Global Convergence Rate of Gradient Flow for Asymmetric Matrix Factorization

Tian Ye YET17@MAILS.TSINGHUA.EDU.CN
Tsinghua University, China
Simon S. Du SSDU@CS.WASHINGTON.EDU
University of Washington, The U.S.

Abstract

We analyze the convergence rate of gradient flow for solving $\min_{U\in\mathbb{R}^{d\times d},V\in\mathbb{R}^{d\times d}} \frac{1}{2}\|UV^\top - M\|_F^2$ in the case M is full-rank, and U and V are randomly initialized. In contrast to previous work, our analysis does not require any balancing regularizer or additive isotropic noise. Our key idea is to couple the trajectory of the gradient flow with an ideal trajectory induced by a symmetric training process. We believe this technique will have applications in other problems.

1. Introduction

This paper studies the convergence rate of applying gradient descent to solve the *asymmetric* matrix factorization

$$\min_{U\in\mathbb{R}^{d\times d},V\in\mathbb{R}^{d\times d}V} \frac{1}{2}\|UV^\top - M\|_F^2$$

The main difficulties are 1) the problem is non-convex and 2) this problem is not smooth with respect to U and V because the magnitudes of them can be imbalanced. This is a prototypical problem that has the difficulty in analyzing the convergence of optimization method for homogeneous models, such as deep neural networks. See [2] for more discussions.

Du et al. [2] showed the global convergence of gradient flow but rate is given, but no rate was given. Their analysis relies on the geometric result that all saddle points in the objective function is strict [3], and then invokes the stable manifold theory used in [5]. However, to prove the polynomial convergence rate, the approach that solely relies on the global geometric will fail because there exists a counter example [1].

Some previous work, e.g., [4] changed the gradient descent algorithm to noisy gradient descent by adding additive isotropic noise, which can help escape strict saddle points and bypass the exponential lower bound in [1], and then added an additional regularizer

$$\frac{1}{8}\|U^\top U - V^\top V\|_F^2$$

to the objective function to ensure balancedness between U and V throughout the training process. With these two artificial modifications, one can prove a polynomial convergence rate. Another line of work, e.g., [7], 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. However, in has been found empirically that these modifications are not necessary. *Randomly initialized* gradient descent without any additional regularization or additive noise converges to the global minimum with a linear rate. See Figure 1 in [2].

To our knowledge, the only result for randomly initialized gradient is by Du et al. [2] who proved the global convergence rate for the case M 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 easily characterized. Unfortunately, it is very difficult to generalize their analysis to the high rank scenario.

In this paper, we take step to understand the global convergence rate of randomly initialized gradient descent. We analyze continuous time gradient descent in the case M has full rank. Our main result is the following.

Theorem 1 *There exists universal constants δ and δ' such that the following statement is true. Suppose U_0 and V_0 are two random matrices whose coefficients u_{ij} and v_{ij} are independent, and are of Gaussian distribution with mean 0 and variance $\sigma_d e^{-\delta\kappa \ln \kappa d}$. Then with high probability, the integral curve generated by (1) and (2), called $U(\cdot)$ and $V(\cdot)$, converges at the global optimal point at rate*

$$f(U(t), V(t)) \leq \epsilon \|\Sigma\|^2, \forall t \geq \delta' \left(\frac{\kappa}{\sigma_d} \ln(\kappa d) + \frac{\ln \frac{1}{\epsilon}}{\sigma_d} \right).$$

To our knowledge, this is the first quantitative global convergence result of gradient descent for asymmetric factorization. While our result only holds for gradient flow, in the full version of the paper, we will present result for gradient descent with a constant step size.

2. Problem Setup

Let $\Sigma \in \mathbb{R}^{d\times d}$ be a non-singular matrix with singular value $\sigma_1 \geq \cdots \geq \sigma_d > 0$, and U and V are two matrices with the same size. We study the objective function

$$f(U, V) := \frac{1}{2}\|\Sigma - UV^\top\|_F^2,$$

and use gradient descent to optimize U and V. In this paper analyze the convergence rate of continuous time gradient descent (gradient descent with stepsize → 0), a.k.a., gradient flow. More precisely, we deal with the ODE

$$\dot{U} = -\tfrac{\partial f}{\partial U} = (\Sigma - UV^\top)V; \tag{1}$$

$$\dot{V} = -\tfrac{\partial f}{\partial V} = (\Sigma - UV^\top)^\top U. \tag{2}$$

Equations (1) and (2) define a smooth vector field of manifold $\mathcal{M} := \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d}$. Suppose $\theta : \mathcal{D} \to \mathcal{M}$ is the maximal flow generated by the vector field, where $\mathcal{D} \subseteq \mathbb{R} \times \mathcal{M}$. Our goal is to prove the global convergence speed of $f \circ \theta(\cdot, M_0)$ for some initial M_0 with large probability, where the probability is induced by the initial distribution.

If Σ is symmetric and $U = V$, the Section 4 will show that the integral curve converges linearly to the global optimum. However, proving the analogue result for the asymmetric case is significantly more difficult.

2.1. Notations

We use $\sigma_1 \geq \cdots \geq \sigma_d$ to represent the singular values of matrix Σ, where Σ is assumed to be a diagonal matrix. With a little abuse of notation, we use $\sigma_i(\cdot)$ and $\lambda_i(\cdot)$ to represent the i^{th} singular value and eigenvalue of a given matrix. Define $\kappa : \frac{\sigma_1}{\sigma_d}$ as the condition number of Σ.

3. Main Difficulty and Technique Overview

Let the singular value decomposition of Σ is $\Sigma = \Phi\Sigma'\Psi^\top$ where $\Sigma' = \text{diag}(\sigma_1, \cdots, \sigma_d)$. Also, we suppose $U = \Phi U'$ and $V = \Psi V'$. We can rewrite (1) and (2) into

$$\begin{aligned}\dot{U'} &= (\Sigma' - U'{V'}^\top)V'; \\ \dot{V'} &= (\Sigma' - U'{V'}^\top)^\top U',\end{aligned}$$

which are exactly the same as the original equations. Hence, we can assume, without loss of generality, $\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_d)$. However, we cannot further assume $U = V$, since Φ and Ψ are unknown.

The first difficulty is that the smallest singular value of U and V are not monotonically increasing anymore, which deprives us of the ability to analyze the singular value initially.

In our approach, we will study two auxiliary curves $A := \frac{U+V}{2}$ and $B := \frac{U-V}{2}$, we have their own evolution equations:

$$\begin{aligned}\dot{A} &= (\Sigma - AA^\top + BB^\top)A - (AB^\top - BA^\top)B; &(3)\\ \dot{B} &= -(\Sigma - AA^\top + BB^\top)B + (AB^\top - BA^\top)A. &(4)\end{aligned}$$

It comes out that the norm of B is monotonically decreasing. However, even if B is extremely small, we cannot simply apply trivial inequalities on the last terms of (3) and (4), since the algorithm will diverge if we reverse the sign of the last terms. We finally divide the whole integral curve into three parts, each of which is carefully analyzed in the corresponding subsections in Section 5.

4. Warm up: symmetric case

First of all, we can give a tight bound on symmetric case, i.e. the case when initial point $U_0 = V_0$. In this case, the symmetry of (1) and (2) implies that $U \equiv V$ during the evolution. Define $S := UU^\top$. Then fortunately, we have

$$\dot{S} = (\Sigma - S)S + S(\Sigma - S). \tag{5}$$

Define $\mathcal{M}_{\text{sym}} := \mathbb{R}^{d\times d}_{\text{sym}}$ as the manifold of symmetric matrices in $\mathbb{R}^{d\times d}$. Then we can get a maximal flow $\vartheta : \mathcal{D}' \to \mathcal{M}_{\text{sym}}$ generated by smooth vector field (5), where $\mathcal{D}' \subseteq \mathbb{R} \times \mathcal{M}_{\text{sym}}$.

Here are two useful lemmas.

Lemma 2 *Suppose $P : (-a, a) \to \mathcal{M}_{sym}$ is a smooth matrix curve. Suppose the differential equation*

$$\dot{S}(t) = P(t)S(t) + S(t)P(t)$$

with "initial" point $S(0) \succeq 0$ has a solution S. Then $\forall t \in (-a, a)$, $S(t) \succeq 0$. Moreover, if $\forall t \in (-a, a)$, $P(t)$ is a positive semi-definite matrix, then the minimal and the maximal singular values of S is non-decreasing.

With lemma 2, we know that the matrices S and $\Sigma - S$ remain positive semi-definite if they are initially PSD. Hence, they are always bounded by Σ, which implies the domain of $\vartheta(\cdot, S)$ is $\mathbb{R}$.

Lemma 3 *Suppose $S_1(0), S_2(0)$ are two matrices in $\mathcal{M}_{sym}$ such that $S_1(0) \preceq S_2(0)$. Define $S_i(t) := \vartheta(t, S_i(0)), \forall i \in \{1,2\}$, then $\forall t$ in domain, we have $S_1(t_1) \preceq S_2(t_1)$.*

Now, given arbitrary positive definite matrix S, we have $\sigma_d(S)I \preceq S$. By applying lemma 3, we have $\forall t \geq 0$, $\vartheta(t, \sigma_d(S)I) \preceq \vartheta(t, S)$. Because I is always commutable with Σ, we can give analytical expression on them and their eigenvalues. Thus the tight bound for the largest singular value of $\Sigma - S$ follows.

Theorem 4 *Suppose α_1 and α_d are the largest and smallest singular value of U_0. Then $\forall t > 0$, we have*

$$\mathit{diag}\left(\frac{1 - \frac{\sigma_i}{\alpha_1^2}}{e^{2\sigma_i t} + \frac{\sigma_i}{\alpha_1^2} - 1}\right)_{i \in [d]} \preceq \Sigma - U(t)U^\top(t) \preceq \mathit{diag}\left(\frac{1 - \frac{\sigma_i}{\alpha_d^2}}{e^{2\sigma_i t} + \frac{\sigma_i}{\alpha_d^2} - 1}\right)_{i \in [d]}. \tag{6}$$

5. Asymmetric case

In this section, we analyze the convergence property of θ by separating the whole process into three stages.

- In the first stage, the initial matrices are quite small, and we will prove that $\frac{\theta(t,U)+\theta(t,V)}{2}$ are quite close to $\theta\left(t, \frac{U+V}{2}\right)$. If this is true, the smallest singular value of the former matrix can be relatively big, while the difference between U and V are quite small.
- In the second stage, we will prove that the smallest singular value of A increases considerably fast, the function value decreases to a small value, while the difference B keeps being small.
- The third stage is the local linear convergence of the integral curve, by using the continuous version of PL inequality.

We only present proof sketch here. The full proof is deferred to appendix.

5.1. Initialization

We use Gaussian distribution to generate $d \times d$ matrices $\mathcal{U}$ and $\mathcal{V}$ element-wisely and independently. According to Corollary 2.3.5 and Theorem 2.7.5in *Topics in random matrix theory*[6], we observe that $\exists c_1 > 0$, with high probability, the smallest singular values of $\mathcal{U}, \mathcal{V}, \frac{\mathcal{U}+\mathcal{V}}{2}, \frac{\mathcal{U}-\mathcal{V}}{2}$ are larger than $\frac{1}{c_1\sqrt{d}}$, while the largest singular values of $\mathcal{U}, \mathcal{V}, \frac{\mathcal{U}+\mathcal{V}}{2}, \frac{\mathcal{U}-\mathcal{V}}{2}$ are smaller than $c_1\sqrt{d}$.

We assume the integral curve starts at the point $(U_0, V_0) := (\varepsilon\mathcal{U}, \varepsilon\mathcal{V})$ with sufficiently small ε.

5.2. Stage 1

In the first stage, we consider about the center of the original curve $(U(t), V(t))$, say $A(t) = \frac{U(t)+V(t)}{2}$. In the previous section, we have analyzed the curve $(\overline{A}(t), \overline{A}(t)) := \theta\left(t, \left(\frac{U_0+V_0}{2}, \frac{U_0+V_0}{2}\right)\right)$. If we can prove that the error $E(t) := A(t) - \overline{A}(t)$ is small, then A will increase almost as fast as $\overline{A}$.

The intuition of the proof is that the velocity vector at every point of $A(\cdot)$ is extremely close to that of the symmetric case. Hence, even if the error increases exponentially, it is still quite small at the first stage. Please see equation (17) in appendix for details. Finally, at the end of stage 1, the difference B becomes much smaller than its initial value, while the smallest singular value of A becomes much larger.

5.3. Stage 2

In this section, we would like to prove three things.

- The smallest singular value of A continuously increases to a constant fraction of $\sqrt{\sigma_d}$.
- The loss of the function f decreases to a small fraction of σ_d^2.
- The norm of B remains small.

These three facts are simple derivations of their corresponding inequality (21), (10) and (23).

5.4. Step 3

We will prove linear convergence in this stage.

First of all, because the function value $\frac{1}{2}\|\Sigma - UV^\top\|^2$ is small, we could prove the smallest eigenvalues of U and V are small if U and V are close to each other, which in turn implies PL inequalities, a sufficient condition of linear convergence.

On the other hand, if the function value is small, we will prove the increasing speed of $U - V$ is small. By combining these two observations together, the proof of linear convergence completed.

6. Conclusion

In this paper we give the first global convergence rate analysis of gradient flow for solving asymmetric matrix factorization. There are two terms, $\frac{\kappa}{\sigma_d}\ln(\kappa d)$ represents the initial stage and $\frac{\ln(1/\epsilon)}{\sigma_d}$ represents the final local convergence stage.

In the full version of this paper, we will further extend the result to the case where one uses a positive step gradient descent to optimize the objective function. Furthermore, we will also study the case M is of low rank and U and V are low rank factors. Our analysis our highlight the importance of using a trajectory-based analysis, which gives a more fine-grained characterization than the one given by the global geometric approach. We believe that based on our result, one can also prove global convergence rate of gradient decent for solving asymmetric matrix sensing, asymmetric matrix completion, and other related problems. Note for these problems, empirically, randomly initialized (stochastic) gradient decent often gives satisfying outcomes, and no additional regularization or additive noise is needed.

References

[1] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Barnabas Poczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points. *arXiv preprint arXiv:1705.10412*, 2017.

[2] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, pages 384–395, 2018.

[3] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

[4] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1233–1242, 2017.

[5] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

[6] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[7] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via Procrustes flow. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 964–973. JMLR. org, 2016.

Appendix A. Proofs for Section 4

Proof [Proof of Lemma 2] It is well-known that the linear differential equation

$$\dot{H}(t) = P(t)H(t)$$

with "initial" point $H(0) = I$ has a solution H. Consider about the curve $S^*(t) := H(t)S(0)H^\top(t)$. We have

$$\begin{aligned} \dot{S}^*(t) &= P(t)S^*(t) + S^*(t)P(t); \\ S^*(0) &= S(0). \end{aligned}$$

By the uniqueness of differential equation, $S(t) \equiv S^*(t) \succeq 0$.

Moreover, define $\sigma_d^s(t) := \sigma_{\min}(S(t))$ and $\sigma_1^s(t) := \sigma_{\max}(S(t))$. Define σ_d^p and σ_1^p similarly. Then $\forall t_0 \in (-a, a)$, $S(t_0+\epsilon) = S(t_0)+\epsilon P(t_0)S(t_0)+\epsilon S(t_0)P(t_0)+o(\epsilon) = (I+\epsilon P(t_0))S(t_0)(I+\epsilon P(t_0)) + o(\epsilon)$, which comes out immediately

$$\begin{aligned} (1+\epsilon\sigma_d^p(t_0))^2\sigma_d^s(t_0) + o(\epsilon) \le \sigma_d^s(t_0+\epsilon) \le (1+\epsilon\sigma_1^p(t_0))^2\sigma_d^s(t_0) + o(\epsilon); \\ (1+\epsilon\sigma_d^p(t_0))^2\sigma_1^s(t_0) + o(\epsilon) \le \sigma_1^s(t_0+\epsilon) \le (1+\epsilon\sigma_1^p(t_0))^2\sigma_1^s(t_0) + o(\epsilon). \end{aligned}$$

In other words,

$$\begin{aligned} 2\sigma_d^p(t_0)\sigma_d^s(t_0) \le \liminf_{\epsilon\to 0} \frac{\sigma_d^s(t_0+\epsilon) - \sigma_d^s(t_0)}{\epsilon} \le 2\sigma_1^p(t_0)\sigma_d^s(t_0); \\ 2\sigma_d^p(t_0)\sigma_1^s(t_0) \le \limsup_{\epsilon\to 0} \frac{\sigma_1^s(t_0+\epsilon) - \sigma_1^s(t_0)}{\epsilon} \le 2\sigma_1^p(t_0)\sigma_1^s(t_0). \end{aligned}$$

Proof follows by writing it into integral form. ■

Proof [Proof of Lemma 3] Define curve $T(t) := S_2(t) - S_1(t)$. Then

$$\begin{aligned} \dot{T} &= (\Sigma - S_2)S_2 - S_2(\Sigma - S_2) - (\Sigma - S_1)S_1 - S_1(\Sigma - S_1) \\ &= \Sigma(S_2 - S_1) + (S_2 - S_1)\Sigma + 2S_1S_1 - 2S_2S_2. \\ &= (\Sigma - S_1 - S_2)(S_2 - S_1) + (S_2 - S_1)(\Sigma - S_1 - S_2). \end{aligned}$$

By applying lemma 2 on T, the proof follows. ■

Appendix B. Proofs for Section 4

B.1. Stage 1

The first stage is the interval $t \in [0, t_1]$, where t_1 is a parameter we will define later.

First of all, define $\overline{A}(t) := \theta\left(t, \left(\frac{U_0+V_0}{2}, \frac{U_0+V_0}{2}\right)\right)$. By applying lemma 3, we could give bound on singular values of $\overline{A}$.

Lemma 5 *If we assume $A(0)A^\top(0) \preceq \sigma_d I$, we have*

$$\sqrt{\frac{\sigma_d e^{2\sigma_d t}}{e^{2\sigma_d t} + \frac{c_1^2 d\sigma_d}{\varepsilon^2} - 1}} \le \sigma_d(\overline{A}) \le \sigma_1(\overline{A}) \le \sqrt{\frac{\sigma_1 e^{2\sigma_1 t}}{e^{2\sigma_1 t} + \frac{\sigma_1}{\varepsilon^2 c_1^2 d} - 1}}. \tag{7}$$

Proof We first compute the analytical formula for singular values of $A_d(t) := \theta\left(t, \frac{\varepsilon}{c_1\sqrt{d}} I\right)$ and $A_1(t) := \theta\left(t, \varepsilon c_1 \sqrt{d} I\right)$. According to lemma 3, we have $A_d A_d^\top \preceq \overline{AA}^\top \preceq A_1 A_1^\top$. Take A_d as an example. Because A_d is commutable with Σ, we can assume A_d and Σ are initially diagonal. Then, for every singular value σ_i of Σ, it corresponds to a singular value of $A_d(0)A_d^\top(0)$, say $a_d^i(0)$, which is exactly $\frac{\varepsilon^2}{c_1^2 d}$. Because in this case, every element on the diagonal is independent with each other, we can simply solve the ODE and obtain

$$a_d^i(t) = \frac{\sigma_i e^{2\sigma_i t}}{e^{2\sigma_i t} + \frac{c_1^2 d \sigma_i}{\varepsilon^2} - 1}. \tag{8}$$

By viewing it as function of σ_i, we have $a_d^1(t) \leq a_d^2(t) \leq \cdots \leq a_d^d(t)$. Actually, define $g(\sigma) := \frac{\sigma e^{2\sigma t}}{e^{2\sigma t} + \frac{c_1^2 d\sigma}{\varepsilon^2} - 1}$, we have $g'(\sigma) = \frac{e^{4\sigma t} - e^{2\sigma t} + 2t\sigma e^{2t\sigma}\left(\frac{c_1^2 d\sigma}{\varepsilon^2} - 1\right)}{\left(e^{2\sigma_i t} + \frac{c_1^2 d\sigma_i}{\varepsilon^2} - 1\right)^2}$, which is larger than 0, since $t > 0$ and $A(0)A^\top(0) \preceq \sigma_d I$.

Hence $a_d^1 I \preceq \overline{AA}^\top$. The upper bound comes out similarly. ■

Remark 6 *To satisfy the assumption in lemma 5, we only need to choose small ε. More precisely, $\varepsilon^2 \leq \frac{\sigma_d}{c_1^2 d}$.*

Suppose $\alpha \in (1, 0)$ is a parameter we will define later, the first stage is defined to be $\{t \geq 0 | a_1^1(t) \leq \alpha\sigma_d\}$, i.e. $t \in [0, t_1]$ where

$$t_1 := \frac{\ln\left(\frac{\sigma_1}{\varepsilon^2 c_1^2 d} - 1\right) + \ln\frac{\kappa - \alpha}{\kappa}}{2\sigma_1}, \tag{9}$$

where $\kappa := \frac{\sigma_1}{\sigma_d}$. We will prove later that $\sigma_d(\overline{A})$ is of order $\varepsilon^{1-\frac{1}{\kappa}}$.

B.1.1. BOUND ON DIFFERENCE

We have observed that $\sigma_d(\overline{A})$ has increased from $O(\varepsilon)$ to $\Theta(\varepsilon^{1-\frac{1}{\kappa}})$, if we can prove the error $E := A - \overline{A}$ is small, we could prove that $\sigma_d(A)$ is also $\Theta(\varepsilon^{1-\frac{1}{\kappa}})$. Before analyzing the error E, we have to bound another quantity B for preparation. According to equation (4), we have

$$\begin{aligned} \|\dot{B}\|^2 &= \left\langle \dot{B}, B \right\rangle + \left\langle B, \dot{B} \right\rangle \\ &= -2\left\langle B, (\Sigma - AA^\top + BB^\top)B - (AB^\top - BA^\top)A \right\rangle \\ &\leq -2\left\langle B, (\Sigma - AA^\top + BB^\top)B \right\rangle, \end{aligned} \tag{10}$$

where the last inequality come from the simple fact that $\forall R \in \mathbb{R}^{d\times d}$, $\mathrm{Tr}(RR^\top - RR) = \sum\limits_{i,j}(r_{i,j} - r_{j,i})^2 \geq 0$.

Now, we assume that $AA^\top \preceq \alpha'\sigma_d I$ in the first stage, where $\alpha' \in (\alpha, 1)$ is a parameter we will define later. Then $\Sigma - AA^\top + BB^\top \succeq (1-\alpha')\sigma_d I$, which implies that $\|\dot{B}\|^2 \leq -2(1-\alpha')\sigma_d\|B\|^2$. By solving the differential equation, we have

$$\|B(t)\|^2 \leq e^{-2(1-\alpha')\sigma_d t}\|B(0)\|^2. \tag{11}$$

B.1.2. BOUND ON ERROR

To analyze the curve of $\overline{A}$, we define another curve ϕ as following.

$$\begin{aligned} \phi(0, M) &= M; & (12)\\ \phi(t, \phi(s, M)) &= \phi(t+s, M); & (13)\\ \left.\frac{\partial \phi}{\partial t}\right|_{(t_0, M)} &= F(\phi(t_0, M)), & (14) \end{aligned}$$

where $F(M) := (\Sigma - MM^\top)M$. By taking derivative of $s = 0$ in equation (13), we obtain an equation we will use later:

$$\left.\frac{\partial \phi}{\partial M}\right|_{(t, M_0)} \circ F(M_0) = F(\phi(t, M_0)). \tag{15}$$

Then, define a function $\gamma_t(s) := \phi(t-s, A(s))$, where t is a constant here.

Now we have

$$\begin{aligned} A(t) - \overline{A}(t) &= \gamma_t(t) - \gamma_t(0) \\ &= \int_0^t \dot{\gamma}_t(s)\mathrm{d}s \\ &= \int_0^t -F(\phi(t-s, A(s))) + \left.\frac{\partial \phi}{\partial M}\right|_{(t-s, A(s))} \circ \dot{A}(s)\mathrm{d}s \\ &= \int_0^t \left.\frac{\partial \phi}{\partial M}\right|_{(t-s, A(s))} \circ \left(\dot{A}(s) - F(A(s))\right)\mathrm{d}s & (16)\\ &= \int_0^t \left.\frac{\partial \phi}{\partial M}\right|_{(t-s, A(s))} \circ \left(BB^\top A - AB^\top B + BA^\top B\right)(s)\mathrm{d}s. & (17) \end{aligned}$$

Here (16) comes from (15), and (17) comes from (3) and (14). The following lemma gives a bound for the largest singular value of $\left.\frac{\partial \phi}{\partial M}\right|_{(t-s, A(s))}$.

Lemma 7 *If* $\forall t \in [0, t_1], \phi(t, M_0)\phi(t, M_0)^\top \preceq \frac{\sigma_d\sqrt{2}}{2}I$, *then* $\forall D \in \mathbb{R}^{d\times d}, t \in [0, t_1]$,

$$\left\| \left.\frac{\partial \phi}{\partial M}\right|_{(t, M_0)} \circ D \right\| \le e^{\sigma_1 t}\|D\|.$$

Proof Because ϕ is the flow generated by gradient field F, it is natural to consider the smoothness of the function $f_{\text{sym}}(U) := \frac{1}{2}f(U, U)$, whose gradient field is exactly F.

Suppose $\mathcal{H}$ is the Hessian operator of f_{sym}, then $\forall \Delta \in \mathbb{R}^{d\times d}$, we have

$$\mathcal{H}(U) \circ (\Delta, \Delta) = -\left\langle \Sigma - UU^\top, \Delta\Delta^\top \right\rangle + \frac{1}{2}\left\| U\Delta^\top + \Delta U^\top \right\|^2, \tag{18}$$

which is upper bounded by $\sigma_1\|\Delta\|^2$ and lower bounded by $-\sigma_1\|\Delta\|^2$ if $0 \preceq UU^\top \preceq \frac{\sigma_d\sqrt{2}}{2}I \preceq \Sigma$. This exactly implies σ_1-smoothness of this function.

Then

$$\begin{aligned}
\left.\frac{\partial \phi}{\partial M}\right|_{(t_0, M_0)} &= \int_0^{t_0} \left.\frac{\partial^2 \phi}{\partial M \partial t}\right|_{(s, M_0)} \mathrm{d}s \\
&= \int_0^{t_0} \mathcal{H}(\phi(s, M_0)) \circ \left.\frac{\partial \phi}{\partial M}\right|_{(s, M_0)} \mathrm{d}s.
\end{aligned}$$

Take the norm and we have

$$\left\| \left.\frac{\partial \phi}{\partial M}\right|_{(t_0, M_0)} \circ D \right\| \leq \int_0^{t_0} \sigma_1 \left\| \left.\frac{\partial \phi}{\partial M}\right|_{(s, M_0)} \circ D \right\| \mathrm{d}s.$$

Because $\left.\frac{\partial \phi}{\partial M}\right|_{(0, M_0)} \circ D = D$, the proof follows by simply solving the ODE above. ■

Now, choose $\alpha' \leq \frac{\sqrt{2}}{2}$ and we can bound equation (17) by

$$\begin{aligned}
\|A(t) - \overline{A}(t)\| &\leq \int_0^t e^{\sigma_1 (t-s)} \cdot 3\sqrt{\alpha' \sigma_d} e^{-2(1-\alpha')\sigma_d s} \|B(0)\|^2 \mathrm{d}s \\
&\leq e^{\sigma_1 t} \frac{3\sqrt{\alpha'}}{\kappa \sqrt{\sigma_d}} \varepsilon^2 c_1^2 d^2.
\end{aligned}$$

Because $e^{\sigma_1 t} \leq e^{\sigma_1 t_1} = \sqrt{\left(\frac{\sigma_1}{\varepsilon^2 c_1^2 d} - 1\right) \frac{\kappa - \alpha}{\kappa}} \leq \frac{1}{\varepsilon c_1} \sqrt{\frac{\sigma_1}{d}}$, we have

$$\|A(t) - \overline{A}(t)\| \leq \frac{3\sqrt{\alpha'}}{\sqrt{\kappa}} \varepsilon c_1 d^{1.5}. \tag{19}$$

B.1.3. SUMMARY FOR THE STAGE 1

Stage 1 is defined to be $t \in [0, t_1]$, where $t_1 := \frac{\ln\left(\frac{\sigma_1}{\varepsilon^2 c_1^2 d} - 1\right) + \ln \frac{\kappa - \alpha}{\kappa}}{2\sigma_1}$ by (9). According to the definition, in this stage, $\sigma_{\max}(\overline{A}) \leq \sqrt{\alpha \sigma_d}$.

By choosing $\varepsilon = \frac{1}{\xi}\sqrt{\frac{\sigma_d}{c_1^2 d}}$ for some $\xi > 2$ and $0 < \alpha < \alpha' \leq \frac{\sqrt{2}}{2}$, we can lower bound $e^{\sigma_1 t_1}$ by $C_1 \sqrt{\frac{\sigma_1}{\varepsilon^2 c_1^2 d \kappa}}$ for some universal constant $C_1 \in (0, 1)$ (independent of ξ). Furthermore, there exists universal constant $C_2 > 0$ such that $\sigma_{\min}(\overline{A}(t_1)) \geq C_2 \frac{\sqrt{\sigma_d} \xi^{\frac{1}{\kappa} - 1}}{c_1^2 d}$.

Recall that $\|A(t) - \overline{A}(t)\| \leq \frac{3\sqrt{\alpha'}}{\sqrt{\kappa}} \varepsilon c_1 d^{1.5} = \frac{1}{\xi} \cdot \frac{3\sqrt{\alpha' \sigma_d}}{\sqrt{\kappa}} d$. To make $\sigma_{\max}(A) \leq \sqrt{\alpha' \sigma_d}$, we only need to choose large ξ such that $\sqrt{\alpha} + \frac{1}{\xi} \frac{3d\sqrt{\alpha'}}{\sqrt{\kappa}} \leq \sqrt{\alpha'}$ by triangle inequality. Hence, choosing $\xi \geq \frac{\sqrt{\alpha}}{\sqrt{\alpha'} - \sqrt{\alpha}} \frac{3d}{\sqrt{\kappa}}$ is appropriate.

Finally, we would like to give a lower bound of $\sigma_{\min}(A(t_1))$. Triangle inequality implies that $\sigma_{\min}(A(t_1)) \geq \frac{1}{\xi}\left(C_2 \xi^{\frac{1}{\kappa}} \frac{\sqrt{\sigma_d}}{c_1^2 d} - \frac{3\sqrt{\alpha' \sigma_d}}{\sqrt{\kappa}} d\right)$. Then, there exists universal constant $C_3, C_4 > 0$, such that $\forall \zeta > C_3$, choosing $\xi = \zeta \left(\frac{3\sqrt{\alpha'} d^2 c_1^2}{C_1 \sqrt{\kappa}}\right)^{\kappa}$ makes $\sigma_{\min}(A(t_1)) \geq C_4 \xi^{\frac{1}{\kappa} - 1} \frac{\sqrt{\sigma_d}}{c_1^2 d}$.

Proposition 8 *$\forall 0 < \alpha < \alpha' \leq \frac{\sqrt{2}}{2}$, if we choose $\varepsilon = \frac{1}{\xi}\sqrt{\frac{\sigma_d}{c_1^2 d}}$, where*

$$\xi \geq \max\left\{2, \frac{\sqrt{\alpha}}{\sqrt{\alpha'} - \sqrt{\alpha}}\frac{3d}{\sqrt{\kappa}}, C_3\left(\frac{3\sqrt{\alpha'}d^2c_1^2}{C_1\sqrt{\kappa}}\right)^{\kappa}\right\},$$

we have $\sigma_{\min}(A(t_1)) \geq C_4 \xi^{\frac{1}{\kappa}-1}\frac{\sqrt{\sigma_d}}{c_1^2 d}$.

B.2. Stage 2

Stage 2 is defined to be $t \in [t_1, t_2]$, where t_2 is a parameter we will define later. In this stage, we hope to prove 1) the smallest eigenvalue of $AA^\top$ becomes a constant multiple of σ_d, 2) the loss function will be smaller than $\frac{\sigma_d^2}{9}$, 3) $\|B\|$ is still pretty small. The intuition is that if we define $S := AA^\top$, we have

$$\dot{S} = PS + SP - QBA^\top + AB^\top Q, \tag{20}$$

where $P := \Sigma - AA^\top + BB^\top$ and $Q := AB^\top - BA^\top$. Define s as the smallest eigenvalue of S, then

$$\dot{s} \geq 2\sigma_d s - 2s^2 - 4\sigma_{\max}(S)\|B\|^2. \tag{21}$$

If we can prove $\sigma_{\max}(S)$ cannot be greatly larger than $2\sigma_1$ and $\|B\|^2$ is extremely small, we can observe that s becomes a constant fraction of σ_d at a fast speed.

Now, at the beginning of the second stage, $s(t_1) \geq C_4^2 \xi^{-2+\frac{2}{\kappa}}\frac{\sigma_d}{c_1^4 d^2}$, while

$$\begin{aligned} \|B(t_1)\|^2 &\leq \left(e^{-2\sigma_d t_1}\right)^{(1-\alpha')}\varepsilon^2 c_1^2 d^2 \\ &\leq C_1^{-\frac{2(1-\alpha')}{\kappa}}\xi^{-2-\frac{2(1-\alpha')}{\kappa}} d\sigma_d \\ &\leq \frac{1}{C_1^2}\cdot \xi^{-2-\frac{2(1-\alpha')}{\kappa}} d\sigma_d, \end{aligned} \tag{22}$$

which is much smaller than $s(t_1)$.

To give an upper bound for $\|B\|^2$ in the second stage, we also need to lower bound the smallest eigenvalue of P according to inequality (10). For convenience, we call P the complementary matrix. Then

$$\begin{aligned} \dot{P} &= -(AA^\top + BB^\top)P - P(AA^\top + BB^\top) \\ &\quad -(AB^\top + BA^\top)Q + Q(AB^\top + BA^\top). \end{aligned} \tag{23}$$

The first line implies something like shrinking, while the second line is extremely small. All these observations suggest that the minimal eigenvalue of P cannot be a huge negative number.

With all preparation above, we can prove a fast convergence speed at stage 2. Please see the subsections for details.

B.2.1. BOUND WITH ASSUMPTIONS

We first make two assumptions, and we will verify these two assumptions later.

- $s(t) \geq s(t_1)$ in the second stage. Also, $AA^\top \preceq 2\sigma_1 I$.
- There exists $C_5 > 0$, such that during the second stage, $\|B\|^2 \leq C_5 \xi^{-2} d\sigma_d$.

Denote p as the smallest eigenvalue of P. Then, once $p \leq 0$, we have the following bound:

$$\begin{aligned} \dot{p} &\geq -2C_4^2 \xi^{-2+\frac{2}{\kappa}} \frac{\sigma_d}{c_1^4 d^2} p - 16\sigma_1 \|B\|^2 \\ &\geq -2C_4^2 \xi^{-2+\frac{2}{\kappa}} \frac{\sigma_d}{c_1^4 d^2} p - 16\sigma_1 C_5 \xi^{-2} d\sigma_d. \end{aligned}$$

Notice that initially $p(t_1) > 0$. Solving the ODE and we have $\exists C_6 > 0, p \geq -C_6 \xi^{-\frac{2}{\kappa}} d^3 \sigma_1$.

Now (21) implies $\dot{s} \geq -2s^2 + 2\sigma_d s - 8C_5 \xi^{-2-\frac{2(1-\alpha')}{\kappa}} d\sigma_1 \sigma_d$. Suppose $x_1 \geq x_2$ are two roots of quadratic function $x^2 - \sigma_d x + 4C_5 \xi^{-2-\frac{2(1-\alpha')}{\kappa}} d\sigma_1 \sigma_d$. Then $\forall \alpha'' \in (0,1)$, we could choose $\xi = \xi_1 := O((\text{poly}(\kappa, d))^{\frac{\kappa}{2}})$ such that $0 < x_2 \leq 4C_5 \xi^{-2-\frac{2(1-\alpha')}{\kappa}} d\sigma_1 < \alpha'' s(t_1)$, thus $\sigma_d - \alpha'' s(t_1) \leq x_1 \leq \sigma_d$. Solving the ODE and we have $s(t_1 + t) \geq x_1 - \frac{x_1 - x_2}{e^{\sigma_d t + c} + 1}$ where $c = \ln \frac{s(t_1) - x_2}{x_1 - s(t_1)} = O\left(\ln\left(\xi^{-2+\frac{2}{\kappa}} \frac{1}{d^2}\right)\right) = -O(\kappa \ln \kappa d)$. By choosing $t' = O\left(\frac{\kappa \ln \kappa d}{\sigma_d}\right)$, we can make sure $\forall t \geq t'$ in the second stage, $s(t_1 + t) \geq \frac{\sigma_d}{2}$.

Finally, based on the lower bound of the eigenvalue of $AA^\top$, we can give an upper bound of maximal eigenvalue of P, say $p_{\max}$, when $t \geq t_1 + t'$:

$$\dot{p}_{\max} \leq -\sigma_d p_{\max} + 16C_5 \xi^{-2} d\sigma_1 \sigma_d. \tag{24}$$

This implies that by choosing $t = t_1 + t' + t'' = O\left(\frac{\kappa \ln \kappa d}{\sigma_d}\right)$, $p_{\max}(t) \leq \frac{\sigma_d}{4\sqrt{d}}$. Then we can easily bound $\|f\|^2$ by $\|P\|^2 + \|Q\|^2 \leq \frac{\sigma_d^2}{16} + 4\|A\|^2\|B\|^2 \leq \frac{\sigma_d^2}{9}$. We define $t_2 := t_1 + t' + t''$ as the end of the second stage.

B.2.2. VERIFY THE ASSUMPTIONS

The assumption on A is straightforward, as we have verified the right hand side of (21) is always non-negative once $s(t) \leq s(t_1)$ by simply choosing $\xi \geq \xi_1$. The upper bound of $AA^\top$ simply comes from the lower bound of P and the upper bound of B.

The assumption on B is based on the fact that we have $\dot{b} \leq -2pb$, where $b := \|B\|^2$, according to (10). Hence $b(t_1 + t) \leq e^{C_6 \xi^{-\frac{2}{\kappa}} d^3 \sigma_1 t} \frac{1}{C_1^2} \cdot \xi^{-2-\frac{2(1-\alpha')}{\kappa}} d\sigma_d$. Because $t_2 = O\left(\frac{\kappa \ln \kappa d}{\sigma_d}\right)$, we can choose $\xi = (\text{poly}\,(\kappa, d))^\kappa$ to ensure the existence of the universal constant C_6.

B.3. Stage 3

The last stage is the simplest part of the proof. We only need to bound $\ell := \|\Sigma - UV^\top\|$ and $\|B\|^2$ simultaneously.

Suppose initially $\ell(t_2) \leq \frac{1}{3}\sigma_d$, and $\|B\|^2 \leq \rho\sigma_d$. Then $\|P\|^2 \leq \|P\|^2 + \|Q\|^2 = \|\Sigma - UV^\top\|^2 \leq \frac{\sigma_d^2}{9}$. Hence $AA^\top = \Sigma - P + BB^\top$ implies the smallest singular value of $AA^\top$ is lower bounded by $\frac{2}{3}\sigma_d$. Then, the smallest singular value of $U = A+B$ and $V = A-B$ is lower bounded by $\frac{\sqrt{\sigma_d}}{2}$, if $\sqrt{\rho} \leq \sqrt{\frac{2}{3}} - \frac{1}{2}$. Hence we could draw a conclusion that

$$\begin{aligned} \|\nabla f(U,V)\|^2 &= \|(\Sigma - UV^\top)V\|^2 + \|(\Sigma - UV^\top)^\top U\|^2 \\ &\geq \frac{\sigma_d}{2} f(U,V), \end{aligned}$$

which is exactly Polyak-Łojasiewicz inequality. With a little abuse of notation, we use θ to denote this integral curve. Then

$$\|\dot{\theta}(t_0)\|^2 \geq \frac{\sigma_d}{2}\left(f(\theta(0)) - \int_0^{t_0} \|\dot{\theta}(t)\|^2\right) \mathrm{d}t. \tag{25}$$

Hence $\ln \frac{1}{f(\theta(t))}\Big|_0^{t_0} \geq \frac{\sigma_d t}{2}$, i.e. $f(\theta(t)) \leq e^{-\frac{\sigma_d t}{2}} f(\theta(0))$.

In other words, $\|\Sigma - U(t_2+t)V^\top(t_2+t)\|^2 \leq e^{-\frac{\sigma_d t}{2}}\|\Sigma - U(t_2)V^\top(t_2)\|^2$. For simplicity, we denote $\|\Sigma - U(t_2)V^\top(t_2)\|^2$ by F_0.

Besides, $\|\dot{B}\|^2 \leq 2\|P\|\|B\|^2$, i.e. $\|B(t_2+t_0)\|^2 \leq e^{\int_0^{t_0} 2\|P(t)\|\mathrm{d}t}\|B(t_2)\|^2$, which is bounded by $e^{\frac{8}{\sigma_d}\sqrt{F_0}}B(t_2)\|^2 \leq e^{\frac{8}{3}}\|B(t_2)\|^2$. Hence, we only need to make $\|B(t_2)\|^2 \leq \frac{\rho\sigma_d}{e^{\frac{8}{3}}}$, which is an obvious condition (because $\|B\|$ is $O\left(\frac{1}{\xi}\right)$).

Notice that we have proved linear convergence. To sum up, by factoring what has been discussed above in, we could draw a conlusion that the total complexity is $O\left(\frac{\kappa}{\sigma_d}\ln(\kappa d) + \frac{\ln\frac{1}{\epsilon}}{\sigma_d}\right)$, here κ is the condition number, σ_d is the smallest singular value of Σ, d is the dimension of the matrix, and ϵ means that our output satisfies $f(U,V) \leq \epsilon\|\Sigma\|^2$.