Preconditioning Matters: Fast Global Convergence of Non-convex Matrix Factorization via Scaled Gradient Descent

Xixi Jia\(^1\) Hailin Wang\(^2\) Jiangjun Peng\(^2\) Xiangchu Feng\(^1\)\(^*\) Deyu Meng\(^2,3\)

\(^1\)School of Mathematics and Statistics, Xidian University
\(^2\)School of Mathematics and Statistics, Xi’an Jiaotong University
\(^3\)Macao Institute of Systems Engineering, Macau University of Science and Technology

\{hsijiaxidian, andrew.pengjj\}@gmail.com; wanghai970163.com
xcfeng@mail.xidian.edu.cn; dymeng@xjtu.edu.cn

Abstract

Low-rank matrix factorization (LRMF) is a canonical problem in non-convex optimization, the objective function to be minimized is non-convex and even non-smooth, which makes the global convergence guarantee of gradient-based algorithm quite challenging. Recent work made a breakthrough on proving that standard gradient descent converges to the \(\varepsilon\)-global minima after \(O(d^3 \kappa^2 \tau^2 \ln \frac{d}{\varepsilon})\) iterations from small initialization with a very small learning rate (both are related to the small constant \(\tau\)). While the dependence of the convergence on the condition number \(\kappa\) and small learning rate makes it not practical especially for ill-conditioned LRMF problem.

In this paper, we show that precondition helps in accelerating the convergence and prove that the scaled gradient descent (ScaledGD) and its variant, alternating scaled gradient descent (AltScaledGD) converge to an \(\varepsilon\)-global minima after \(O(\ln \frac{d}{\delta} + \ln \frac{d}{\varepsilon})\) iterations from general random initialization. Meanwhile, for small initialization as in gradient descent, both ScaledGD and AltScaledGD converge to \(\varepsilon\)-global minima after only \(O(\ln \frac{d}{\varepsilon})\) iterations. Furthermore, we prove that as a proximity to the alternating minimization, AltScaledGD converges faster than ScaledGD, its global convergence does not rely on small learning rate and small initialization, which certifies the advantages of AltScaledGD in LRMF.

1 Introduction

Low-rank matrix factorization aims to approximate a given rank \(d\) matrix \(M \in \mathbb{R}^{m \times n}\) by the product of two factor matrices \(U \in \mathbb{R}^{m \times d}, V \in \mathbb{R}^{n \times d}\), which plays a fundamental and essential role in low-rank matrix recovery such as matrix completion \cite{Jain2013, Ge2016, Sun2016}, matrix sensing \cite{Chi2019, Zhao2015, Charisopoulos2021}, robust principal component analysis \cite{Candes2011, Cai2021}, and the theoretical analysis of deep neural network \cite{Du2018}. Meanwhile, low-rank matrix factorization is also viewed as a canonical problem in non-convex optimization as the objective function to be minimized is non-convex and even non-smooth. Mathematically, we are to solve

\[
\min_{U \in \mathbb{R}^{m \times d}, V \in \mathbb{R}^{n \times d}} f(U, V) := \frac{1}{2} \|UV^T - M\|_F^2, \tag{1}
\]

\(^*\)Corresponding author.

where \( d \ll \min(m, n) \). Though problem (1) is not difficult to solve, the study of this problem has great significance to the gradient-based algorithm for low-rank matrix recovery [Hou et al. 2020, Li et al. 2019b, Chen et al. 2019, Tong et al. 2021], as it is exactly the population loss of low-rank matrix recovery models. Meanwhile, from the perspective of non-convex optimization, problem (1) is an ideal test bed for the theoretical analysis of the asymptotic convergence of gradient-based algorithm for non-convex optimization.

The theoretical guarantee for the global convergence of gradient-based algorithm for problem (1) is challenging, which is due to the following reasons: 1) the problem is non-convex with respect to the variables \( U \) and \( V \), and there are infinitely many local minima and saddle points. Specifically, if \( U^* \) and \( V^* \) is an optimal solution of problem (1), then \( U^* Q \) and \( V^* Q^{-1} \) is also an optimal solution for any invertible matrix \( Q \); 2) the problem is non-smooth with respect to the variables \( U \) and \( V \) and is not coercive due to \( f(\alpha U \frac{1}{\sqrt{d}} V^\top) = f(UV^\top) \) where the scalar \( \alpha \) can be arbitrarily large or small. In theory, gradient-based algorithm is only able to find critical points, while practically, gradient descent algorithm has been verified to converge to the global minima of problem (1) efficiently.

To close the gap between theory and practice, Li et al. [2019a], Ge et al. [2017], Chi et al. [2019] proved that even-though the loss in Eq. (1) is non-convex its loss landscape has some nice property: all local minima are global optima and all the saddle points are strict saddles. Therefore gradient descent algorithms can be guarantee to converge to the global minima. To help escape the strict saddles, Jin et al. [2017] proposed perturbed gradient descent by adding isotropic noise to the gradient at each iteration, they prove that with high probability, perturbed gradient descent converges to the global minima from random initialization at a linear rate. While as analyzed in Ye and Du [2021] and verified by experiments that the gradient perturbation is not really necessary for problem (1).

In contrast to the perturbed gradient descent, Du et al. [2018] studied the naive gradient descent for problem (1), they exploited the balancedness of the two factors \( \|U\|_F^2 + \|V\|_F^2 \), maintained by gradient flow and proved polynomial convergence rate of gradient descent for problem (1) when \( d = 1 \). Furthermore, Ye and Du [2021] improved the results of Du et al. [2018] to rank \( d \) case, they proved that gradient descent converges to the \( \varepsilon \)-global minima of problem (1) after \( O(\frac{d\kappa^2}{\varepsilon^2} \ln \frac{d\kappa^2}{\varepsilon} + \frac{d\kappa^2}{\varepsilon^2} \ln \frac{d\kappa^2}{\varepsilon}) \) iterations. Ye and Du [2021] divided the convergence process into two stages: warm-up phase which takes \( O(\frac{d\kappa^2}{\varepsilon^2} \ln \frac{d\kappa^2}{\varepsilon}) \) iterations and local convergence phase which takes \( O(\frac{d\kappa^2}{\varepsilon^2} \ln \frac{d\kappa^2}{\varepsilon}) \) iterations. The warm-up phase in Ye and Du [2021] is actually the saddle avoid phase after which the gradient descent escapes all the saddle regions as shown in Fig. 1.

It can be seen from Fig. 1 and proved by Ye and Du [2021] that both the saddle avoid phase and local convergence phase of gradient descent highly rely on the condition number \( \kappa \) of the matrix \( M \). If the condition number \( \kappa \) is large, then gradient descent takes long time to escape the saddle regions and also converges slowly. It is therefore very important to know can we improve the gradient-based algorithm such that the global convergence is independent of the condition number? Besides, the global convergence of both Du et al. [2018] and Ye and Du [2021] require very small learning rate (related to the small constant \( \tau \), which seriously limits the application of gradient descent algorithm for ill-conditioned LRMF problem.

Recently, the scaled gradient descent algorithm (ScaledGD) [Apuroop 2012, Mishra and Sepulchre 2016, Tanner and Wei 2016], has been proved by Tong et al. [2021], Tong [2022] to converge very fast from specialized initialization (spectral initialization) to the global minima of problem (1) and the convergence rate is independent of the condition number (Theorem 5 in Tong et al. [2021]). Yet the convergence result provided by Tong et al. [2021] is only local, whether the scaled gradient algorithm can escape saddle regions efficiently for the non-convex problem (1) is still not clear.

In this paper, we are the first to prove that ScaledGD as well as AltScaledGD converge to the global minima of problem (1) from general random Gaussian initialization, and the convergence rate is independent of the condition number of the matrix \( M \). Moreover, we show that the global convergence

![Figure 1: Illustration of the global convergence of GD and AltScaledGD from random initialization.](image-url)
results of ScaledGD and AltScaledGD do not rely on small initialization, the global convergence of AltScaledGD does not even require a small learning rate, which significantly improves the result of [Ye and Du 2021]. To sum up, the contributions of this paper are three-fold:

1. We provide a very simple proof framework for the convergence of ScaledGD and AltScaledGD from general random initialization. Specifically, we divide the optimization process into three phases: initial phase, saddle avoid phase and linear convergence phase. We prove that the loss decreases at rate $$(1 - \eta)^{\chi_k}$$ in the initial phase, and further decreases linearly as $$(1 - \chi_k)^{\kappa}$$ in the saddle avoid phase and the linear convergence phase, where $\eta$ and $\chi_k$ are independent of the condition number $\kappa$, while $\chi_k$ is monotonically increasing from $\frac{\eta^2}{(2-\eta)^2}$ to $\eta$.

2. We prove that if the scale of the random initialization is smaller than a given constant (small initialization), then the loss decrease linearly as $$(1 - \eta)^k$$ from such small initialization to the global minima for both ScaledGD and AltScaledGD.

3. We show that AltScaledGD is a significant improvement of the ScaledGD in that it converges fast with large learning rate up to 1. While in contrast the learning rate of ScaledGD should be smaller than a constant $c_\eta$ that is much smaller than 1.

The organization of this paper is as follow. In Section 2 we introduce the related works on ScaledGD and AltScaledGD. In Section 3 we present our main results, then we give more detailed theoretical analysis on the proof sketch in Section 4. Finally, we conclude this work in Section 5.

2 Related work

In this section, we introduce the ScaledGD and the AltScaledGD as specified in [Apuroop 2012, Mishra and Sepulchre 2016, Tanner and Wei 2016, Tong et al. 2021]. We show that our work is a significant improvement to these existing works on the global convergence analysis.

2.1 Scaled gradient descent

Different to the gradient descent algorithm which takes the negative gradient direction as the descent direction, scaled gradient descent is designed to accelerate the convergence process by scaled the gradient with a preconditioning matrix. Specifically, the ScaledGD updates the variables $U_k$, $V_k$ as

$$
\begin{align*}
U_{k+1} &= U_k - \eta \nabla U_k f(U_k, V_k) (V_k^T V_k)^{-1} \\
V_{k+1} &= V_k - \eta \nabla V_k f(U_k, V_k) (U_k^T U_k)^{-1}
\end{align*}
$$

where $\eta$ is the learning rate, $U_k^T U_k$ and $V_k^T V_k$ are matrices of $d \times d$, and $d \ll \min(m, n)$ therefore the computation of ScaledGD is comparable to that of gradient descent. The inverse matrices $(V_k^T V_k)^{-1}$ and $(U_k^T U_k)^{-1}$ is the preconditioning for the gradient descent. If we denote $X = (U, V)$, then Eq. (2) corresponds to

$$X_{k+1} = X_k - \eta \nabla f_X(X_k) H_k$$

where

$$H_k = \begin{bmatrix}
(V_k^T V_k)^{-1} & 0 \\
0 & (U_k^T U_k)^{-1}
\end{bmatrix}.$$

[Apuroop 2012, Mishra and Sepulchre 2016] proved that ScaledGD is derived by imposing a new metric on the tangent space of the Riemannian manifold. They verified empirically that ScaledGD converges much faster than gradient descent while there is no rigorous convergence rate analysis. Recently, [Tong et al. 2021] is the first to prove the linear convergence property of Eq. (2) for problem [1], while their proof relies on specialized initialization that $\text{dist}(X_0, X_*) \leq 0.1 \sigma_d(M)$, where $X^* = (U^*, V^*)$ and $U^* V^*\dagger = M$ (Theorem 5 [Tong et al. 2021]). The local convergence guarantee is far from satisfactory to understand the convergence of ScaledGD for the non-convex optimization problem [1].
2.2 Alternating scaled gradient descent

The Gaussian-Seidel version of ScaledGD is the following alternating scaled gradient descent which writes

\[
\begin{align*}
U_{k+1} &= U_k - \eta \nabla U_k f(U_k, V_k)(V_k^T V_k)^{-1} \\
V_{k+1} &= V_k - \eta \nabla V_k f(U_{k+1}, V_k)(U_{k+1}^T U_{k+1})^{-1}
\end{align*}
\]

Eq. (4) was studied as scaled alternating steepest descent algorithm in Tanner and Wei [2016], and it is also closely related to the alternating minimization algorithm for the minimization problem (1) when \( \eta = 1 \), and other matrix recovery problem as [Wen et al. 2012], Jain et al. [2013], Chandrasekher et al. [2022]. Different to alternating minimization, the AltScaledGD presented in Eq. (4) can be broadly used in lots of low-rank matrix recovery problem where alternating minimization is computationally prohibitive, such as matrix completion [Zilber and Nadler 2022], Sun and Luo [2016], matrix sensing [Ma et al. 2021]. Existing works for Eq. (4) only proved convergence to critical point as [Wen et al. 2012], Tanner and Wei [2016], yet global convergence analysis of AltScaledGD is still vague.

In this paper, we provide rigorous proofs for the global convergence of ScaledGD Eq. (2) and AltScaledGD Eq. (4), and we show that both ScaledGD and AltScaledGD converge linearly for random Gaussian initialization after saddle avoid phase. Meanwhile, we show that AltScaledGD is robust to the learning rate \( \eta \) which can be set as large as 1, while large \( \eta \) can seriously deteriorates the convergence property of ScaledGD as illustrated by Fig. 6, which sheds light on the superiority of AltScaledGD Eq. (4) over ScaledGD Eq. (2) on problem (1) as well as more low-rank matrix recovery problem.

3 Main results

In this section, we present our main theorems on the convergence of ScaledGD and AltScaledGD for two different random initialization: general random initialization and small initialization. These two initializations are both random Gaussian initialization with zero mean but different variances. Small initialization is widely used in the convergence analysis of low rank matrix factorization problem Stöger and Soltanolkotabi [2021], Ye and Du [2021], Ma and Fattahi [2022], while small initialization is skin to spectral initialization which does not help us fully understand the global convergence of the non-convex problem. In this paper, we provide both the global convergence analysis of general random initialization and small initialization.

3.1 Global convergence of ScaledGD

If the matrix \( M \) is rank one, i.e., \( d \) in Eq. (1) is 1, then Eq. (2) is exactly gradient descent with adaptive step-size. We show that such specialized gradient descent for \( d \geq 1 \) converges linearly to the global minima after an initial decreasing phase and the convergence rate is independent of the singular value of \( M \).

**Theorem 1** (General random initialization). Let \( U_0 \in \mathbb{R}^{m \times d} \) and \( V_0 \in \mathbb{R}^{n \times d} \) be random Gaussian that follow \( \mathcal{N}(0, \sigma) \) for \( \sigma > c_{\text{init}} \) (\( c_{\text{init}} \) is a positive constant), and \( U_k, V_k \) are updated by Eq. (2).

If \( \eta \leq c_\eta < 1 \) for small constant \( c_\eta \), we have that the objective function of problem (1) decreases linearly after \( T_1 = O \left( \ln \frac{d}{\delta} \right) \) iterations, namely

\[
\|U_{k+T_1} V_{k+T_1}^T - M\|_F \leq \alpha_1 (1 - \chi_{k+T_1})^k \|M\|_F, \forall k \geq 0
\]

where \( \chi_{k+T_1} \) is monotonically increasing from \( \frac{\eta^2}{(2-\eta)^2} \) to \( \eta, \delta \) is a sufficiently small constant, \( \alpha_1 \) is a constant.

The Theorem 1 indicates that the global convergence of ScaledGD can be divided into three phases: the initial phase that lasts \( T_1 \) iterations, the saddle avoid phase in which \( \chi_{k+T_1} \) increases from \( \frac{\eta^2}{1-\eta^2} \) to \( \eta \) and the final linear convergence phase with convergence rate \( 1 - \eta \). While if the scale of the initialization \( U_0 \) and \( V_0 \) are very small (with small \( \sigma \)), then the following theorem shows that the ScaledGD converges linearly without entering the saddle regions.

**Theorem 2** (Small initialization). Let \( U_0 \in \mathbb{R}^{m \times d} \) and \( V_0 \in \mathbb{R}^{n \times d} \) be random Gaussian that follow \( \mathcal{N}(0, \sigma) \), with \( \sigma \leq c_{\text{init}} \) and \( U_k, V_k \) are updated by Eq. (2). If \( \eta \leq c_\eta < 1 \) for small constant \( c_\eta \), we have that the objective function of problem (1) decreases linearly, namely

\[
\|U_k V_k^T - M\|_F \leq \alpha_2 (1 - \eta)^k \|M\|_F
\]

(6)
where \( c_{\text{init}} \) is a small constant and \( \alpha_2 \) is a constant.

### 3.2 Global convergence of AltScaledGD

We now present the main convergence results of AltScaledGD.

**Theorem 3** (General random initialization). Let \( U_0 \in \mathbb{R}^{m \times d} \) and \( V_0 \in \mathbb{R}^{n \times d} \) be random Gaussian that follow \( \mathcal{N}(0, \sigma) \) for \( \sigma > c_{\text{init}} \). \( U_k, V_k \) are updated by Eq. (11). we have that the objective function of problem (11) decreases linearly after \( T_1 = O(\ln \frac{1}{\delta}) \) iterations, namely

\[
\|U_k + T_1 V_k^T - M\|_F \leq \alpha_1 (1 - \chi_{k+T_1})^k \|M\|_F
\]

where \( \chi_{k+T_1} \) is monotonically increasing from \( \frac{\eta^2}{(2-\eta)^2} \) to \( \eta \), \( 0 < \eta \leq 1 \) and \( \alpha_1 \) is a constant.

**Theorem 4** (Small initialization). Let \( U_0 \in \mathbb{R}^{m \times d} \) and \( V_0 \in \mathbb{R}^{n \times d} \) be random Gaussian that follow \( \mathcal{N}(0, \rho) \), with \( \sigma \leq c_{\text{init}} \). \( U_k, V_k \) are updated by Eq. (11) then we have that the objective function of problem (11) decreases linearly, namely

\[
\|U_k V_k^T - M\|_F \leq \alpha_2 (1 - \eta)^k \|M\|_F
\]

where \( 0 < \eta \leq 1 \) is the step size, \( \alpha_2 \) is a constant and \( c_{\text{init}} \) is a small constant.

The convergence results of ScaledGD and AltScaledGD are almost the same with differences in that for ScaledGD the learning \( \eta \) should be smaller than a constant \( c_\eta \) which is much less than 1. The small constant \( c_\eta \) greatly restricts the convergence rate of ScaledGD. While for AltScaledGD the learning \( \eta \) can be as large as 1, which indicates the superiority of AltScaledGD over ScaledGD in convergence as \( \eta \) is crucial to the convergence rate. From the above theorems, it can be easily deduced that both ScaledGD and AltScaledGD converge to an \( \varepsilon \)-global minima after \( O(\ln \frac{1}{\delta} + \ln \frac{1}{\varepsilon}) \) iterations from general random initialization. While for small initialization, these two algorithms only need \( O(\ln \frac{1}{\varepsilon}) \) iterations to converge to an \( \varepsilon \)-global minima. More detailed analysis and proofs are provided in Section 4 the proof sketch part and the supplementary materials.

### 3.3 Why does preconditioning help?

In previous works, preconditioning has been used to improve the condition of the optimization problem [Saad, 2003; Zhang et al., 2021]. In this paper, we analyze how the preconditioning helps improving the convergence by analyzing the effect of condition number \( \kappa \) on the learning rate \( \eta \) as the convergence rate highly depends on the learning rate. For simplicity, we take one step of the AltScaledGD as example to analyze the effect of the preconditioning. Since \( f(U, V) \) is quadratic on \( U \), then it can be verified

\[
f(U_{k+1}, V_k) \leq f(U_k, V_k) + \langle \nabla f(U_k, V_k), \Delta \rangle + \frac{1}{2} \|\Delta\|^2_{A_k}
\]

where \( \Delta = U_{k+1} - U_k \) and \( \| \cdot \|_{A_k} \) is a local norm defined by \( \| \Delta \|_{A_k} = \langle \Delta V_k^T V_k, \Delta \rangle \).

For gradient descent, we take \( \Delta = -\eta \nabla f(U_k, V_k) \), then Eq. (9) becomes

\[
f(U_{k+1}, V_k) \leq f(U_k, V_k) - \eta \|\nabla f(U_k, V_k)\|^2 + \frac{\eta^2 \sigma_2^2(V_k)}{2} \|\nabla f(U_k, V_k)\|^2.
\]

\[
\leq f(U_k, V_k) - 2\eta \sigma_2^2(V_k) f(U_k, V_k) + \eta^2 \sigma_1^4(V_k) f(U_k, V_k)
\]

\[
\leq (1 - 2\eta \sigma_2^2(V_k) - \eta^2 \sigma_1^4(V_k)) f(U_k, V_k)
\]

Similarly, it holds

\[
f(U_{k+1}, V_{k+1}) \leq (1 - 2\eta \sigma_2^2(U_{k+1}) - \eta^2 \sigma_1^4(U_{k+1})) f(U_{k+1}, V_k)
\]

Therefore, we have

\[
f(U_{k+1}, V_{k+1}) \leq (1 - 2\eta \sigma_2^2(U_{k+1}) - \eta^2 \sigma_1^4(U_{k+1})) (1 - 2\eta \sigma_2^2(V_k) - \eta^2 \sigma_1^4(V_k)) f(U_k, V_k)
\]
We show in Fig. 2 that the convergence of ScaledGD and AltScaledGD are independent of the number of the matrix $M$ with different initialization. These results are also true for small initialization as shown in the right subfigure of Fig. 2. These observations certificate that preconditioning in Eq. (2) and Eq. (4) indeed help accelerating the convergence such that the convergence rate is independent of the condition number of the matrix $M$.

To guarantee the linear convergence of gradient descent, it is required that $2\eta\sigma^2(V_k) \geq \eta^2\sigma^4(V_k)$ and $2\eta\sigma^2(U_{k+1}) \geq \eta^2\sigma^4(U_{k+1})$ which implies $\eta \leq \min\left\{\frac{1}{2\sigma^2(V_k)}, \kappa(U_{k+1})^2\right\}$. In contrast, if we take $\Delta = -\eta\nabla f(U_k, V_k)(V_k^T V_k)^{-1}$, then

$$f(U_{k+1}, V_k) \leq f(U_k, V_k) - \eta \sigma_r(V_k) f(U_k, V_k)$$

and similarly

$$f(U_{k+1}, V_{k+1}) \leq f(U_{k+1}, V_k) - \eta \sigma_r(U_{k+1}) f(U_{k+1}, V_k)$$

Thus

$$f(U_{k+1}, V_{k+1}) \leq \left(1 - \eta \sigma_r(V_k) f(U_k, V_k)\right) \left(1 - \eta \sigma_r(U_{k+1}) f(U_k, V_k)\right)$$

To guarantee the linear convergence, we only need $0 < \eta < \frac{\sigma_r(V_k)}{2}$, $\sigma_r(U_{k+1})$ as well as $\sigma_r(V_k)$ is strictly larger than 0 ($V_k$ is the orthogonal row subspace of $U_k V_k^T - M$ and $V_k$ is the orthogonal subspace of $V_k$), which indicates that the linear convergence rate is independent of the condition number of matrix $M$.

We show in Fig. 2 that the convergence of ScaledGD and AltScaledGD are independent of the condition number $\kappa$ of the matrix $M$ with general random initialization and small initialization. In Fig. 3 we set the rank of the matrix $M$ as 5, with condition number $\kappa$ ranging from 10, 100, 200. It can be seen from the left subfigure of Fig. 2 that for general random initialization, the error curves of ScaledGD with different $\kappa$ are the exactly the same, and the error curves of ScaledGD also coincide with that of the AltScaledGD. These results are also true for small initialization as shown in the right subfigure of Fig. 2. These observations certificate that preconditioning in Eq. 2 and Eq. 4 indeed help accelerating the convergence such that the convergence rate is independent of the condition number of the matrix $M$.

\[\text{Since } U_k V_k^T \rightarrow M, k \rightarrow \infty \text{ and the same analysis is applied on Eq. 10 with respect to } V, \text{ we know that } \eta \leq \text{cmax.}\]
4 Theoretical analysis – proof sketch

In this section, we provide the proof sketch of our results in Section 3. For simplicity, we present the theoretical analysis for rank one matrix factorization where \( U \in \mathbb{R}^{m \times 1} \) and \( V \in \mathbb{R}^{n \times 1} \). More detailed proofs for the main theorems are provided in the supplementary material.

4.1 Convergence of ScaledGD

It can be deduced that the objective function of problem (1) is upper bounded by four terms as

\[
\begin{align*}
\|U_{k+1}V_{k+1}^\top - M\|_F & \leq (1 - \eta)^2 \|U_kV_k^\top - M\|_F + \eta (1 - \eta) \|M\|_F \|V_{k+1}^\top V_k\|_2 \tag{1} \\
+ \eta (1 - \eta) \|M\|_F \|U_{k+1}^\top U_k\|_2 \tag{2} + \eta^2 \|M\|_F \left(1 - \frac{\|M\|_F \cos \theta_k \cos \theta_k^*}{\|U_kV_k^\top\|_F}\right) \tag{3} \\
\end{align*}
\]

where \( U_k \) and \( V_k \) correspond to the orthogonal basis of the column space of \( U_k \) and \( V_k \), \( U_{k+1} \) and \( V_{k+1} \) are the orthogonal complements of the left and right singular vector matrices of \( M \) (i.e. \( U_0, V_0 \)), and \( \cos \theta_k \) is cosine value of the angle between the vectors \( U_k \) and \( V_k \). The upper-bound depicts the differences between \( U_{k+1}V_{k+1} \) and \( M \) in two aspects:

- The angle between the subspace of \( U_kV_k^\top \) and \( M \): \( \|U_{k+1}^\top U_k\|_2, \|V_{k+1}^\top V_k\|_2 \);
- The difference of the length (norm) between \( U_kV_k^\top \) and \( M \): \( \|M\|_F - \|U_kV_k^\top\|_F \).

The term (2) and (3) are related to the angle between the subspace of \( M \) and \( U_kV_k^\top \), the term (4) is related to the difference between the norm of \( M \) and \( U_kV_k^\top \), as given by the following lemma.

**Lemma 1.** If \( \langle M, U_kV_k^\top \rangle \geq \|U_kV_k^\top\|_F^2 \), then there is constant \( C_u \geq 0 \) such that

\[
\left|1 - \frac{\|M\|_F \cos \theta_k \cos \theta_k^*}{\|U_kV_k^\top\|_F}\right| \leq C_u \left(\|M\|_F - \|U_kV_k^\top\|_F\right) \tag{17}
\]

According to Eq. (16), we know that the decrease of the objective function in problem (1) is decided by the decrease of the distance between the subspace (2 and 3) and the difference between the norm of \( M \) and \( U_kV_k^\top \) (4). The following lemma further reveals that the distance between the subspace of \( U_kV_k^\top \) and \( M \) decreases.

**Lemma 2.** (Convergence of the distance between subspaces) For the ScaledGD (2), if \( \|M\|_F \cos \theta_k \cos \theta_k^* \geq \|U_kV_k^\top\|_F \), then the following holds

\[
\|U_{k+1}^\top U_k\|_2 \leq (1 - \eta)\|U_{k+1}^\top U_k\|_2, \quad \|V_{k+1}^\top V_k\|_2 \leq (1 - \eta)\|V_{k+1}^\top V_k\|_2 \tag{18}
\]

The Lemma 2 indicates that the term (2) and (3) in Eq. (16) decrease linearly if the norm of \( \|U_kV_k^\top\|_F \) is smaller than norm of the projection of \( M \) onto the column and row spaces of \( U_kV_k^\top \). At the same time, the condition \( \|M\|_F \cos \theta_k \cos \theta_k^* \geq \|U_kV_k^\top\|_F \) also guarantees the linear convergence of the differences between the norm of \( U_kV_k^\top \) and \( M \).

**Theorem 5.** (Convergence of the matrix norm) For the ScaledGD (2), if \( \|M\|_F \cos \theta_k \cos \theta_k^* \geq \|U_kV_k^\top\|_F \) for all \( k \geq 0 \), then we have

\[
\|M\|_F - \|U_{k+1}V_{k+1}^\top\|_F \leq (1 - \eta)^2 k C_\alpha \tag{19}
\]

where \( C_\alpha \) is a constant and \( \eta \) is the step length \( 0 \leq \eta < 1 \).

Both Lemma 2 and Theorem 5 are built on the condition that \( \|M\|_F \cos \theta_k \cos \theta_k^* \geq \|U_kV_k^\top\|_F \) for all \( k \geq 0 \), while this is not a trivial condition for ScaledGD. The following lemma guarantees that the condition can be satisfied if the step length \( \eta \) is smaller than a constant.

**Lemma 3.** Let \( \eta \leq \eta < 1 \) with \( c_\eta \) a small constant, if \( \|M\|_F \cos \theta_k \cos \theta_k^* \geq \|U_0V_0^\top\|_F \) then the following is true

\[
\|M\|_F \cos \theta_k \cos \theta_k^* \geq \|U_kV_k^\top\|_F, \forall k > 0. \tag{20}
\]
The above results guarantee the local linear convergence of the term 2, 3 and 4 on the condition that
\[ \| M \|_F \cos \theta^v_0 \cos \theta^v_0 \geq \| U_0 V_0^T \|_F \] (21)
which is critical for our analysis on random Gaussian initialization and small initialization.

4.1.1 Small initialization

In practice, the condition \( \| M \|_F \cos \theta^v_0 \cos \theta^v_0 \geq \| U_0 V_0^T \|_F \) can be easily satisfied by very small (near zero) initialization. According to the random matrix theory Theorem 2.7.5 in Tao [2012], for Gaussian initialization there exists \( \nu > 0 \) such that with high probability \( \cos \theta^v_0 \) and \( \cos \theta^v_0 \) is lower bounded by constant \( 1/\nu \), therefore one can simply set the norm of \( U_0 \) and \( V_0 \) to be sufficiently small such that the inequality (21) holds. In consequence small initialization can guarantee the global linear convergence of ScaledGD, as shown in Fig. 3. While small initialization is very special, it can not help us fully understand the global convergence property of ScaledGD from arbitrary initialization for the non-convex objective (1), even though small initialization has been widely used in the global convergence analysis of gradient descent algorithms Stöger and Soltanolkotabi [2021], Ye and Du [2021], Ma and Fattahi [2022] and ScaledGD for symmetric low rank matrix recovery problems Xu et al. [2023], Zhang et al. [2021].

4.1.2 General random initialization

In order to understand the optimization path of ScaledGD for the non-convex objective (1), we present the theoretical analysis of ScaledGD from random Gaussian initialization that may not satisfy the condition in Eq. (21). As shown in Fig. 3, when initialized with \( \sigma = 1 \) ScaledGD iterations are also attracted by the saddle point thus enter the saddle region (zoomed region marked by red rectangle), while it can escape saddle region very fast. To rigorously characterize the saddle avoid phase, we first show and prove that the norm of matrices \( U_k \) and \( V_k \) decrease if \( \| M \|_F \max\{\cos \theta_k^v, \cos \theta_k^v\} < \| U_k V_k^T \|_F \) as given by the following lemma and shown in Fig. 4.

**Lemma 4.** If the condition \( \| M \|_F \max\{\cos \theta_k^v, \cos \theta_k^v\} < \| U_k V_k^T \|_F \) is satisfied then we have
\[ \| U_{k+1} \|_F < \| U_k \|_F \quad \text{and} \quad \| V_{k+1} \|_F < \| V_k \|_F. \] (22)

Furthermore, if the condition \( \| M \|_F \cos \theta_k^v \cos \theta_k^v \geq \| U_k V_k^T \|_F \) is satisfied then we have
\[ \| U_{k+1} \|_F \geq \| U_k \|_F \quad \text{and} \quad \| V_{k+1} \|_F \geq \| V_k \|_F. \] (23)

In general, if we initialize the matrices \( U \) and \( V \) as \( N(0, \sigma) \) with large \( \sigma \), then in the initial phase, with high probability we have \( \| M \|_F \max\{\cos \theta_0^v, \cos \theta_0^v\} < \| U_0 V_0^T \|_F \). According to Lemma 3 and Lemma 4, we know that the norm of the matrices \( U_k \) and \( V_k \) decreases with the increase of \( k \) until it reaches the condition \( \| M \|_F \cos \theta_k^v \cos \theta_k^v \geq \| U_k V_k^T \|_F \), which is also illustrated in Fig. 4. In Fig. 4, we plot the changes of the norm of matrices \( U \) and \( V \), the nested subfigure illustrates the decrease in the matrix norm in log scale. It is very interesting to study the changes of the matrix norm with respect to the optimization path. Generally, if \( \| U_0 \|_F \) and \( \| V_0 \|_F \) is initialized very large, then the decrease of the norm will decrease the objective function (1). Meanwhile, \( U = 0 \) and \( V = 0 \) is a saddle point of the objective function (1), the results in Lemma 4 thus indicate that the matrices \( U_k \) and \( V_k \) are updated toward the saddle point zero. While interestingly, as shown in Fig. 4, the matrix norm decreases to a magnitude which is strictly larger than zero, then the matrix norm begins to increase. These observation indicates that ScaledGD can escape from the saddle point zero, the saddle avoid phase is also illustrated in Fig. 5.

**Analysis on the entire iteration process.** It can be easily deduced from Eq. (16) that
\[ \| U_k V_k^T - M \|_F < (1 - \eta)2^k \| U_0 V_0^T - M \|_F + \| M \|_F, \] (24)
4.2 Convergence of AltScaledGD

The convergence analysis of the AltScaledGD is similar to that of ScaledGD, while different to ScaledGD, the objective function (1) is upper-bounded by three terms in AltScaledGD as

\[
\|U_{k+1}V_{k+1}^T - M\|_F \leq (1 - \eta)^2 \|U_kV_k^T - M\|_F + (\eta - \eta^2) \|M\|_F \|V_k^\top V_k\|_2 + \eta \|M\|_F \|U_k^\top U_{k+1}\|_2
\]

\(1\) \hspace{1cm} \(2\) \hspace{1cm} \(3\)

Since in this period of time, the norm of the matrices \(U_k\) and \(V_k\) decreases, while once \(\chi_k = \eta\) (equivalently \(\tau_k = 1\)), according to Lemma 4 and Lemma 5 the norm of the matrices \(U_k\) and \(V_k\) begin to increase, which indicates that the matrices \(U_k\) and \(V_k\) are escaping from the saddle point zero.
therefore, the analysis of AltScaledGD for $\{4\}$ is much easier than that of the ScaledGD in Eq. (16). Specifically, we only need to guarantee that the distance between subspaces of $U_k$ and $U_k^*$ ($\|U_k^\perp U_k\|_2$, $V_k$ and $V_k^*$ ($\|V_k^\perp V_k\|_2$) decrease linearly. The Lemma 2 also holds for AltScaledGD as long as $\eta_k \leq \frac{c}{\theta_0}$.

**Lemma 6.** (Convergence of the distance between subspaces) For AltScaledGD $\{4\}$, if $\|M\|_F \cos \theta_k^0 \cos \theta_k^1 \geq \|U_k V_k^T\|_F$ and $0 < \eta \leq 1$, then the following holds

$$\|U_k^\perp U_{k+1}\|_2 \leq (1-\eta)\|U_k^\perp U_k\|_2, \quad \|V_k^\perp V_{k+1}\|_2 \leq (1-\eta)\|V_k^\perp V_k\|_2.$$  

The condition in Lemma 2 can be satisfied $\forall k$ if $\|M\|_F \cos \theta_k^0 \cos \theta_k^1 \geq \|U_0 V_0^T\|_F$ and $0 < \eta \leq 1$ as specified by the following lemma. The condition is mild compared to the condition in Lemma 3.

**Lemma 7.** For AltScaledGD $\{4\}$, if $\|M\|_F \cos \theta_k^0 \cos \theta_k^1 \geq \|U_0 V_0^T\|_F$ and $0 < \eta \leq 1$, then the following is true

$$\|M\|_F \cos \theta_k^0 \cos \theta_k^1 \geq \|U_k V_k^T\|_F, \forall k > 0.$$  

The convergence analysis of AltScaledGD is the same as that of the ScaledGD in Section 4.1 with small initialization and general Gaussian initialization (the three phases convergence). The main difference between ScaledGD and AltScaledGD is that $\eta$ in ScaledGD should be small such that $\eta \leq c_\eta < 1$, while for AltScaledGD $\eta$ can be as large as 1. It can be seen from Eq. (29) that if $\eta = 1$, the the AltScaledGD Eq. (4) converges to the global minima in just one iteration. We also illustrate the convergence of ScaledGD Eq. (2) and AltScaledGD Eq. (4) with respect to the learning rate $\eta$ in Fig. 6. It can be seen from Fig. 6 that for small learning rate $\eta = 0.1$, the convergence property (the loss curve) of AltScaledGD is almost exactly the same as ScaledGD, while for large learning rate $\eta = 0.8$, the AltScaledGD converges very fast, in contrast the ScaledGD does not converge as the condition $\eta \leq c_\eta$ is not satisfied according to Lemma 3. These results certifies the superiority of AltScaledGD over ScaledGD, since both ScaledGD and AltScaledGD converges fast with large $\eta$, while the learning rate $\eta$ is upper-bounded by a small constant $c_\eta$ in ScaledGD.

5 Conclusion

In this work, we are the first to rigorously prove the global convergence of ScaledGD and AltScaledGD for the non-convex low rank matrix factorization problem and show that thanks to the preconditioning the global convergence rate of ScaledGD and AltScaledGD are independent of the condition number of the matrix $M$, thus they converge faster than gradient descent algorithm for ill-conditioned problem. We further prove that ScaledGD and AltScaledGD converges linearly from both small initialization and general random initialization, which is in contrast to the existing global convergence analysis that are only applicable to small initialization. Meanwhile, we show that compared to ScaledGD, AltScaledGD is more practical as it enables larger learning rate thus converges fast.

**Limitations.** This paper concerns low-rank matrix factorization which is the population loss of the more general low-rank matrix recovery problem, such as matrix completion and matrix sensing. While the empirical loss is different to the population loss in that the number of the samples is limited, therefore our results can not directly applied to general low-rank matrix recovery. Our further work is to study the empirical loss with the help of RIP condition for matrix sensing and the sampling lower-bound for matrix completion.

**Acknowledgements**

This research was supported by the National Key R&D Program of China (2020YFA0713900), the China NSFC projects under contracts 62372359, 61721002, 12226004, the Macao Science and Technology Development Fund under Grant 061/2020/A2, and the Fundamental Research Funds for the Central University under Grant ZYTS23056.
References


