are only included here to simplify the presentation of theoretical analysis. QR decomposition ensures the orthonormal property: given an orthonormal matrix X (or Y), the objective function $F(X, Y)$ is strongly convex and smooth with respect to Y (or X), as shown in the appendix. However, it has been proved (Jain, Netrapalli, and Sanghavi 2013) that when $\omega = 0.5$, alternating minimization with and without QR decomposition are equivalent. The same conclusion also holds for all ω. Therefore, in performance evaluation, we do not have to and did not apply QR decomposition.

The subproblems in Step 5 and Step 7 can be solved efficiently with standard quadratic program (QP) solvers after some reformulation. We now illustrate such equivalence to QP for Step 5, which minimizes $F(\bar{X}, Y)$ given $\bar{X}$. Let $r_i^+ := \max(r_i, 0)$ denote the positive part of residual r_i, and $r_i^- := -\min(r_i, 0)$ denote the negative part of r_i. We have $r_i = r_i^+ - r_i^-$, and the asymmetric least squares loss can be rewritten as

$$\rho_\omega^{[2]}(r_i) = \omega (r_i^+)^2 + (1-\omega)(r_i^-)^2.$$

Given $\bar{X}$, we have

$$\mathcal{A}(\bar{X}Y^\mathsf{T}) = \{\langle A_i, \bar{X}Y^\mathsf{T}\rangle\}_{i=1}^p = \{\langle A_i^\mathsf{T}\bar{X}, Y\rangle\}_{i=1}^p.$$

Let $r^+ = (r_1^+, \dots, r_p^+)^\mathsf{T}$ and $r^- = (r_1^-, \dots, r_p^-)^\mathsf{T}$. For simplicity, let $\mathcal{A}_1(Y) := \mathcal{A}(\bar{X}Y^\mathsf{T})$. Then, minimizing $F(\bar{X}, Y)$ given $\bar{X}$ in Step 5 is equivalent to the following QP:

$$\begin{aligned} \min_{Y \in \mathbb{R}^{n\times k}, r^+, r^- \in \mathbb{R}_+^p} \quad & \omega \|r^+\|_2^2 + (1-\omega)\|r^-\|_2^2 \\ \text{s.t.} \quad & r^+ - r^- = b - \mathcal{A}_1(Y). \end{aligned} \tag{3}$$

Similarly, Step 7 can be reformulated as a QP as well.

Steps 5 and 7 can be solved even more efficiently in the matrix completion case, which aims at recovering an incomplete low-rank matrix from a few observed entries and is a special case of the matrix estimation problem under discussion, where each b_i is simply an observation of a matrix element (possibly with noise). In matrix completion, we can decompose the above QP in Steps 5 and 7 by updating each row of X (or Y), whose time complexity in practice is similar to conventional alternating least squares, e.g., (Koren, Bell, and Volinsky 2009), which also solve QPs.

We now show that the proposed algorithm for expectile matrix factorization retains the optimality for any $\omega \in (0, 1)$ when observations are noiseless, i.e., the produced $M^{(T)}$ will eventually approach the true low-rank matrix M^* to be recovered. We generalize the recent result (Zhao, Wang, and Liu 2015) of the optimality of alternating minimization for matrix estimation under the symmetric least squares loss function (corresponding to $\omega = 0.5$ in EMF) to a general class of "asymmetric least squares" loss functions with any $\omega \in (0, 1)$.

We assume that the linear mapping $\mathcal{A}$ satisfies the well-known $2k$-RIP condition (Jain, Netrapalli, and Sanghavi 2013):

Assumption 1 ($2k$-RIP)**.** *There exists a constant $\delta_{2k} \in (0, 1)$ such that for any matrix M with rank at most $2k$, the following property holds:*

$$(1-\delta_{2k})\|M\|_F^2 \le \|\mathcal{A}(M)\|_2^2 \le (1+\delta_{2k})\|M\|_F^2.$$

A linear mapping $\mathcal{A}$ satisfying the RIP condition can be obtained in various ways. For example, if each entry of A_i is independently drawn from the sub-Gaussian distribution, then $\mathcal{A}$ satisfies $2k$-RIP property with high probability for $p = \Omega(\delta_{2k}^{-2} kn \log n)$ (Jain, Netrapalli, and Sanghavi 2013).

Clearly, Algorithm 1 involves minimizing a weighted sum of squared losses in the form of

$$\mathcal{F}(X, Y) = \textstyle\sum_{i=1}^p w_i (b_i - \langle A_i, XY^\mathsf{T}\rangle)^2,$$

although the weight w_i depends on the sign of residual r_i and may vary in each iteration. We show that if the weights w_i are confined within a closed interval $[w_-, w_+]$ with constants $w_-, w_+ > 0$, then the alternating minimization algorithm for the weighted sum of squared losses will converge to an optimal point. Without loss of generality, we can assume that $w_- \le 1/2 \le w_+$ and $w_- + w_+ = 1$ by weight normalization.

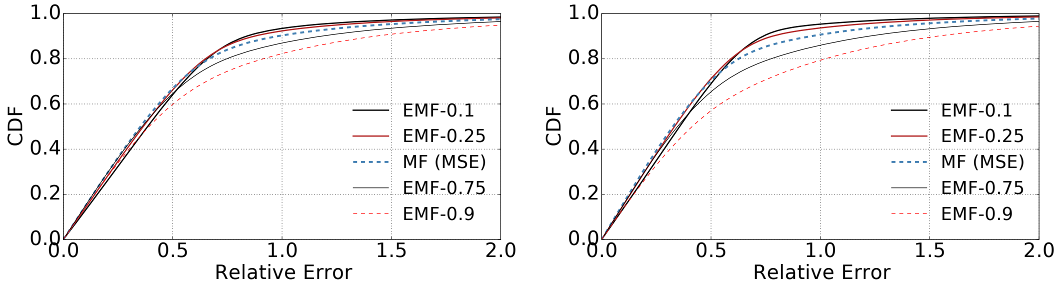
First, we show the geometric convergence of alternating minimization for weighted matrix factorization, if all weights belong to $[w_-, w_+]$ in each iteration:

Theorem 1. *Assume that the linear mapping $\mathcal{A}(\cdot)$ satisfies $2k$-RIP condition with $\delta_{2k} \le C_1/k \cdot w_-^2/w_+^2$ for some small constant C_1, and assume that the singular values of M^* are bounded in the range of $[\Sigma_{\min}, \Sigma_{\max}]$, where $\Sigma_{\min}$ and $\Sigma_{\max}$ are constants and do not scale with the matrix size. Suppose the weights in $\mathcal{F}(X, Y)$ are bounded by two positive finite constants, i.e., $w_i \in [w_-, w_+]$ with $0 < w_- \le 1/2 \le w_+ < 1$ and $w_- + w_+ = 1$. Then, given any desired precision ε, there exists a constant C_2 such that by applying alternating minimization to $\mathcal{F}(X, Y)$, the solution $M^{(T)}$ satisfies $\|M^{(T)} - M^*\|_F \le \varepsilon$ for all $T \ge O(\log(C_2/\varepsilon) + \log(w_-/w_+))$.*

The detailed proof of the above theorem is quite involving and is included in the supplemental material. Theorem 1 implies that the weighted matrix factorization can geometrically converge to a global optimum. Note that the negative term $\log(w_-/w_+)$ does not imply that weighted matrix factorization converges faster, since the value of C_2 for two w's may differ. In fact, due to the lower RIP constant δ_{2k}, the convergence rate in the case of $w_- \ne w_+$ is usually slower than that in the case of $w_- = w_+$.

In Algorithm 1 for expectile matrix factorization, the weight w_i in each iteration for residual r_i is ω if $r_i \ge 0$, and is $1 - \omega$ otherwise. Although w_i is changing across iterations, we can choose $w_- = \min(\omega, 1-\omega)$ and $w_+ = \max(\omega, 1-\omega)$, both satisfying the assumptions in Theorem 1, to bound all w_i. Then we can derive the following main result directly from Theorem 1.

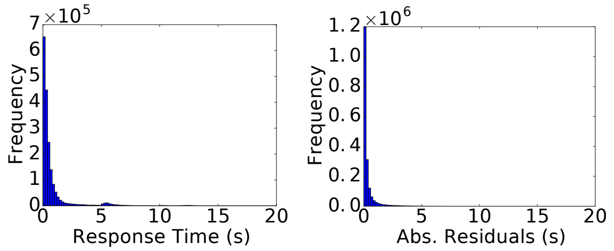
Theorem 2 (Optimality of Algorithm 1)**.** *Suppose $\omega \le 1/2$. Assume that the linear mapping $\mathcal{A}(\cdot)$ satisfies $2k$-RIP condition with $\delta_{2k} \le C_3/k \cdot (1-\omega)^2/\omega^2$ for some small constant C_3, and assume that the singular values of M^* are bounded in the range of $[\Sigma_{\min}, \Sigma_{\max}]$, where $\Sigma_{\min}$ and*



(a) Sampling rate $R = 0.05$

(b) Sampling rate $R = 0.1$

Figure 2: CDF of relative errors via expectile matrix factorization on synthetic 1000×1000 matrices with skewed noise.



(a) Response times

(b) Residuals from MSE-MF

Figure 3: Histograms of a) response times between 5825 web services and 339 service users; b) the residuals of estimates from MSE-based matrix factorization applied on the complete matrix.

$\Sigma_{\max}$ *are constants and do not scale with the matrix size. Then, given any desired precision* ε*, there exists a constant* C_4 *such that Algorithm 1 satisfies* $\|M^{(T)} - M^*\|_F \leq \varepsilon$ *for all* $T \geq O(\log(C_4/\varepsilon) + \log(\omega/(1-\omega)))$*. If* $\omega > 1/2$*, we can get the same result by substituting* ω *with* $1 - \omega$*.*

Additionally, the number of observations needed for exact recovery is $p = \Omega\big(\frac{(1-\omega)^2}{\omega^2} k^3 n \log n\big)$, if the entries of A_i are independently drawn from a sub-Gaussian distribution with zero mean and unit variance, since we require $\delta_{2k} \leq C/k \cdot (1-\omega)^2/\omega^2$. This also matches the sampling complexity of conventional matrix factorization (Jain, Netrapalli, and Sanghavi 2013).

Experiments

In this section, we evaluate the performance of EMF in comparison to the state-of-the-art MSE-based matrix factorization based on both skewed synthetic data and a real-world dataset containing web service response times between 339 users and 5825 web services collected worldwide (Zheng, Zhang, and Lyu 2014). In both tasks, we aim to estimate a true matrix M^* based on partial observations. We define the relative error (RE) as $|M^*_{i,j} - \hat{M}_{i,j}|/M^*_{i,j}$ for all the missing entries (i, j). We use RE to evaluate the prediction accuracy of different methods under a certain sampling rate R (the fraction of known entries).

Experiments on Skewed Synthetic Data

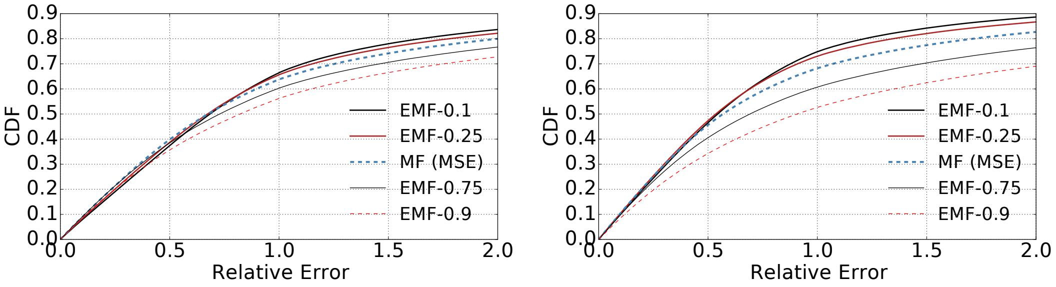
We randomly generate a 1000×1000 matrix $M^* = XY^\top$ of rank $k = 10$, where $X \in \mathbb{R}^{m \times k}$ and $Y \in \mathbb{R}^{n \times k}$ have independent and uniformly distributed entries in $[0, 1]$. Then, we contaminate M^* by a skewed noise matrix $0.5N$, where N contains independent *Chi-square* entries with 3 degrees of freedom. The 0.5 is to make sure the noise does not dominate. We observe some elements in the contaminated matrix and aim to recover the underlying true low-rank M^* under two sampling rates $R = 0.05$ and $R = 0.1$, respectively, where R is the fraction of elements observed. The experiment is repeated for 10 times for each R. We plot the CDF of relative errors in terms of recovering the missing elements of M^* in Fig. 2. We can see that expectile matrix factorization outperforms the conventional MSE-based algorithm (EMF with $\omega = 0.5$) in terms of recovery from skewed noise, with $\omega = 0.1$ yielding the best performance, under both $R = 0.05$ and $R = 0.1$. When more observations are available with $R = 0.1$, EMF with $\omega = 0.1$ demonstrates more benefit as it is more robust to the heavy-tailed noise in data.

Experiments on Web Service Latency Estimation

In these experiments, we aim to recover the web service response times between 339 users and 5825 web services (Zheng, Zhang, and Lyu 2014) distributed worldwide, under different sampling rates.

Fig. 3(a) shows the histogram of all the response times measured between 339 users and 5825 web services. While most entries are less than 1 second, some response times may be as high as 20 seconds due to network delay variations, software glitches and even temporary service outages. The mean latency is 0.91 second, whereas the median is only 0.32 second. This implies that the mean is heavily impacted by the few tail values, while the 0.1-th expectile, which is 0.3 second, is closer to the median of the data. Therefore, if we use the conventional MSE-based matrix factorization to recover this skewed data, the result can be far away from the central area, while EMF with $\omega = 0.1$ may better explain the central tendency.

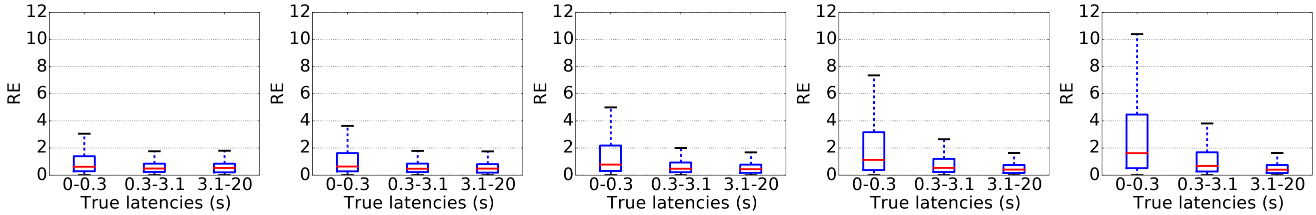
We further performed the MSE-based matrix factorization for the complete response time matrix, which boils down to singular value decomposition (SVD) and we plot the resid-



(a) Sampling rate $R = 0.05$

(b) Sampling rate $R = 0.1$

Figure 4: CDF of relative errors via expectile matrix factorization for web service response time estimation under different sampling rates and ω.



(a) EMF–0.1

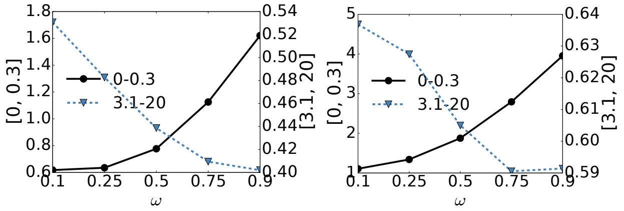
(b) EMF–0.25

(c) MF–MSE (EMF–0.5)

(d) EMF–0.75

(e) EMF–0.9

Figure 5: Box plots of relative errors for different bins of true latencies in the test sets.



(a) Median of REs

(b) IQR of REs

Figure 6: The medians and IQRs of relative errors for different bins as ω varies.

ual histogram in Fig. 3(b). In this figure, 90% of residuals are less than 0.8, while the largest residual can be up to 19.73. Since the residuals are still highly skewed, the conditional means do not serve as good estimates for the most probable data.

In Fig. 4, we plot the relative errors of recovering missing response times with EMF under different ω. Note that EMF-0.5 is essentially the conventional MSE-based matrix factorization. In Fig. 4, we can see that EMF-0.1 performs the best under both sampling rates, and EMF-0.9 performs the worst, because the 0.1-th expectile is the closest to the median, while both the mean and 0.9-th expectile are far away from the central area of data distribution.

To take a closer look at the performance EMF on different segments of data, we divide the testing response times into three bins: 0-0.3s containing 47.5% of all entries, 0.3-3.1s containing 45.4% of all entries, and 3.1-20s containing only 7.1% of all entries. We show the relative errors for testing samples from different bins in box plots in Fig. 5 under different ω. In addition, in Fig. 6, we plot the median of REs and the interquartile range (IQR, the gap between the upper and lower quartiles) of REs when $R = 0.1$, as ω varies for the lower latency bin and the higher latency bin, respectively.

We can observe that EMF with a lower ω achieves higher accuracy in the lower range 0-0.3s, while EMF with a higher ω can predict better in the higher end 3.1-20s. This observation conforms to the intuition illustrated in Fig. 1: an $\omega < 0.5$ penalizes negative residuals, pushing the estimates to be more accurate on the lower end, where most data are centered around. From Fig. 6(a) and Fig. 6(b), we can see that EMF-0.1 predicts the best for the lower range, while EMF-0.9 performs the best for the higher range. However, since most data are distributed in the lower range, EMF-0.1 is better at predicting the central tendency and achieves the best overall accuracy.

Concluding Remarks

In this paper, we propose the expectile matrix factorization approach (EMF) which introduces the "asymmetric least squares" loss function of expectile regression analysis originated in statistics and econometrics into matrix factorization for robust matrix estimation. Existing matrix factorization techniques aim at minimizing the mean squared error and essentially estimate the conditional means of matrix entries. In contrast, the proposed EMF can yield the ωth conditional expectile estimates of matrix entries for any $\omega \in (0, 1)$, accommodating the conventional matrix factorization as a special case of $\omega = 0.5$. We propose an efficient alternating minimization algorithm to solve EMF and theoretically prove its convergence to the global optimality in the noiseless case. Through evaluation based on both synthetic data and a dataset containing real-world web service response times, we show that EMF achieves better recovery than conventional matrix factorization when the data is skewed or contaminated by skewed noise. By using a flexible ω, EMF is not only more robust to outliers but can also be tuned to obtain a more comprehensive understanding of data distribution in a matrix, depending on application requirements.

References

Cai, J.-F.; Candès, E. J.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* 20(4):1956–1982.

Candes, E. J., and Plan, Y. 2010. Matrix completion with noise. *Proc. IEEE* 98(6):925–936.

Candès, E. J., and Tao, T. 2010. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Info. Theory (TIT)* 56(5):2053–2080.

Chen, P., and Suter, D. 2004. Recovering the missing components in a large noisy low-rank matrix: Application to SFM. *IEEE Trans. Pattern Anal. Mach. Intell.* 26(8):1051–1063.

Hu, Y.; Koren, Y.; and Volinsky, C. 2008. Collaborative filtering for implicit feedback datasets. In *Proc. IEEE ICDM*, 263–272. Ieee.

Jain, P.; Meka, R.; and Dhillon, I. S. 2010. Guaranteed rank minimization via singular value projection. In *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 937–945.

Jain, P.; Netrapalli, P.; and Sanghavi, S. 2013. Low-rank matrix completion using alternating minimization. In *Proc. ACM STOC*, 665–674. ACM.

Koenker, R., and Bassett Jr, G. 1978. Regression quantiles. *Econometrica* 33–50.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer* (8):30–37.

Liao, Y.; Du, W.; Geurts, P.; and Leduc, G. 2013. DMF-SGD: A decentralized matrix factorization algorithm for network distance prediction. *IEEE/ACM Trans. Netw. (TON)* 21(5):1511–1524.

Liu, Z., and Vandenberghe, L. 2009. Interior-point method for nuclear norm approximation with application to system identification. *SIAM J. Matrix Anal. Appl.* 31(3):1235–1256.

Liu, B.; Niu, D.; Li, Z.; and Zhao, H. V. 2015. Network latency prediction for personal devices: Distance-feature decomposition from 3D sampling. In *Proc. IEEE INFOCOM*, 307–315. IEEE.

Ma, S.; Goldfarb, D.; and Chen, L. 2011. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming* 128(1-2):321–353.

Newey, W. K., and Powell, J. L. 1987. Asymmetric least squares estimation and testing. *Econometrica* 819–847.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proc. UAI*, 452–461. AUAI Press.

Singh, A. P., and Gordon, G. J. 2008. A unified view of matrix factorization models. In *ECML PKDD*, 358–373. Springer.

Su, X., and Khoshgoftaar, T. M. 2009. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence* 2009:4.

Sun, R., and Luo, Z.-Q. 2015. Guaranteed matrix completion via nonconvex factorization. In *Proc. IEEE FOCS*, 270–289. IEEE.

Zhao, T.; Wang, Z.; and Liu, H. 2015. A nonconvex optimization framework for low rank matrix estimation. In *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 559–567.

Zheng, Z.; Zhang, Y.; and Lyu, M. R. 2014. Investigating QoS of real-world web services. *IEEE Trans. Service Comput.* 7(1):32–39.

Preliminaries

Lemma 1 (Lemma B.1 of (Jain, Netrapalli, and Sanghavi 2013))**.** *Suppose $\mathcal{A}(\cdot)$ satisfies $2k$-RIP. For any $X, U \in \mathbb{R}^{m \times k}$ and $Y, V \in \mathbb{R}^{n \times k}$, we have*

$$|\langle \mathcal{A}(XY^\mathsf{T}), \mathcal{A}(UV^\mathsf{T}) \rangle - \langle X^\mathsf{T}U, Y^\mathsf{T}V \rangle| \leq 3\delta_{2k} \|XY^\mathsf{T}\|_F \cdot \|UV^\mathsf{T}\|_F$$

Lemma 2 (Lemma 2.1 of (Jain, Meka, and Dhillon 2010))**.** *Let $b = \mathcal{A}(M^*) + \varepsilon$, where M^* is a matrix with the rank of k, $\mathcal{A}$ is the linear mapping operator satisfies $2k$-RIP with constant $\delta_{2k} < 1/3$, and ε is a bounded error vector. Let $M^{(t+1)}$ be the $t+1$-th step iteration of SVP, then we have*

$$\|\mathcal{A}(M^{(t+1)}) - b\|_2^2 \leq \|\mathcal{A}(M^*) - b\|_2^2 + 2\delta_{2k}\|\mathcal{A}(M^{(t)}) - b\|_2^2.$$

Lemma 3 (Lemma 4.5 of (Zhao, Wang, and Liu 2015))**.** *Suppose that $Y^{(t+0.5)}$ in Alg. 1 satisfies $\|Y^{(t+0.5)} - V^{(t)}\|_F \leq \sigma_k/4$. Then, there exists a factorization of matrix $M^* = U^{(t+1)}\bar{V}^{(t+1)\mathsf{T}}$ such that $V^{(t+1)} \in \mathbb{R}^{n \times k}$ is an orthonomal matrix, and satisfies*

$$\|\bar{Y}^{(t+1)} - \bar{V}^{(t+1)}\|_F \leq 2/\sigma_k \cdot \|Y^{(t+0.5)} - V^{(t)}\|_F.$$

Proofs

Roadmap

The first step is to prove strongly convexity and smoothness of $\mathcal{F}(X,Y)$ if one variable is fixed by a orthonormal matrix as follows:

Lemma 4. *Suppose that δ_{2k} and $\bar{X}^{(t)}$ satisfy*

$$\delta_{2k} \leq \frac{\sqrt{2}w_-^2(1-\delta_{2k})^2\sigma_k}{24\xi w_+ k(1+\delta_{2k})\sigma_1}. \tag{4}$$

and

$$\|\bar{X}^{(t)} - \bar{U}^{(t)}\|_F \leq \frac{w_-(1-\delta_{2k})\sigma_k}{2\xi w_+(1+\delta_{2k})\sigma_1} \tag{5}$$

Then we have:

$$\|Y^{(t+0.5)} - V^{(t)}\|_F \leq \frac{\sigma_k}{2\xi}\|\bar{X}^{(t)} - \bar{U}^{(t)}\|_F.$$

Clearly, Algorithm 1 involves minimizing a weighted sum of squared losses in the form of $\mathcal{F}(X,Y) = \sum_{i=1}^p w_i(b_i - \langle A_i, XY^\mathsf{T}\rangle)^2$, although the weight w_i depends on the sign of residual r_i and may vary in each iteration. We show that the if the weights w_i are confined in a closed interval $[w_-, w_+]$ with constants $w_-, w_+ > 0$, then the alternating minimization algorithm for the weighted sum of squared losses will converge to the optimal point. Without loss of generality, we can assume that $w_- \leq 1/2 \leq w_+$ and $w_- + w_+ = 1$ by weight normalization. For notation simplicity, we denote a finite positive constant $\xi > 1$ throughout this paper.

Lemma 5. *Suppose the linear operator $\mathcal{A}(\cdot)$ satisfies $2k$-RIP with parameter δ_{2k}. For any orthonormal matrix $\bar{X} \in \mathbb{R}^{m \times k}$, the function $\mathcal{F}(\bar{X}, Y)$ with bounded weights is strongly convex and smooth. In particular, if any weight w_i in $\mathcal{F}(\bar{X}, Y)$ belongs to $[w_-, w_+]$, the value of*

$$\mathcal{F}(\bar{X}, Y') - \mathcal{F}(\bar{X}, Y) - \langle \nabla_Y \mathcal{F}(\bar{X}, Y), Y' - Y \rangle$$

is bounded by

$$[w_-(1-\delta_{2k})\|Y' - Y\|_F^2, w_+(1+\delta_{2k})\|Y' - Y\|_F^2]$$

for all Y, Y'.

Lemma 5 shows that $\mathcal{F}(X,Y)$ can be block-wise strongly convex and smooth if the weights w_i belongs to $[w_-, w_+]$. In the following, we use U and V to denote the optimal factorization of $M^* = UV^\mathsf{T}$. Note that U and V are unique up to orthogonal transformations. The following lemma shows that by taking the block-wise minimum, the distance between the newly updated variable $Y^{(t+0.5)}$ and its "nearby" $V^{(t)}$ is upper bounded by the distance between $X^{(t)}$ and its corresponding neighbor $U^{(t)}$.

Lemma 6. *Suppose that δ_{2k} satisfies*

$$\delta_{2k} \leq \frac{w_-^2(1-\delta_{2k})^2\sigma_k^4}{48\xi^2 k w_+^2(1+\delta_{2k})^2\sigma_1^4}.$$

We have $\|\bar{Y}^{(t+1)} - \bar{V}^{(t+1)}\|_F \leq \frac{1}{\xi}\|\bar{X}^{(t)} - \bar{U}^{(t)}\|_F$.

With the above three lemmas, we can prove Theorem 1 by iteratively upper bounded the distance $\|\bar{Y}^{(t)} - \bar{V}^{(t)}\|_F$ as well as $\|\bar{X}^{(t)} - \bar{U}^{(t)}\|_F$.

Proof of Lemma 5

Now we begin to prove these lemmas. Note that a similar technique has also been used by (Zhao, Wang, and Liu 2015). Since we should fix $X^{(t)}$ or $Y^{(t)}$ as orthonormal matrices, we perform a QR decomposition after getting the minimum. The following lemma shows the distance between $\bar{Y}^{(t+1)}$ and its "nearby" $\bar{V}^{(t+1)}$ is still under control. Due to the page limit, we leave all the proofs in the supplemental material.

Proof. Since $\mathcal{F}(\bar{X}, Y)$ is a quadratic function, we have

$$\begin{aligned}\mathcal{F}(\bar{X}, Y') &= \mathcal{F}(\bar{X}, Y) + \langle \nabla_Y \mathcal{F}(\bar{X}, Y), Y' - Y\rangle \\ &+ \frac{1}{2}(\mathrm{vec}(Y') - \mathrm{vec}(Y))^\mathsf{T} \nabla_Y^2 \mathcal{F}(\bar{X}, Y)(\mathrm{vec}(Y') - \mathrm{vec}(Y)),\end{aligned}$$

and it suffices to bound the singular values of the Hessian matrix $S_\omega := \nabla_Y^2 \mathcal{F}(\bar{X}, Y)$ so that

$$\begin{aligned}\mathcal{F}(\bar{X}, Y') - \mathcal{F}(\bar{X}, Y) - \langle \nabla_Y \mathcal{F}(\bar{X}, Y), Y' - Y\rangle &\leq \frac{\sigma_{\max}(S_\omega)}{2}\|Y' - Y\|_F^2 \\ \mathcal{F}(\bar{X}, Y') - \mathcal{F}(\bar{X}, Y) - \langle \nabla_Y \mathcal{F}(\bar{X}, Y), Y' - Y\rangle &\geq \frac{\sigma_{\min}(S_\omega)}{2}\|Y' - Y\|_F^2.\end{aligned}$$

Now we proceed to derive the Hessian matrix S_ω. Using the fact $\mathrm{vec}(AXB) = (B^\mathsf{T} \otimes A)\mathrm{vec}(X)$, we can write S_ω as follows:

$$\begin{aligned}S_\omega &= \sum_{i=1}^p 2w_i \cdot \mathrm{vec}(A_i^\mathsf{T} \bar{X})\mathrm{vec}^\mathsf{T}(A_i^\mathsf{T} \bar{X}) \\ &= \sum_{i=1}^p 2w_i \cdot (I_k \otimes A_i^\mathsf{T})\mathrm{vec}(\bar{X})\mathrm{vec}^\mathsf{T}(\bar{X})(I_k \otimes A_i).\end{aligned}$$

Consider a matrix $Z \in \mathbb{R}^{n\times k}$ with $\|Z\|_F = 1$, and we denote $z = \mathrm{vec}(Z)$. Then we have

$$\begin{aligned}z^\mathsf{T} S_\omega z &= \sum_{i=1}^p 2w_i \cdot z^\mathsf{T}(I_k \otimes A_i^\mathsf{T})\mathrm{vec}(\bar{X})\mathrm{vec}^\mathsf{T}(\bar{X})(I_k \otimes A_i) \\ &= \sum_{i=1}^p 2w_i \cdot \mathrm{vec}^\mathsf{T}(A_i Z)\mathrm{vec}(\bar{X})\mathrm{vec}^\mathsf{T}(\bar{X})\mathrm{vec}(A_i Z) \\ &= \sum_{i=1}^p 2w_i \cdot \mathrm{tr}^2(\bar{X}^\mathsf{T} A_i Z) = \sum_{i=1}^p 2w_i \cdot \mathrm{tr}^2(A_i^\mathsf{T} \bar{X} Z^\mathsf{T}).\end{aligned}$$

From the $2k$-RIP property of $\mathcal{A}_\rangle$, we have

$$\begin{aligned}z^\mathsf{T} S_\omega z &\leq \sum_{i=1}^p 2w_+ \mathrm{tr}^2(\bar{X}^\mathsf{T} A_i Z) \\ &\leq 2w_+(1 + \delta_{2k})\|\bar{X} Z^\mathsf{T}\|_F \\ &= 2w_+(1 + \delta_{2k})\|Z^\mathsf{T}\|_F = 2w_+(1 + \delta_{2k}).\end{aligned}$$

Similarly, we also have

$$z^\mathsf{T} S_\omega z \geq 2w_-(1 - \delta_{2k}).$$

Therefore, the maximum singular value $\sigma_{\max}$ is upper bounded by $2w_+(1+\delta_{2k})$ and the minimum singular value $\sigma_{\min}$ is lower bounded by $2w_-(1 - \delta_{2k})$, and the Lemma has been proved. □

Proof of Lemma 4

We prove this lemma by introducing a divergence function as follows.

$$\mathcal{D}(Y^{(t+0.5)}, Y^{(t+0.5)}, \bar{X}^{(t)}) = \left\langle \nabla_Y \mathcal{F}(\bar{U}^{(t)}, Y^{(t+0.5)}) - \nabla_Y \mathcal{F}(\bar{X}^{(t)}, Y^{(t+0.5)}), \frac{Y^{(t+0.5)} - V^{(t)}}{\|Y^{(t+0.5)} - V^{(t)}\|_F} \right\rangle.$$

Lemma 7. *Under the same condition in Lemma 4, we have*

$$\mathcal{D}(Y^{(t+0.5)}, Y^{(t+0.5)}, \bar{X}^{(t)}) \leq \frac{3(1 - \delta_{2k})\sigma_k}{2\xi} \cdot \frac{w_+^2}{w_-}\|\bar{X}^{(t)} - \bar{U}^{(t)}\|. \tag{6}$$

Proof of Lemma 7. In this proof we omit the iteration superscript, and Y stands particularly for $Y^{(t+0.5)}$. Since b_i is measured by $\langle A_i, \bar{U}V^\mathsf{T} \rangle$, we have

$$\mathcal{F}(\bar{X}, Y) = \sum_{i=1}^{p} w_i(\langle A_i, \bar{X}Y^\mathsf{T} \rangle - \langle A_i, \bar{U}V^\mathsf{T} \rangle)^2.$$

By taking the partial derivatives on Y we have

$$\begin{aligned} \nabla_Y \mathcal{F}(\bar{X}, Y) &= \sum_{i=1}^{p} 2w_i(\langle A_i, \bar{X}Y^\mathsf{T} \rangle - \langle A_i, \bar{U}V^\mathsf{T} \rangle)A_i^\mathsf{T} X \\ &= \sum_{i=1}^{p} 2w_i(\langle A_i^\mathsf{T} \bar{X}, Y \rangle - \langle A_i^\mathsf{T} \bar{U}, V \rangle)A_i^\mathsf{T} X \end{aligned}$$

Let $x := \mathrm{vec}(\bar{X})$, $y := \mathrm{vec}(Y)$, $u := \mathrm{vec}(\bar{U})$, and $v := \mathrm{vec}(V)$. Since Y minimizes $\mathcal{F}(\bar{X}, \hat{Y})$, we have

$$\begin{aligned} \mathrm{vec}(\nabla_Y \mathcal{F}(\bar{X}, Y)) &= \sum_{i=1}^{p} 2w_i(\langle A_i^\mathsf{T} \bar{X}, Y \rangle - \langle A_i^\mathsf{T} \bar{U}, V \rangle)A_i^\mathsf{T} x \\ &= \sum_{i=1}^{p} 2w_i((\mathrm{vec}(A_i^\mathsf{T} \bar{X}) \cdot \langle A_i^\mathsf{T} \bar{X}, Y \rangle - \mathrm{vec}(A_i^\mathsf{T} \bar{X}) \cdot \langle A_i^\mathsf{T} \bar{X}, Y \rangle)) \\ &= \sum_{i=1}^{p} 2w_i((I_k \otimes A_i^\mathsf{T})xx^\mathsf{T}(I_k \otimes A_i)y - (I_k \otimes A_i^\mathsf{T})xu^\mathsf{T}(I_k \otimes A_i)v) \end{aligned}$$

We denote

$$S_\omega = \sum_{i=1}^{p} 2w_i \cdot (I_k \otimes A_i^\mathsf{T})xx^\mathsf{T}(I_k \otimes A_i),$$

and

$$J_\omega = \sum_{i=1}^{p} 2w_i \cdot (I_k \otimes A_i^\mathsf{T})xu^\mathsf{T}(I_k \otimes A_i),$$

So the equation becomes $S_\omega y - J_\omega v = 0$ and since S_ω is invertible we have $y = (S_\omega)^{-1} J_\omega v$. Meanwhile, we denote

$$G_\omega = \sum_{i=1}^{p} 2w_i \cdot (I_k \otimes A_i^\mathsf{T})uu^\mathsf{T}(I_k \otimes A_i)$$

as the Hessian matrix of $\nabla_Y^2 \mathcal{F}(\bar{U}, Y)$. Then, the partial gradient $\nabla_Y \mathcal{F}(\bar{U}, Y)$ can be written as

$$\begin{aligned} \mathrm{vec}(\nabla_Y \mathcal{F}(\bar{U}, Y)) &= \sum_{i=1}^{p} 2w_i(\langle A_i^\mathsf{T} \bar{U}, Y \rangle - \langle A_i^\mathsf{T} \bar{U}, V \rangle)(I_k \otimes A_i^\mathsf{T})u \\ &= \sum_{i=1}^{p} 2w_i((I_k \otimes A_i^\mathsf{T})uu^\mathsf{T}(I_k \otimes A_i)y - (I_k \otimes A_i^\mathsf{T})uu^\mathsf{T}(I_k \otimes A_i)v) \\ &= G_\omega(y - v) \\ &= G_\omega(S_\omega^{-1} J_\omega - I_{nk})v. \end{aligned}$$

Since we have $\mathrm{vec}(\nabla_Y \mathcal{F}(\bar{X}, Y)) = 0$, the divergence $\mathcal{D} = \langle \nabla_Y(\bar{U}, Y), (Y - V)/\|(\|Y - V) \rangle_F$. So we need to bound $\nabla_Y \mathcal{F}(\bar{U}, Y)$. Let $K := \bar{X}^\mathsf{T} \bar{U} \otimes I_n$. To get the estimate of $S_\omega^{-1} J_\omega - I_{nk}$, we rewrite it as

$$S_\omega^{-1} J_\omega - I_{nk} = K - I_{nk} + S_\omega^{-1}(J_\omega - S_\omega K).$$

We firstly bound the term $(K - I_{nk})v$. Recall $\mathrm{vec}(AXB) = (B^\mathsf{T} \otimes A)\mathrm{vec}(X)$, we have

$$\begin{aligned} (K - I_{nk})v &= ((\bar{X}^\mathsf{T} \bar{U} - I_k) \otimes I_n)v = \mathrm{vec}(V(\bar{U}^\mathsf{T} X - I_k)) \\ \|(K - I_{nk})v\|_2 &= \|V(\bar{U}^\mathsf{T} \bar{X} - I_k)\|_F \leq \sigma_1 \|\bar{U}^\mathsf{T} \bar{X} - I_k\|_F \\ &\leq \sigma_1 \|(\bar{X} - \bar{U})^\mathsf{T}(\bar{X} - \bar{U})\|_F \leq \sigma_1 \|\bar{X} - \bar{U}\|_F^2 \end{aligned}$$

We then bound the term $J_\omega - S_\omega K$. For any two matrices $Z_1, Z_2 \in \mathbb{R}^{n\times k}$, we denote $z_1 := \mathrm{vec}(Z_1)$ and $z_2 := \mathrm{vec}(Z_2)$. Then we have:

$$\begin{aligned}
& z_1^\mathsf{T}(S_\omega K - J_\omega) z_2 \\
= & \sum_{i=1}^{p} 2w_i z_1^\mathsf{T}(I_k \otimes A_i^\mathsf{T}) x\{x^\mathsf{T}(I_k \otimes A_i)(\bar{X}^\mathsf{T}\bar{U} \otimes I_n)) - u^\mathsf{T}(I_k \otimes A_i)\} z_2 \\
= & \sum_{i=1}^{p} 2w_i \langle Z_1, A_i^\mathsf{T}\bar{X}\rangle \cdot (x^\mathsf{T}(\bar{X}^\mathsf{T}\bar{U} \otimes A_i) z_2 - \langle \bar{U}, A_i Z\rangle) \\
= & \sum_{i=1}^{p} 2w_i \langle A_i, \bar{X}Z_1^\mathsf{T}\rangle (\langle A_i, \bar{X}\bar{X}^\mathsf{T} - I_m)\bar{U}Z_2^\mathsf{T}\rangle \\
\leq & \; 2w_+ \langle \mathcal{A}(\bar{X}Z_1^\mathsf{T}), \mathcal{A}((\bar{X}\bar{X}^\mathsf{T} - I_m)\bar{U}Z_2^\mathsf{T})\rangle
\end{aligned}$$

Since $\bar{X}^\mathsf{T}(\bar{X}\bar{X}^\mathsf{T} - I_m)\bar{U} = 0$, by Lemma 1 we have

$$\begin{aligned}
& z_1^\mathsf{T}(S_\omega K - J_\omega) z_2 \\
\leq & \; 2w_+ \cdot 3\delta_{2k} \|\bar{X}Z_1^\mathsf{T}\|_F \|(\bar{X}\bar{X}^\mathsf{T} - I_m)\bar{U}Z_2^\mathsf{T}\|_F \\
\leq & \; 6w_+\delta_{2k} \|Z_1\|_F \sqrt{\|\bar{U}^\mathsf{T}(\bar{X}\bar{X}^\mathsf{T} - I_m)\bar{U}\|_F \|Z_2^\mathsf{T} Z_2\|_F} \\
= & \; 6w_+\delta_{2k} \sqrt{\|\bar{U}^\mathsf{T}(\bar{X}\bar{X}^\mathsf{T} - I_m)\bar{U}\|_F} \\
\leq & \; 6w_+\delta_{2k}\sqrt{2k}\|\bar{X} - \bar{U}\|_F.
\end{aligned}$$

Thus, the spectral norm of this term is upper bounded by $6w_+\delta_{2k}\sqrt{2k}\|\bar{X} - \bar{U}\|_F$ and finally we have

$$\begin{aligned}
\|\mathrm{vec}(\nabla_Y \mathcal{F}(\bar{U}, Y))\|_2 &= \|G_\omega(S_\omega^{-1}J_\omega - I_{nk})v\|_2 \\
&\leq w_+(1+\delta_{2k})(\sigma_1\|\bar{X} - \bar{U}\|_F^2 + \frac{1}{(1-\delta_{2k})w_-}\|S_\omega K - J_\omega\|_2\|V\|_F) \\
&\leq w_+(1+\delta_{2k})(\sigma_1\|\bar{X} - \bar{U}\|_F^2 + \frac{\sigma_1\sqrt{k}}{(1-\delta_{2k})w_-}\|S_\omega K - J_\omega\|_2) \\
&\leq w_+(1+\delta_{2k})\sigma_1(\|\bar{X} - \bar{U}\|_F^2 + \frac{\sqrt{k}\cdot 6w_+\delta_{2k}\sqrt{2k}}{(1-\delta_{2k})w_-}\|\bar{X} - \bar{U}\|_F) \\
&\leq w_+(1+\delta_{2k})\sigma_1(\|\bar{X} - \bar{U}\|_F^2 + \frac{6\sqrt{2}\cdot w_+\delta_{2k}k}{(1-\delta_{2k})w_-}\|\bar{X} - \bar{U}\|_F).
\end{aligned}$$

Under the given condition, we can upper bound $\|\bar{X} - \bar{U}\|$ and δ_{2k} and we go to the final step as follows:

$$\begin{aligned}
\|\mathrm{vec}(\nabla_Y \mathcal{F}(\bar{U}, Y))\|_2 &\leq \frac{(1-\delta_{2k})\sigma_k w_-}{2\xi} + \frac{(1-\delta_{2k})\sigma_k w_-}{2\xi} \\
&= \frac{(1-\delta_{2k})\sigma_k w_-}{\xi}
\end{aligned}$$

Thus, the divergence $\mathcal{D}(Y, Y, \bar{X})$ can be upperbounded by

$$\mathcal{D}(Y, Y, \bar{X}) \leq \|\mathrm{vec}(\nabla_Y \mathcal{F}(\bar{U}, Y))\|_2 \leq \frac{(1-\delta_{2k})\sigma_k w_-}{\xi}\|\bar{X}^{(t)} - \bar{U}^{(t)}\|_F. \tag{7}$$

□

Lemma 8.

$$\|Y^{(t+0.5)} - V^{(t)}\|_F \leq \frac{1}{2w_-(1-\delta_{2k})}\mathcal{D}(Y^{(t+0.5)}, Y^{(t+0.5)}, \bar{X}^{(t)}). \tag{8}$$

Proof of Lemma 8. Here we utilize the strongly convexity of $\mathcal{F}(X, Y)$ given a orthonormal matrix X. By Lemma 5, we have

$$\mathcal{F}(\bar{U}, V) \geq \mathcal{F}(\bar{U}, Y) + \langle \nabla_Y \mathcal{F}(\bar{U}, Y), V - Y\rangle + w_-(1-\delta_{2k})\|V - Y\|_F^2. \tag{9}$$

Since V minimizes the function $\mathcal{F}(\bar{U}, \hat{V})$, we have $\langle \nabla_Y \mathcal{F}(\bar{U}, V), Y - V \rangle \geq 0$ and thus

$$\begin{aligned} \mathcal{F}(\bar{U}, Y) &\geq \mathcal{F}(\bar{U}, V) + \langle \nabla_Y \mathcal{F}(\bar{U}, V), Y - V \rangle + (1 - \delta_{2k}) w_- \|V - Y\|_F^2 \\ &\geq \mathcal{F}(\bar{U}, V) + w_-(1 - \delta_{2k}) \|V - Y\|_F^2. \end{aligned} \tag{10}$$

Add (9) and (10) we have

$$\langle \nabla_Y \mathcal{F}(\bar{U}, Y), Y - V \rangle \geq 2 w_- (1 - \delta_{2k}) \|V - Y\|_F^2. \tag{11}$$

Since Y also minimizes $\mathcal{F}(\bar{X}, \hat{Y})$, we have $\langle \nabla_Y \mathcal{F}(\bar{X}, V), V - Y \rangle \geq 0$ and thus

$$\begin{aligned} \langle \nabla_Y \mathcal{F}(\bar{U}, Y) - \nabla_Y \mathcal{F}(\bar{X}, Y), Y - V \rangle &\geq \langle \nabla_Y \mathcal{F}(\bar{U}, Y), Y - V \rangle \\ &\geq 2 w_- (1 - \delta_{2k}) \|V - Y\|_F^2. \end{aligned} \tag{12}$$

Therefore, we have

$$\|V - Y\|_F \leq \frac{1}{2 w_- (1 - \delta_{2k})} \mathcal{D}(Y, Y, \bar{X}) \tag{13}$$

□

Given Lemma 7 and Lemma 8, we can now bound $\|Y^{(t+0.5)} - V^{(t)}\|_F$ and thus prove Lemma 4.

Proof of Lemma 4. From Lemma 7, we have

$$\mathcal{D}(Y^{(t+0.5)}, Y^{(t+0.5)}, \bar{X}^{(t)}) \leq \frac{(1 - \delta_{2k}) \sigma_k w_-}{\xi} \|\bar{X}^{(t)} - \bar{U}^{(t)}\|_F,$$

and from Lemma 8, we have

$$\|Y^{(t+0.5)} - V^{(t)}\|_F \leq \frac{1}{2 w_- (1 - \delta_{2k})} \mathcal{D}(Y^{(t+0.5)}, Y^{(t+0.5)}, \hat{X}^{(t)}).$$

Therefore,

$$\begin{aligned} & \|Y^{(t+0.5)} - V^{(t)}\|_F && (14) \\ \leq \;& \frac{(1 - \delta_{2k}) \sigma_k w_-}{\xi} \cdot \frac{1}{2 w_- (1 - \delta_{2k})} \|\bar{X}^{(t)} - \bar{U}^{(t)}\|_F && (15) \\ = \;& \frac{\sigma_k}{2\xi} \|\bar{X}^{(t)} - \bar{U}^{(t)}\|_F && (16) \end{aligned}$$

□

Proof of Lemma 6

From Lemma 4, we have

$$\begin{aligned} \|Y^{(0.5)} - V^{(t)}\|_F \;&\leq\; \frac{\sigma_k}{2\xi} \|\bar{X}^{(t)} - \bar{U}_F^{(t)}\| && (17) \\ &\leq\; \frac{(1 - \delta_{2k}) \sigma_k w_-}{2\xi^2 (1 + \delta_{2k}) \sigma_1 w_+} \leq \frac{\sigma_k}{4}, && (18) \end{aligned}$$

where (18) is from $\xi > 1$. Thus, we can see from Lemma3 and we obtain that

$$\|\bar{Y}^{(t+1)} - \bar{V}^{(t+1)}\|_F \leq \frac{2}{\sigma_k} \|Y^{(0.5)} - V^{(t)}\|_F \leq \frac{1}{\xi} \|\bar{X}^{(t)} - \bar{U}^{(t)}\| \leq \frac{(1 - \delta_{2k}) \sigma_k w_-}{2\xi (1 + \delta_{2k}) \sigma_1 w_+}. \tag{19}$$

Proof of Theorem 1

Lemma 9. *Suppose that δ_{2k} satisfies*

$$\delta_{2k} \leq \frac{w_-^2 (1 - \delta_{2k})^2 \sigma_k^4}{48 \xi^2 k w_+^2 (1 + \delta_{2k})^2 \sigma_1^4}.$$

Then there exists a factorization of $M^* = \bar{U}^0 V^{(0)\top}$ such that $\bar{U}^{(0)} \in \mathbb{R}^{m \times k}$ is an orthonormal matrix, and satisfies

$$\|\bar{X}^{(0)} - \bar{U}^{(0)}\|_F \leq \frac{w_- (1 - \delta_{2k}) \sigma_k}{2 \xi w_+ (1 + \delta_{2k}) \sigma_1}.$$

Proof of Lemma 9. The initialization step can be regarded as taking a step iterate of singular value projection (SVP) as taking $M^{(t)} = 0$ and the next iterate with the step size $1/(1+\delta_{2k})$ will result $M^{(t+1)} = \bar{X}^{(0)}D^{(0)}\bar{Y}^{(0)}/(1+\delta_{2k})$, where $\bar{X}^{(0)}$,$D^{(0)}$ and $\bar{Y}^{(0)}$ are from the top k singular value decomposition of $\sum_{i=1}^{p} b_i A_i$.

Then, by Lemma 2 and the fact that $\varepsilon = 0$, we have

$$\left\|\mathcal{A}(\frac{\bar{X}^{(0)}D^{(0)}\bar{Y}^{(0)}}{(1+\delta_{2k})}) - \mathcal{A}(M^*)\right\|_2^2 \leq 4\delta_{2k}\|0 - \mathcal{A}(M^*)\|_2^2. \tag{20}$$

From the $2k$-RIP condition, we have

$$\begin{aligned} \left\|\frac{\bar{X}^{(0)}D^{(0)}\bar{Y}^{(0)}}{(1+\delta_{2k})}\right\| &\leq \frac{1}{1-\delta_{2k}}\left\|\mathcal{A}(\frac{\bar{X}^{(0)}D^{(0)}\bar{Y}^{(0)}}{(1+\delta_{2k})}) - \mathcal{A}(M^*)\right\|_2^2 \\ &\leq \frac{4\delta_{2k}}{1-\delta_{2k}}\|\mathcal{A}(M^*)\|_2^2 \\ &\leq \frac{4\delta_{2k}(1+\delta_{2k})}{1-\delta_{2k}}\|M^*\|_F^2 \leq 6\delta_{2k}\|M^*\|_F^2. \end{aligned}$$

Then, we project each column of M^* into the column subspace of $\bar{X}^{(0)}$ and obtain

$$\|(\bar{X}^{(0)}\bar{X}^{(0)\mathsf{T}} - I)M^*\|_F^2 \leq 6\delta_{2k}\|M^*\|_F^2.$$

We denote the orthonormal complement of $\bar{X}^{(0)}$ as $\bar{X}_{\perp}^{(0)}$. Then, we have

$$\frac{6\delta_{2k}k\sigma_1^2}{\sigma_k^2} \geq \|\bar{X}_{\perp}^{(0)\mathsf{T}}\bar{U}^*\|_F^2,$$

where $\bar{U}^*$ is from the singular value decomposition of $M^* = \bar{U}D\bar{V}\mathsf{T}$. Then, there exists a unitary matrix $O \in \mathbb{R}^{k\times k}$ such that $O^{\mathsf{T}}O = I_k$ and

$$\|\bar{X}^{(0)} - \bar{U}^*O\|_F \leq \sqrt{2}\|\bar{X}_{\perp}^{(0)\mathsf{T}}\bar{U}^*\|_F \leq 2\sqrt{3\delta_{2k}\frac{\sigma_1}{\sigma_k}}.$$

By taking the condition of δ_{2k}, we have

$$\|\bar{X}^0 - \bar{U}^*\|_F \leq \frac{(1-\delta_{2k})\sigma_k w_-}{2\xi(1+\delta_{2k})\sigma_1 w_+}. \tag{21}$$

□

Proof of Theorem 1. The proof of Theorem 1 can be done by induction. Firstly, we note that Lemma 9 ensures that the initial $\bar{X}^{(0)}$ is close to a $\bar{U}^{(0)}$. Then, by Lemma 3 we have the following sequence of inequalities for all T iterations:

$$\|\bar{Y}^{(T)} - \bar{V}^{(T)}\|_F \leq \frac{1}{\xi}\|\bar{X}^{(T-1)} - \bar{U}^{(T-1)}\|_F \leq \cdots \leq \frac{1}{\xi^{2T-1}}\|\bar{X}^{(0)} - \bar{U}^{(0)}\|_F \leq \frac{(1-\delta_{2k})\sigma_k w_-}{2\xi^{2T}(1+\delta_{2k})\sigma_1 w_+}. \tag{22}$$

Therefore, we can bound the right most term by $\varepsilon/2$ for any given precision ε. By algebra, we can derive the required number of iterations T as:

$$T \geq \frac{1}{2}\log\left(\frac{(1-\delta_{2k})\sigma_k w_-}{2\varepsilon(1+\delta_{2k})\sigma_1 w_+}\right)\log^{-1}\xi.$$

Similarly, we can also bound $\|X^{(T-0.5)} - U^{(T)}\|_F$,

$$\|X^{(T-0.5)} - U^{(T)}\|_F \leq \frac{\sigma_k}{2\xi}\|\bar{Y}^{(T)} - \bar{V}^{(T)}\|_F \leq \frac{(1-\delta_{2k})\sigma_k^2 w_-}{4\xi(1+\delta_{2k})\sigma_1 w_+}. \tag{23}$$

To make it smaller than $\varepsilon\sigma_1/2$, we need the number of iterations as

$$T \geq \frac{1}{2}\log\left(\frac{(1-\delta_{2k})\sigma_k^2 w_-}{4\varepsilon(1+\delta_{2k})\sigma_1 w_+}\right)\log^{-1}\xi.$$

Combining all results we have

$$\begin{aligned} \|M^{(T)} - M^*\|_F &= \|X^{(T-0.5)}\bar{Y}^{(T)\mathsf{T}} - U^{(T)}\bar{V}^{(T)\mathsf{T}}\|_F \\ &= \|X^{(T-0.5)}\bar{Y}^{(T)\mathsf{T}} - U^{(T)}\bar{Y}^{(T)\mathsf{T}} + U^{(T)}\bar{Y}^{(T)\mathsf{T}} - U^{(T)}\bar{V}^{(T)\mathsf{T}}\|_F \\ &\leq \|\bar{Y}^{(T)\mathsf{T}}\|_2\|X^{(T-0.5)} - U^{(T)}\|_F + \|U^{(T)}\|_2\|\bar{Y}^{(T)} - \bar{V}^{(T)}\|_F \leq \varepsilon. \end{aligned} \tag{24}$$

Here we use the fact that the orthonormal matrix $\bar{V}^{(T)}$ leads to $\|\bar{V}^{(T)}\|_2 = 1$, and $\|M^*\|_2 = \|U^{(T)}\bar{V}^{(T)\mathsf{T}}\|_2 = \|U^{(T)}\|_2 = \sigma_1$. Now we complete the proof of Theorem 1. □