Global Convergence Rate of Gradient Flow for Asymmetric Matrix Factorization

Tian Ye, Simon S. Du

Institute for Interdisciplinary Information Sciences, Tsinghua University
Paul G. Allen School of Computer Science & Engineering, University of Washington

yet17@mails.tsinghua.edu.cn
ssdu@cs.washington.edu

Resumo

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.

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 1.1 *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*

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

2. Problem Setup

Let $\Sigma\in\mathbb{R}^{d\times d}$ be a non-singular matrix with singular value $\sigma_1\ge\cdots\ge\sigma_d>0$, 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 $\to 0$), a.k.a., gradient flow. More precisely, we deal with the ODE

$$\dot{U} = -\frac{\partial f}{\partial U} = (\Sigma - UV^\top)V; \quad (1)$$
$$\dot{V} = -\frac{\partial f}{\partial V} = (\Sigma - UV^\top)^\top U. \quad (2)$$

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 next section 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 $\sigma_1\ge\cdots\ge\sigma_d$ 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. 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

$$\dot{S} = (\Sigma - S)S + S(\Sigma - S). \quad (3)$$

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 (3), where $\mathcal{D}'\subseteq\mathbb{R}\times\mathcal{M}_{\text{sym}}$.

Here are two useful lemmas.

Lemma 3.1 *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 3.1, 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 3.2 *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 3.2, we have $\forall t\ge 0$, $\vartheta(t,\sigma_d(S)I)\preceq\vartheta(t,S)$. 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 3.3 *Suppose* α_1 *and* α_d *are the largest and smallest singular value of* U_0*. Then* $\forall t>0$*, we have*

$$\text{diag}\left(\frac{1-\frac{\sigma_i}{\alpha_1^2}}{e^{2\sigma_i t}+\frac{\sigma_i}{\alpha_1^2}-1}\right)_{i\in[d]}\preceq P(t)\preceq\text{diag}\left(\frac{1-\frac{\sigma_i}{\alpha_d^2}}{e^{2\sigma_i t}+\frac{\sigma_i}{\alpha_d^2}-1}\right)_{i\in[d]} \quad (4)$$

where $P(t):=\Sigma - U(t)U(t)^\top$*.*

4. Asymmetric Case

In asymmetric case, we divide the whole process into three stages.

- In the first stage, the initial matrices are quite small, and we will prove that $A(t):=\frac{\theta(t,U)+\theta(t,V)}{2}$ 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.

4.1 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 3.2, we could give exact analytical bound on singular values of $\overline{A}$.

Lemma 4.1 *If we assume* $A(0)A^\top(0)\preceq\sigma_d I$*, we have*

$$\sqrt{\frac{\sigma_d e^{2\sigma_d t}}{e^{2\sigma_d t}+\frac{c_1^2 d\sigma_d}{\varepsilon^2}-1}}\le\sigma_d(\overline{A})\le\sigma_1(\overline{A})\le\sqrt{\frac{\sigma_1 e^{2\sigma_1 t}}{e^{2\sigma_1 t}+\frac{\sigma_1}{\varepsilon^2 c_1^2 d}-1}} \quad (5)$$

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

$$t_1:=\frac{\ln\left(\frac{\sigma_1}{\varepsilon^2 c_1^2 d}-1\right)+\ln\frac{\kappa-\alpha}{\kappa}}{2\sigma_1}, \quad (6)$$

where $\kappa:=\frac{\sigma_1}{\sigma_d}$. We will prove later that $\sigma_d(\overline{A})$ is of order $\varepsilon^{1-\frac{1}{\kappa}}$. We choose such kind of small t_1 is to prevent the largest singular value of A from being too large, so as to ensure the monotonousness of $\|B\|$ where $B=\frac{U-V}{2}$, by using

$$\|\dot{B}\|^2\le -2\left\langle B,(\Sigma - AA^\top + BB^\top)B\right\rangle. \quad (7)$$

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

$$\phi(0,M) = M; \quad (8)$$
$$\phi(t,\phi(s,M)) = \phi(t+s,M); \quad (9)$$
$$\left.\frac{\partial\phi}{\partial t}\right|_{(t_0,M)} = F(\phi(t_0,M)), \quad (10)$$

where $F(M):=(\Sigma - MM^\top)M$. We could observe that velocity vector of A and ϕ are close to each other. Hence we can bound $A(t)-\overline{A}(t)$ by

$$\int_0^t\left.\frac{\partial\phi}{\partial M}\right|_{(t-s,A(s))}\circ\left(\dot{A}(s)-F(A(s))\right)\mathrm{d}s \quad (11)$$
$$=\int_0^t\left.\frac{\partial\phi}{\partial M}\right|_{(t-s,A(s))}\circ\left(BB^\top A - AB^\top B + BA^\top B\right)(s)\mathrm{d}s \quad (12)$$

which is relatively small, since B is extremely small.

Now, we can make a brief summary about the first stage. At the end of stage 1, the smallest singular value of A becomes $\Theta\left(\varepsilon^{1-\frac{1}{\kappa}}\right)$, and the norm of B is $o(\varepsilon)$.

4.2 Stage 2

Stage 2 is defined to be $t\in[t_1,t_2]$, where t_2 is a parameter we will define later. During the second stage, we mainly consider about three equations.

$$\dot{S} = PS + SP - QBA^\top + AB^\top Q, \quad (13)$$
$$\|\dot{B}\|^2\le -2\left\langle B,(\Sigma - AA^\top + BB^\top)B\right\rangle, \quad (14)$$
$$\dot{P} = -(AA^\top + BB^\top)P - P(AA^\top + BB^\top) - (AB^\top + BA^\top)Q + Q(AB^\top + BA^\top), \quad (15)$$

where $S:=AA^\top$, $P:=\Sigma - AA^\top + BB^\top$ and $Q:=AB^\top - BA^\top$. Here (13) gives the lower bound of the smallest singular value of S, (14) gives upper bound of $\|B\|$ and (15) gives the upper bound for the largest singular value of P and the lower bound for the smallest singular value of P. We need to bound these quantities concurrently. A brief summary of stage 2 is that $f=f(U,V)$ and B becomes relatively small, and we will not consider about S anymore, since its lower bound can be derived by f and U.

4.3 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 $\ell(t_2)\le\frac{1}{3}\sigma_d$, and $\|B\|^2\le\rho\sigma_d$. Then $\|P\|^2\le\|P\|^2+\|Q\|^2=\|\Sigma - UV^\top\|^2\le\frac{\sigma_d^2}{9}$. 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 $\sqrt{\rho}\le\sqrt{\frac{2}{3}}-\frac{1}{2}$. Hence we could draw a conclusion that

$$\|\nabla f(U,V)\|^2 = \|(\Sigma - UV^\top)V\|^2 + \|(\Sigma - UV^\top)^\top U\|^2 \ge\frac{\sigma_d}{2}f(U,V),$$

which is exactly Polyak-Łojasiewicz inequality. Hence we have immediately $\|\Sigma - U(t_2+t)V^\top(t_2+t)\|^2\le e^{-\frac{\sigma_d t}{2}}\|\Sigma - U(t_2)V^\top(t_2)\|^2$. For simplicity, we denote $\|\Sigma - U(t_2)V^\top(t_2)\|^2$ by F_0.

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