Global Convergence of Gradient Descent for Asymmetric Low-Rank Matrix Factorization

Tian Ye
Institute for Interdisciplinary Information Sciences
Tsinghua University
yet17@mails.tsinghua.edu.cn

Simon S. Du
Paul G. Allen School of Computer Science and Engineering
University of Washington
ssdu@cs.washington.edu

Abstract

We study the asymmetric low-rank factorization problem:
\[ \min_{U \in \mathbb{R}^{m \times d}, V \in \mathbb{R}^{n \times d}} \frac{1}{2} \| UV^\top - \Sigma \|_F^2, \]
where \( \Sigma \) is a given matrix of size \( m \times n \) and rank \( d \). This is a canonical problem that admits two difficulties in optimization: 1) non-convexity and 2) non-smoothness (due to unbalancedness of \( U \) and \( V \)). This is also a prototype for more complex problems such as asymmetric matrix sensing and matrix completion. Despite being non-convex and non-smooth, it has been observed empirically that the randomly initialized gradient descent algorithm can solve this problem in polynomial time. Existing theories to explain this phenomenon all require artificial modifications of the algorithm, such as adding noise in each iteration and adding a balancing regularizer to balance the \( U \) and \( V \).

This paper presents the first proof that shows randomly initialized gradient descent converges to a global minimum of the asymmetric low-rank factorization problem with a polynomial rate. For the proof, we develop 1) a new symmetrization technique to capture the magnitudes of the symmetry and asymmetry, and 2) a quantitative perturbation analysis to approximate matrix derivatives. We believe both are useful for other related non-convex problems.

1 Introduction

This paper studies the asymmetric low-rank matrix factorization problem:
\[ \min_{U \in \mathbb{R}^{m \times d}, V \in \mathbb{R}^{n \times d}} f(U, V) := \frac{1}{2} \| UV^\top - \Sigma \|_F^2, \]
where \( \Sigma \in \mathbb{R}^{m \times n} \) is a given matrix of rank \( d \). While solving this optimization problem is not hard (e.g., using power method), in this paper, we are interested in using randomly initialized gradient descent to solve this problem:
\[ U_{t+1} = U_t + \eta(\Sigma - U_t V_t^\top)V_t; \]
\[ V_{t+1} = V_t + \eta(\Sigma - U_t V_t^\top)^\top U_t, \]
where $\eta > 0$ is the learning rate and $U_0$, $V_0$ are randomly initialized according to some distribution. Empirically, gradient descent with a constant learning rate can efficiently solve this problem (see, e.g., Figure 1 in Du et al. [2018]). Somehow surprisingly, there is no global convergence proof of this generic algorithm, let alone convergence rate analysis. The main difficulties are 1) the problem is non-convex and 2) this problem is not smooth with respect to $(U, V)$ because the magnitudes of them can be highly unbalanced.

To motivate the study of gradient descent for this optimization problem, we note that this is a prototypical optimization problem that illustrates the gap between practice and theory. In particular, the prediction function $UV^\top$ is homogeneous: if we multiply a factor by a scalar $c$ and divide another factor by $c$, the prediction function remains the same. This homogeneity also exists in deep learning models. Therefore, progress made in understand $\ddot{U}$ can further help us gain understanding on other non-convex problems, such as asymmetric matrix sensing, asymmetric matrix completion, and deep learning optimization. We refer readers to Du et al. [2018] for more discussions.

For Problem (1), Du et al. [2018] showed gradient flow (gradient descent with the step size $\eta \to 0$),

$$\dot{U} = (\Sigma - UV^\top) V \quad \text{and} \quad \dot{V} = (\Sigma - UV^\top)^\top U,$$

converges to the global minimum but no rate was given. Here $U : \mathbb{R} \to \mathbb{R}^{m \times d}$, $U = U(t)$ and $V : \mathbb{R} \to \mathbb{R}^{n \times d}$, $V = V(t)$ are an integral curve over manifold $\mathbb{R}^{m \times d} \times \mathbb{R}^{n \times d}$, and $\bar{U} := \frac{dU}{dt}$, $\bar{V} := \frac{dV}{dt}$. Key in their proof is an invariance maintained by gradient flow:

$$\frac{d}{dt} (U^\top U - V^\top V) = 0.$$ 

This invariance implies that if initially the difference between the magnitudes of $U_0$ and $V_0$ is small, then the difference remains small. This in turn guarantees the smoothness on the gradient flow trajectory. Du et al. [2018] further uses a geometric result (all saddle points in the objective function are strict and all local minima are global minima [Ge et al. 2015, 2017b, 2016, Li et al. 2019b]), and then invokes the stable manifold theorem to show the global convergence of gradient flow [Lee et al. 2016, Panageas and Piliouras 2016].

However, to prove a polynomial convergence rate, the approach that solely relies on the geometry will fail because there exists a counter example [Du et al. 2017]. Furthermore, for gradient descent with $\eta > 0$, the key invariance no longer holds [Du et al. 2018] also studied gradient descent with decreasing step sizes $\eta_t = O \left(t^{-1/2}\right)$, and obtained an “approximate global optimality result”: if the magnitude of the initialization is $O(\delta)$, then gradient descent converges to a $\delta$-optimal solution, i.e., this result does not establish that gradient descent converges to a global minimum. And again, there was no convergence rate. Furthermore, their result crucially relies on $\eta_t$ is of order $O \left(t^{-1/2}\right)$ to ensure the second order term does not diverge and thus does not apply to gradient descent with a constant learning rate.

Some previous works, e.g., [Ge et al. 2015, Jin et al. 2017], modified the gradient descent algorithm to the perturbed gradient descent algorithm by adding an isotropic noise at each iteration, which can help escape strict saddle points and bypass the exponential lower bound in [Du et al. 2017]. To deal with the non-smooth problem, they also added a balancing regularization term [Park et al. 2017, Tu et al. 2016, Ge et al. 2017a, Li et al. 2019b]. $rac{1}{2} \| U^\top U - V^\top V \|_F^2$ to the objective function to ensure balancedness between $U$ and $V$ throughout the optimization process. With these two modifications, one can prove a polynomial convergence rate. However, experiments suggest that the isotropic noise and the balancing regularizer may be proof artifacts, because vanilla gradient descent applies to the original objective function $\ddot{U}$ without any regularizer finds a global minimum efficiently. From a practical point of view, one does not want to add noise or additional regularization because it may require more hyper-parameter tuning.

The only global quantitative analysis for randomly initialized gradient is by Du et al. [2018] who proved the global convergence rate for the case where $\Sigma$ has rank 1, and $U$ and $V$ are two vectors. In this case, one can reduce the problem to the dynamics of 4 variables, which can be easily analyzed. Unfortunately, it is very difficult to generalize their analysis to the general rank setting. Recent work also studies this case with noise injected to the gradient update [Liu et al. 2021].

In this paper, we develop new techniques to overcome the technical difficulties and obtain the first polynomial convergence of randomly initialized gradient descent for solving the asymmetric low-rank

---

1. While the invariance can still hold approximately in some way, characterizing the approximation error is highly non-trivial, and this is one of our key technical contributions.
matrix factorization problem. Most importantly, our analysis is completely different from existing ones: we give a thorough characterization of the entire trajectory of gradient descent.

Before presenting our main results, we emphasize that the goal of this paper is not to provide new provably efficient algorithms to solve Problem 1, but to provide a rigorous analysis of an intriguing and practically relevant phenomenon on gradient descent. This is of the same flavor as the recent breakthrough on understanding Burer-Monier method for solving semidefinite programs.\cite{Cifuentes2019}.

1.1 Main Results

Our main result is below.

**Theorem 1.1.** Suppose each entry of $U_0$ and $V_0$ are initialized using Gaussian distribution with mean 0 and variance $\varepsilon^2$, where $\varepsilon = \tilde{O} \left( \frac{\sigma_d}{\sqrt{d^3 \sigma_1 (m+n)}} \right)^2$. Then there exists $T_{\text{total}}(\delta, \eta) = O \left( \frac{1}{\eta} \ln \frac{\delta a_d}{\varepsilon} + \frac{1}{\eta \varepsilon^2} \ln \frac{\sigma_d}{\eta} \right)$ such that for any $\delta > 0$ and learning rate $\eta = O \left( \frac{\sigma_d a^2}{d \varepsilon^2} \right)$, we have that with high probability over the initialization, when $t > T_{\text{total}}(\delta, \eta)$, $f(U_t, V_t) \leq \delta$.

Here, $\sigma_1$ and $\sigma_d$ are the largest and the smallest singular values of $\Sigma$, respectively. Notably, in sharp contrast to the result in \cite{Du2018}, which requires the initialization depends on $\delta$, our initialization does not depend on the target accuracy. To our knowledge, this is the first global convergence result for gradient descent in solving Problem 1. Furthermore, we give a polynomial rate. The first term in $T_{\text{total}}(\delta, \eta)$ represents a warm-up phase and the second term represents the local linear convergence phase, which will be clear in the analysis sections. On the other hand, while we believe $T_{\text{total}}(\delta, \eta)$ is nearly tight, our requirement for $\eta$ is loose. An interesting future direction is further relax this requirement.

Now by taking $\eta \to 0$, we have the following corollary for gradient flow.

**Corollary 1.2.** Given $\delta > 0$, there exists $T = O \left( \frac{1}{\sigma_d} \ln \frac{\sigma_d}{\varepsilon} + \frac{1}{\sigma_d} \ln \frac{\sigma_d}{\eta} \right)$, such that with high probability over the initialization, for all $t \geq T$, we have $f(U_t, V_t) \leq \delta$.

This is also the first convergence rate result of randomly initialized gradient flow for asymmetric matrix factorization. We note that our analysis on gradient flow is nearly tight. To see this, consider the ordinary differential equation $\dot{s}_t = (\sigma_d - a^2) s_t$ with initial point $s_0 > 0$, then $s = a^2$ has analytical solution $s_t = \frac{\sigma_d a^2}{a_0^2 + \frac{\sigma_d a^2}{a_0^2} e^{-\frac{t}{\sigma_d a_0^2}} - 1}$. Hence, to achieve a $\delta$ optimal solution, i.e. $|\sigma_d - a^2| \leq \delta$, we need $\sigma_d \frac{2a_0^2}{\sigma_d} - 1 > \frac{1}{2} \frac{2a_0^2}{\sigma_d} + \frac{\sigma_d}{a_0^2} - 1$. Hence $T = O \left( \frac{1}{\sigma_d} \ln \frac{\sigma_d}{\eta} \right)$ is necessary.

**Remark 1.2.** We note that the weights will also converge. In section 3.3, we will prove that $U$ and $V$ are always bounded, which implies the gradient norm is bounded in terms of the loss. Since the loss function converges linearly, we have that the gradient norm converges to 0 linearly. The convergence of $U$ and $V$ follows from that the trajectory of $(U_t, V_t)$ forms a Cauchy sequence.

1.3 Additional Related Work

Here we discuss additional related work. First, in the symmetric setting, e.g., $\min_{U} \|UU^T - \Sigma\|_F^2$, global convergence of randomized initialized gradient has been established in various settings.\cite{Jain2017, Li2018, Chen2019} However, as has been highlighted in \cite{Li2019a, Park2017, Tu2016}, generalization to the asymmetric case is highly non-trivial. The major technical difficulty is to deal with the unbalancedness between $U$ and $V$. To prevent this, additional balancing regularization is often added.\cite{Li2019b, Park2017, Tu2016, Sun2016} Though empirically this has been shown to be unnecessary.

Another line of work showed one can first uses spectral initialization to find a near-optimal solution, then starting from there, gradient descent converges to an optimum with a linear rate.\cite{Tu2016}

\footnote{\(\tilde{O}\) hides logarithmic terms.}

\footnote{In gradient flow, $t$ is a continuous time index.}

\footnote{In Appendix B, we show the dynamics of gradient flow actually admits a closed form, and thus can be easily analyzed.}
We define such that when we omit the subscript (iteration number) of a matrix, then it represents that this matrix at any iteration. To proceed, we will analyse the principal space and the complement space separately. We denote the algorithm, non-bold letters, e.g., $U$, $V$, $\Sigma$, are for our analysis. For a matrix $W$ with rank $r$, denote $\sigma_i(W)$ as the $i$th largest singular value of $W$, $\forall i \in [r]$. Furthermore, if $W$ is symmetric, denote $\Lambda_i(W)$ as the $i$th largest eigenvalue of $W$. Let $\Sigma \in \mathbb{R}^{m \times n}$ be a rank-$d$ matrix with singular value $\sigma_1 \geq \cdots \geq \sigma_d > 0$, and define its conditional as $\kappa := \frac{\sigma_1}{\sigma_d}$. Our goal is to factorize $\Sigma$ into $UV^\top$.

## 2 Main Difficulties and Technique Overview

### 2.1 A Reduction to Principal and Complement Spaces.

The starting point is the Polyak-Łojasiewicz condition: if we can establish that $\max \{\sigma_d(U_t), \sigma_d(V_t)\}$ is lower bounded by a considerable constant $c_{\text{max}}$, then we have $\|\nabla f(U_t, V_t)\| \geq c_{\text{max}} \sqrt{2f(U_t, V_t)}$, which implies a linear convergence. However, the $d$th singular values of $U$ and $V$ are not monotonic with $t$, and they can even decrease to an extremely small value.

To deal with this issue, we consider the following transformation. Let the singular value decomposition of $\Sigma$ is $\Sigma = \Phi \Sigma' \Psi^\top$, where $\Phi \in \mathbb{R}^{m \times m}$ and $\Psi \in \mathbb{R}^{n \times n}$ are unitary matrices, and $\Sigma'$ is diagonal matrix. Define $U'_t := \Phi^{-1} U_t$ and $V'_t := \Psi^{-1} V_t$. Then we can rewrite equations (2) and (3) as

\[
U'_{t+1} = U'_t + \eta(\Sigma' - U'_t V'_t^\top) V'_t;
\]

\[
V'_{t+1} = V'_t + \eta(\Sigma' - U'_t V'_t^\top)^\top U'_t.
\]

Hence, without loss of generality, we can assume $\Sigma$ is a diagonal matrix with $\Sigma_{i,i} = \sigma_i$, $\forall i \in [d]$, and $\Sigma_{i,j} = 0$ otherwise.

To proceed, we will analyse the principal space and the complement space separately. We denote the upper $d \times d$ matrix of $\Phi$ as $U$ and denote the lower $(m-d) \times d$ matrix of $\Phi$ as $J$. Similarly, we define the upper $d \times d$ matrix of $\Psi$ as $V$ and the lower $(n-d) \times d$ matrix as $K$. Define $\Sigma := \text{diag}(\sigma_1, \cdots, \sigma_d)$. We can write out the dynamics of these matrices:

\[
U_{t+1} = U_t + \eta(\Sigma - U_t V_t^\top) V_t - \eta U_t K_t^\top K_t; \quad (6)
\]

\[
V_{t+1} = V_t + \eta(\Sigma - U_t V_t^\top)^\top U_t - \eta V_t J_t^\top J_t; \quad (7)
\]

\[
J_{t+1} = J_t - \eta J_t (V_t^\top V_t + K_t^\top K_t); \quad (8)
\]

\[
K_{t+1} = K_t - \eta K_t (U_t^\top U_t + J_t^\top J_t). \quad (9)
\]

**Additional Notations** Throughout this paper, we have some notation conventions. First of all, if we omit the subscript (iteration number) of a matrix, then it represents that this matrix at any iteration $t$. If some matrices without subscripts appear in the same equation, it means the equation holds for arbitrary iteration $t$, and the subscript for each matrix should be the same. For example, if we define $A := \frac{U_t + V_t}{2}$, it means we define $A_t := \frac{U_t + V_t}{2}$, $\forall t \geq 0$. Similarly, we define $B := \frac{U_t - V_t}{2}$.

Besides $A$, $B$, $J$ and $K$, there are some other special capital letters used to represent specific matrices throughout this paper. Here is a list.

\[
S = AA^\top;
\]

\[
P = \Sigma - AA^\top + BB^\top;
\]

\[
Q = AB^\top - BA^\top.
\]

We define such $S$ is because in symmetric case ($B \equiv 0$), although it is hard to find analytical solution for $A$ in continuous time case, we do find analytical form for $S$, which contains all information about the singular values of $A$. 

[2016 Zheng and Lafferty] [2016 Zhao et al.] [2015 Bhojanapalli et al.], though in practice random initialization often suffices. Recently, [Ma et al. 2021] proved that if $t$ the initialization is close to a global minimum and 2) $U$ and $V$ are balanced, then without adding additional balancing regularizer, gradient descent converges to a global minimum. Our stage two’s analysis is similar to theirs. However, their result cannot be directly applied to our analysis because they require a more stringent initialization than our stage two’s initial point. There is also other work studying the balancing effect in terms of the learning rate [Wang et al. 2021].
P and Q are just the symmetric and skew-symmetric part of matrix $\Sigma - UV^\top$. Hence the linear convergence of gradient descent is equivalent the linearly diminishing of P and Q by Pythagorean theorem. We will mention their definitions every time we use them.

2.2 Symmetrization

Our key observation is that although the singular values of U and V may not have monotonic property, the symmetrized matrix has this property. Formally, we define

$$A := \frac{U + V}{2} \quad \text{and} \quad B := \frac{U - V}{2}.$$ 

Here, A represents the magnitude in the principal space and B represents the magnitude of asymmetry. Empirically, we can observe that by choosing a sufficiently small learning rate $\eta$, we have two desired properties:

1. The smallest singular value of A is almost monotonically increasing;
2. The norms of $B, J, K$ are almost monotonically decreasing.

The first property ensures we are learning the “signal”, $\Sigma$, and the second property ensures the “noise” is disappearing. Therefore, if we can establish these two properties, we can prove the global convergence.

2.3 Two Stage Analysis

The analysis for asymmetric low rank case is divided into two stages. In the first stage we mainly focus on the increasing rate of $\sigma_d(A)$. We will prove that in gradient descent method $\sigma_d(A_t)$ increases exponentially fast to $\frac{2}{c^d}$ and then $\|P\|_{op}$ drops exponentially fast to $\frac{2}{c^d}$, while preserving $\|B\|_F, \|J\|_{op}$ and $\|K\|_{op}$ small. In the second stage, we will use the large $\sigma_d(A)$ to lower bound the convergence speed of $\|\Sigma - UV^\top\|_2^2$. We will prove that, once gradient descent starts at a point with small $\|P\|_{op}, \|B\|_F, \|J\|_{op}$ and $\|K\|_{op}$, it will converge to global optimal point exponentially fast.

3 Proof Sketch of Theorem 1.1

3.1 Initialization

We first use a Gaussian distribution to generate matrices $U, V, J, K$ element-wisely and independently. By standard random matrix theory (Corollary 2.3.5 and Theorem 2.7.5 of Tao [2012]), we know that $c > 0$, such that with high probability, the smallest singular value of $\frac{U + V}{2}$ is larger than $\frac{1}{c^d}$, the largest singular value of $\frac{U - V}{2}$ is smaller than $c\sqrt{d}$, the Frobenius norm of B is less than $cd$ and the operator norms of J and K are less than $c\sqrt{\max\{m', d\}}$ and $c\sqrt{\max\{n', d\}}$, respectively, where $m' = m - d, n' = n - d$.

The initializations $U_0, V_0, J_0, K_0$ are then scaled by $\varepsilon$ where $\varepsilon$ specified in Theorem 1.1.

3.2 Stage One: Warm-Up Phase

In this stage, we would like to prove the following theorem.

**Theorem 3.1.** By choosing $\varepsilon = \tilde{O}\left(\frac{\sqrt{d}}{\sigma_0} \frac{\sigma_d}{\sigma_0 (m + n)} \right)$ and $\eta = O\left(\frac{\sigma_0}{\sqrt{d}}\right)$, we have that there exists $T_0 = O\left(\frac{1}{\eta^2} \ln \frac{\sigma_d}{\sqrt{d}} \right)$, such that $\forall t \leq T_0$,

- $A_tA_t^\top \preceq 2\Sigma$;
- $\|B_t\|_F \leq 2cd\varepsilon$;

\[\text{Strictly speaking, we cannot make any assumption on U, V, J, K since they need information of singular value decomposition of } \Sigma. \text{ However, a random generation of } U \text{ and } V \text{ implies a random generation of } U, V, J, K \text{ because we use unitary transformations.}\]
We first give some intuitions about the five conditions in Theorem 3.1. The first condition represents the "signal" is properly bounded from below and above throughout stage one. The second condition shows the magnitude of asymmetry is small throughout stage one. We note that it is crucial to study the Frobenius norm of \( B \) instead of operator norm, because Frobenius norm admits a nice expansion for analysis. The third condition is an important one, which guarantees after \( T_0 \) iterations, we have enough "signal" strength in the principal space. The fourth condition is a technical one, which represents the symmetric error is small after \( T_0 \) iterations. The fifth condition represents the magnitude of the complement space remains small.

**Proof Sketch.** The proof of Theorem 3.1 is quite challenging and require new technical ideas and careful calculations, which we explain below.

To analyze the dynamic of \( \sigma_d(A) \), let us recall how it behaves in the continuous-time case. Our main idea is that, instead of analyzing \( A \) itself, we consider the symmetric matrix \( S := AA^\top \). Then \( \frac{dS}{dt} \approx (\Sigma - S)S + S(\Sigma - S) \) plus some small perturbation terms about \( B, J \) and \( K \).

If we only consider a differential equation \( \dot{S} = (\Sigma - S)S + S(\Sigma - S) \), a well-known theorem (Theorem 12 in Lax [2007]) shows that if the singular values of \( S \) are different from each other, and \( \xi \) is the singular vector that \( S\xi = \sigma_d(S)\xi \), then the derivative of \( \sigma_d(S) \) is exactly \( \dot{\xi}^\top \dot{S} \xi \), which is lower bounded by \( 2(\sigma_d - \sigma_d(S))\sigma_d(S) \). To adapt it to discrete case, we prove the following lemma.

**Lemma 3.2.** Suppose \( S, \Sigma \in \mathbb{R}^{d \times d} \) are two definite positive matrices, \( \eta > 0 \), and \( S' = (I + \eta(\Sigma - S))(I + \eta(\Sigma - S)) \). Suppose \( \sigma_1(S) \leq \sigma_2, \sigma_d(\Sigma) \geq \sigma_d \) and \( \sigma_1(\Sigma) \leq \sigma_1 \). Define \( s = \sigma_d(S) \) and \( s' = \sigma_d(S') \). Then \( \forall \beta \in (0, 1) \) and \( \eta \leq \frac{\beta}{8\sigma_1} \),

\[
  s' \geq (1 + \eta(\sigma_d - s))^2 s - \frac{8 + 6\beta}{1 - \beta} \sigma_1^2 \eta^2.
\]

This lemma shows if we ignore perturbations from \( B, J \) and \( K \), then for small \( \eta \) (when \( \eta^2 \) is of smaller order than the first term), the least eigenvalue of \( S \) increases at a geometric rate.

However, there are also some small perturbation terms about \( B, J \) and \( K \) while doing analysis. \( ||J||_{op} \) and \( ||K||_{op} \) are easy to give an upper bound, since by \( (8)\) and \( (2) \), we know that by choosing small enough \( \eta \), they are monotonically decreasing. However, the dynamic of \( B \) is highly non-trivial. After some careful calculations (cf. (13)), we find that the increasing rate of \( ||B||_{op}^2 \) is related to the smallest eigenvalue of \( P := \Sigma - AA^\top + BB^\top \), if \( \max\{0, -\lambda_d(P)\} \) is small, then \( ||B||_{op}^2 \) increases slowly.

Now we would like to give a lower bound on \( \lambda_d(P) \). Inspired by gradient flow case, \( P \) and \( S \) is almost complementary of each other, and their dynamic behaves similarly. Hence we have \( P \approx -(\Sigma - P)P - P(\Sigma - P) \) with some small perturbation terms about \( B, J \) and \( K \). Hence we can use lemma \( (9) \) to give a lower bound in discrete case.

**Lemma 3.3.** Suppose \( P, \Sigma \in \mathbb{R}^{d \times d} \) are two symmetric matrices, \( \eta > 0 \), and \( P' = (I - \eta(\Sigma - P))(I - \eta(\Sigma - P)) \). Suppose \( \sigma_1(P) \leq 2\sigma_1 \) and \( \sigma_2(\Sigma) \leq \sigma_1 \leq \sigma_1 \). Define \( p = \lambda_d(P) \) and \( p' = \lambda_d(P') \). Then \( \forall \beta \in (0, 1) \) and \( \eta \leq \frac{\beta}{8\sigma_1} \),

\[
  p' \geq \left\{ \begin{array}{ll}
    (1 - \eta\sigma_d)^2 p - \frac{8 + 6\beta}{1 - \beta} \sigma_1^2 \eta^2, & \text{if } p < 0; \\
    0, & \text{if } p \geq 0.
  \end{array} \right.
\]

Notice that we use \( B \) while analyzing \( P \) and use \( P \) while analyzing \( B \). Hence, during the whole process, we need to bound both of them inductively.

Finally, once \( \sigma_d(A) \) increases to a relatively large amount, we can use it to prove that \( ||P||_{op} \) will decrease exponentially fast to \( \frac{\sigma_d^2}{4d} \). One cannot simply prove that \( P \) converges to zero in this stage, since the perturbation term \( B \) will never converge to zero.

Below we give more details.
3.2.1 Assumptions

We make some assumptions on \( A \) and \( B \) in iterations \( t \leq T_0 \), where \( T_0 \) will be defined at the end of subsubsection 3.2.4 and we will verify the assumptions in the end.

1. \( \frac{\sigma_d}{\sigma_a^2} I \leq AA^\top \leq 2\Sigma \).
2. The Frobenius norm of \( B \) is bounded by \( e_b dz \) for some \( e_b \geq c \), where \( e_b \) will be determined later. Hence its operator norm is also bounded by \( e_b dz \).

3.2.2 Dynamics on \( A, B \) and \( P \)

The dynamics on \( J \) and \( K \) is trivial, since by equations (11) and (12), i.e.

\[
J_{t+1} = J_t - \eta J_t(V_t^TV_t + K_t^TK_t); \\
K_{t+1} = K_t - \eta K_t(U_t^TU_t + J_t^TJ_t),
\]

we know that if we choose \( \eta \leq \frac{1}{3\sigma_1} \), one can inductively proved that \( 0 \leq V_t^TV_t + K_t^TK_t \leq 3\sigma_1 I \) and \( 0 \leq U_t^TU_t + J_t^TJ_t \leq 3\sigma_1 I \) by using the first two assumptions in subsection 3.2.1 and then it follows that the operator norms of \( J \) and \( K \) are monotonically decreasing in this stage.

However, it is non-trivial to prove that \( \| B \|_{op} \) keeps small. We will analyze the dynamics of \( A, B \) and \( P := \Sigma - AA^\top + BB^\top \) together inductively.

First of all, from equations (11) and (12), we can write down the dynamics of \( A := \frac{U+V}{2} \) and \( B := \frac{U-V}{2} \) as following:

\[
A_{t+1} = A_t + \eta(\Sigma - A_tA_t^\top + B_tB_t^\top)A_t - \eta(A_tB_t^\top - B_tA_t^\top)B_t - \eta A_tK_t^TK_t - \eta J_t^TJ_t \\
B_{t+1} = B_t - \eta(\Sigma - A_tA_t^\top + B_tB_t^\top)B_t + \eta(A_tB_t^\top - B_tA_t^\top)A_t - \eta A_tK_t^TK_t + \eta J_t^TJ_t.
\]

We can further calculate

\[
P_{t+1} = P_t - \eta P_t(\Sigma - P_t) - \eta(\Sigma - P_t)P_t + \eta^2(P_tP_tP_t - P_t\Sigma P_t) - 2\eta B_tB_t^\top P_t - 2\eta P_tB_tB_t^\top - \eta(A_t + \eta P_tA_t)C_t^\top - \eta C_t(A_t + \eta P_tA_t)^\top - \eta^2C_tC_t^\top + \eta(B_t + \eta P_tB_t)D_t^\top + \eta D_t(B_t + \eta P_tB_t)^\top + \eta^2D_tD_t^\top \tag{12}
\]

where

\[
C_t := -A_tB_t^\top B_t + B_tA_t^\top B_t - A_tK_tK_t - J_t^TJ_t \tag{13} \\
D_t := +A_tB_t^\top A_t - B_tA_t^\top A_t - A_tK_tK_t + J_t^TJ_t \tag{14}
\]

are two small perturbation terms.

3.2.3 Dynamics on \( A \)

Given (10), we can give a lower bound for the minimal singular value of \( A_{t+1} \).

\[
\sigma_d(A_{t+1}) \geq \sigma_d(A_t + \eta(\Sigma - A_tA_t^\top)A_t) - \eta \| B_tB_t^\top A_t - A_tB_t^\top B_t + B_tA_t^\top B_t - A_tK_tK_t + J_t^TJ_t \|_{op}. \tag{15}
\]

\*\*We will show later that it is appropriate to choose \( e_b = 2c \).\*\*
For the first part, we could define $S_t := A_t A_t^T$, and $S_{t+1} := \left( I + \eta(\Sigma - S_t) \right) S_t \left( I + \eta(\Sigma - S_t) \right)$. Then according to lemma 3.2, by choosing $\beta = \frac{1}{2}$ and $\eta \leq \frac{1}{\sigma \tau}$, we have
\[
\sigma_d(A_{t+1}) \geq \sqrt{(1 + \eta(\sigma_d - \sigma_d(A_t))^2)\sigma_d(A_t)^2 - 22\sigma_d^3\eta^2 - 1.5\sqrt{2}\sigma_d\eta(e^2 + c^2)\varepsilon^2(m + n)d}.
\] (15)

For simplicity, we denote $\sigma_d(A_t)$ by $a_t$, and define $s_t = \sigma_d(S_t) = a_t^2$.

After some routine computations, we can prove that it takes at most $T_1 := O \left( \frac{1}{\eta \sigma d} \ln \frac{2\sigma_d}{\varepsilon^2} \right)$ iterations to make $a_t$ to at least $\sqrt{\frac{\sigma_d}{2\varepsilon^2}}$, and additional computations show that, if $a_t$ is always bounded by $\sqrt{2\sigma_d}$, then once $a_t$ becomes larger than $\sqrt{\frac{\sigma_d}{2\varepsilon^2}}$, it is always larger than $\sqrt{2\sigma_d}$.

3.2.4 Dynamics on $P$

To bound $P_t$ by equation (12), we need to first bound the norms of $C_t$ and $D_t$ by (13) and (14). By simple triangle inequalities we have
\[
\|C_t\|_\text{op} \leq \sqrt{2\sigma_d}(e^2 + c^2)(m + n)d\varepsilon^2;
\]
\[
\|D_t\|_\text{op} \leq 8\sigma_d e_d d\varepsilon e,
\]
where the last inequality holds when choosing $\varepsilon \leq \frac{\sqrt{\sigma_d} e_d}{c(m + n)}$. Then we can conclude that
\[
P_{t+1} = (I - \eta(\Sigma - P_t))P_t(I - \eta(\Sigma - P_t)) + E_t,
\] (16)
where $E_t$ is a matrix with operator norm less than $O(\eta^2\sigma_d^3 + \eta e_d^2 \varepsilon^2(m + n)d\sigma_1 + \eta^2\sigma_d^2 e_d^2 \varepsilon^2(m + n)d)$. By choosing $\varepsilon \leq \frac{\sqrt{\sigma_d}}{c e_d}$ and $\eta \leq \frac{e_d^2}{\sigma_d (m + n)d\sigma_1}$, we have $\|E_t\|_\text{op} \leq O(\eta e_d^2 \varepsilon^2(m + n)d\sigma_1)$. Further more, by choosing $\beta = \frac{1}{2}$ and $\eta \leq \frac{1}{\sigma \tau_1}$ in lemma 3.3 we have
\[
\lambda_d(P_{t+1}) \geq \max \left\{ (1 - \eta\sigma_d)\lambda_d(P_t) - O(\eta e_d^2 \varepsilon^2(m + n)d\sigma_1), -O(\eta e_d^2 \varepsilon^2(m + n)d\sigma_1) \right\}.
\]

Because $P_0$ is initially positive, we know that
\[
\lambda_d(P_t) \geq -O(e_d^2 \varepsilon^2(m + n)d\kappa).
\] (17)

This lower bound verifies the assumption that $A_t A_t^T \leq 2\Sigma$, since $A_t A_t^T = \Sigma + P + B_t B_t^T \leq \Sigma + O(\eta e_d^2 \varepsilon^2(m + n)d\kappa)I + e_d^2 \varepsilon^2 dI \leq 2\Sigma$ by choosing $e_d^2 \varepsilon^2 = O \left( \frac{\sigma_d}{(m + n)d\kappa} \right)$.

On the other hand, we can also analyze the operator norm of $P$ by using formula (16), since $\sigma_d(A_t) \geq \sqrt{\frac{\sigma_d}{2}}$ for $t \geq T_1$. This implies that
\[
\sigma_1(P_{t+1}) \leq \left( 1 - \frac{\eta \sigma_d}{2} \right)^2 \sigma_1(P_t) + O(\eta e_d^2 \varepsilon^2(m + n)d\sigma_1),
\]
and it follows that
\[
\sigma_1(P_{t+t_1}) \leq \left( 1 - \frac{\eta \sigma_d}{2} \right)^{2t} \sigma_1(P_{t_1}) + O(\eta e_d^2 \varepsilon^2(m + n)d\kappa).
\] (18)

Inequality (18) shows that we only need at most $T_2 := O \left( \frac{1}{\eta \sigma d} \ln \kappa \right)$ iterations after $T_1$ to make $\sigma_1(P_t) \leq \frac{\sigma_d}{4}$. Because $\varepsilon^2 \leq \sigma_d$, we have the total number of iteration $T_0 := T_1 + T_2 = O \left( \frac{1}{\eta \sigma d} \ln \frac{d\sigma_d}{\varepsilon^2} \right)$.

---

1Please see (10) for the full steps for this inequality.
2Please see section D for details.
3Please see (11) for full steps for this inequality.
3.2.5 Dynamics on $B$

To verify the assumption about $\|B\|_F$ made in subsection 3.2.1, we cannot simply use the equation (1), since the error term $(AB^T - BA^T)A_{op}$ is approximately $O(\sigma_1 \|B\|_{op})$, which will perturb the analysis seriously. Inspired by the continuous case that $\|B\|_F^2 = 2 \langle B, \dot{B} \rangle = \text{Tr}(B^T PB) - \frac{1}{2} \|Q\|_F^2 \leq \text{Tr}(B^T PB)$, where $Q = AB^T - BA^T$ if we assume $J = K = 0$. In this inequality, we hide the term $(AB^T - BA^T)AB^T$ in $-\frac{1}{2} \|Q\|_F^2$, and wipe it completely in our analysis.

Hence, for discrete case, we have the following inequality:

$$
\|B_{t+1}\|_F^2 - \|B_t\|_F^2 \leq -2\eta \lambda_d(P_t) \|B_t\|_F^2 + \eta \|B_t^T A_t \|_F \|K_t^T K_t - J_t^T J_t\|_F \\
+ \eta^2 \|\Sigma - A_t A_t^T + B_t B_t^T\|_F \|B_t + (A_t^T B_t - B_t A_t^T) A_t \|_F \\
- A_t \frac{K_t^T K_t - J_t^T J_t}{2} - B_t \frac{K_t^T J_t + J_t^T K_t}{2} \|_F^2 \\
\leq O(\eta \sigma_b^2 \epsilon^2 (m + n) d \kappa) \|B_t\|_F^2 + O(\eta \sqrt{\sigma_1} \epsilon_b (m + n) d^2 \epsilon^3),
$$

(19)

where the last equation is because we have chosen $\eta = O\left(\frac{\sigma_d \epsilon}{\sqrt{\sigma_1}}\right)$ and $\epsilon_b \epsilon^2 = O\left(\frac{\sigma_d}{\sqrt{\sigma_1} \epsilon_b (m + n) d} \right)$.

By some routine calculations in section E, we have that by choosing $\epsilon = \tilde{O}\left(\frac{\sigma_d}{\sqrt{\sigma_1} \epsilon_b (m + n)}\right)$, it is appropriate to choose $\epsilon_b = 2c$, so that $\|B_t\|_F^2 \leq \epsilon_b^2 d^2 \epsilon^2$, and induction holds.

3.3 Stage Two: Local Convergence Phase

We have proved in theorem 3.1 that the gradient descent achieved a pretty good point at $T_0$, i.e. $\|B_{T_0}\|_F \leq 2 \epsilon d \epsilon^3$ and $\sigma_1(P_{T_0}) \leq \frac{\sigma_d}{4}$. In this subsection, we will prove that start from this point, the gradient descent will converge linearly to the global optimal point. Then theorem 1.1 follows.

We will prove inductively on the following conditions:

1. $\|B\|_F = O\left(\frac{\sigma_d}{\sqrt{\sigma_1}}\right)$;
2. $\Delta_t := \|\Sigma - U_{T_0 + t} \Sigma_{T_0 + t}^T\|_{op} \leq \left(1 - \frac{n \sigma_d}{2}\right)^t \frac{\sigma_d}{2}$;
3. $\sigma_d(U), \sigma_d(V) \geq \sqrt{\frac{\sigma_1}{n}}$.

Intuitively, the (1) guarantees the magnitude of asymmetry remains small; (2) guarantees that in the principal space, the error converges to 0 with a geometric rate; and (3) guarantees the “signal” in the principal space remains lower bounded.

**Proof Sketch.** First of all, it is easy to prove linear convergence of $J$ and $K$ by using assumption (3). Now we can verify the assumptions inductively.

1. $1 + 2 \Rightarrow 3$: We can prove $\sigma_d(UV^T) = \Theta(\sigma_d)$. Because $U - V$ is small, (3) follows by triangle inequality.
2. $1 + 3 \Rightarrow 2$: Consider continuous-time case, if we assume $J = K = 0$, the time derivative of $\Sigma - UV^T$ is $-(\Sigma - UV^T)VV^T - UV^T (\Sigma - UV^T)$. Hence the convergence rate is lower bounded by $\sigma_d(U)$ and $\sigma_d(V)$. Because the perturbation term $J$ and $K$ decreases exponentially, assumption (2) follows naturally. We transform this intuition to the discrete-time case.
3. $2 + 3 \Rightarrow 1$: Again, we use (19) to show that the increasing rate of $\|B\|_F^2$ is bounded by $\Delta$. Because $\Delta$ decreases exponentially, $\|B\|_F^2$ cannot diverge to infinity, but increase by a poly$(m, n, \kappa)$ factor. Then by taking $\epsilon$ sufficiently small we can verify the assumption (1).

The full proof is deferred to appendix.

\[\text{(19)}\]

Please see (22) for full steps for this inequality.
Summary of Stage 2. To sum up, we have $\|\Sigma - \mathbf{U}_t^T \mathbf{V}_t^T \|_F^2 = \|\Sigma - \mathbf{U}_t^T \mathbf{V}_t^T \|_F^2 + \|\mathbf{U}_t^T \mathbf{V}_t^T \|_F^2 + \|\mathbf{J}_t \mathbf{V}_t^T \|_F^2 + \|\mathbf{J}_t \mathbf{K}_t^T \|_F^2$, which can be further bounded by

$$\|\Sigma - \mathbf{U}_{T_0+t}^T \mathbf{V}_{T_0+t}^T \|_F^2 \leq \left(1 - \frac{\eta \sigma_d}{2}\right)^t \frac{\sigma_d}{2} + 2(e^2 + c^t) \sigma_1 (m + n) d \left(1 - \frac{\eta \sigma_d}{2}\right)^{2t}$$

$$\leq C \left(1 - \frac{\eta \sigma_d}{2}\right)^t \sigma_d,$$

for some universal constant $C$. Hence one only needs $T_f := O\left(\frac{\ln d}{\eta \sigma_d}\right)$ iterations after $T_0$ to achieve an $\delta$-optimal point.

4 Conclusion

This paper proved that randomly initialized gradient descent converges to a global minimum of the asymmetric low-rank matrix factorization problem with a polynomial convergence rate. This result explains the empirical phenomena observed in prior work, and confirms that gradient descent with a constant learning rate still enjoys the auto-balancing property as argued in [Du et al. 2018].

We believe our requirement of the step size $\eta$ is loose and a tighter analysis may improve the running time of gradient descent. Another interesting direction is to apply our techniques to other related problems such as asymmetric matrix sensing, asymmetric matrix completion and linear neural networks.\[11\]

Acknowledgement

Simon S. Du gratefully acknowledges funding from NSF Award’s IIS-2110170 and DMS-2134106.

References


\[11\] For instance, we refer readers to section for a possible way to extend our technique to matrix sensing.


