

Understanding Incremental Learning with Closed-form Solution to Gradient Flow on Overparameterized Matrix Factorization

Hancheng Min and René Vidal

Abstract—Many theoretical studies on neural networks attribute their excellent empirical performance to the implicit bias or regularization induced by first-order optimization algorithms when training networks under certain initialization assumptions. One example is the incremental learning phenomenon in gradient flow (GF) on an overparameterized matrix factorization problem with small initialization: GF learns a target matrix by sequentially learning its singular values in decreasing order of magnitude over time. In this paper, we develop a quantitative understanding of this incremental learning behavior for GF on the symmetric matrix factorization problem, using its closed-form solution obtained by solving a Riccati-like matrix differential equation. We show that incremental learning emerges from some time-scale separation among dynamics corresponding to learning different components in the target matrix. By decreasing the initialization scale, these time-scale separations become more prominent, allowing one to find low-rank approximations of the target matrix. Lastly, we discuss the possible avenues for extending this analysis to asymmetric matrix factorization problems.

I. INTRODUCTION

Why do deep neural networks [1]–[3] work so well in practice? At the center of this mystery lies the fact that practical neural networks are typically *overparameterized*, i.e., their number of parameters is several orders of magnitude larger than the number of available training examples. As a consequence, there are infinitely many parameter choices that can perfectly fit the data. However, different parameters have different generalization properties, and only a few choices lead to networks that make correct predictions for new samples. Behind the success of deep learning lies the ability of training algorithms, such as gradient descent, to find those networks that generalize well, which over past years has motivated many theoretical studies on such an *implicit bias* [4] of learning algorithms.

An interesting one is the *simplicity bias* of gradient descent or gradient flow under certain initial conditions, typically random initialization with small variance, through which the learned network model possesses certain notions of low complexity. Depending on the network architecture, such

notions range from classic ones, such as low-rankness [5]–[8] or sparsity [9]–[11] of the network weights, to complexity measures in some function space [12]. In many scenarios, the analyses for such simplicity biases often call for theoretical understandings of the *incremental learning* [6], [7], [13] phenomenon, which describes the evolution of the network model during training as sequential learning of components of the underlying target ground truth function (suppose there is one) in some canonical decompositions, without learning any spurious components that overfit to the training data.

One important testbed for understanding incremental learning is the *matrix factorization* [14] problem, studied in the context of linear regression [6], [13], [15]–[17], matrix sensing [5], [7], [18], [19], and recently the low-rank adaptation [20], [21]. Under random initialization with small variance, training with gradient descent/flow on the factorized model finds a stationary point that corresponds exactly (noiseless) or approximately (noisy) to the underlying ground truth matrix, following a particular dynamic trajectory. During training, the factorized model grows its singular components to fit/learn the ground truth matrix in a sequential manner such that the growth happens for one singular component at a time to fit one component in the ground truth, and the ground truth component with the largest singular values get learned first, then followed by the second largest and so on so forth. As such, the training algorithm finds the ground truth by every time *incrementally* learning one of its components.

A body of theoretical work on incremental learning in matrix factorization is motivated by some of the early works [5], [13] that investigate the implicit bias of deep learning algorithms. One initial line of attempts studies *closed-form* solutions: If one carefully initializes the singular vectors of the factorized model to align with those of the ground truth, then the training dynamics are decoupled into multiple scalar dynamics, each corresponding to one principal component in the ground truth [6], [13], [15]. In this case, understanding incremental learning reduces to studying the time-scale separation among these scalar dynamics. However, these results are achieved under very restrictive conditions on initialization.

To analyze this phenomenon under general initial conditions, recent work focuses on a more fine-grained *trajectory analysis* [7], [19] that introduces some orthogonal decomposition of the parameter space into signal and error subspaces based on the ground truth. This analysis shows that, at certain time epoch, the components within the signal subspace learn some principal components of the ground

This work was done when H. Min was a postdoc at the University of Pennsylvania. The authors acknowledge the support of the NSF under grants 2031985 and 2212457, the Simons Foundation under grant 814201, and the ONR MURI Program under grant 503405-78051. H. Min thanks Arthur C. B. de Oliveira for the insightful discussion during the early stage of this work.

H. Min is with the Institute of Natural Sciences (INS) and the School of Mathematical Sciences (SMS) at the Shanghai Jiao Tong University. hanchengmin@sjtu.edu.cn.

R. Vidal is with the Department of Electrical and Systems Engineering (ESE) and the Department of Radiology in the Perelman School of Medicine at the University of Pennsylvania. vidalr@upenn.edu

truth, together with upper bounds on the growth within the error subspace. While the trajectory analysis produces the most general results [7] for matrix factorization, it does not fully answer many quantitative questions. For example, how small does the variance of the random initialization need to be in order for incremental learning to happen? If one seeks one of the low-rank approximations of the ground truth, within what time interval does the algorithm need to stop? These questions have become increasingly important alongside the growing interest (for example, [22]) in utilizing the incremental mechanism to obtain low-complexity models, as opposed to the classic regularized approach.

In this paper, we extend the closed-form solution-based line of work to gradient flow on matrix factorization under general initial conditions, facilitating a rich quantitative understanding of the incremental phenomenon. Our theorems aim to exactly address several questions previously mentioned by showing the precise dependence of incremental learning on the variance of the random initialization and providing time intervals within which the factorized model serves as a good low-rank approximation of the ground truth. Although our results are primarily for symmetric matrix factorization, we outline the path toward extending them to the asymmetric matrix factorization problems, opening up new opportunities to utilize closed-form solutions to study a broader class of problems.

The organization of this paper is as follows. In Section II, we introduce the gradient flow dynamics on the symmetric matrix factorization problem, the primary focus of our theoretical results, and we discuss the closed-form solution of the factorized model to the gradient flow. This forms the basis of our main results in Section III on the incremental learning in the symmetric matrix factorization problem. Then, in Section IV, we conclude by briefly discussing the path toward extending current results to the asymmetric matrix factorization problems.

Notations: For an $n \times m$ matrix A , we let $\|A\|$ and $\|A\|_F$ denote the spectral and Frobenius norm of A , respectively. We write $A \succeq 0$ ($A \preceq 0$) when A is positive semi-definite (negative semi-definite). Then, we let $\text{diag}\{a_i\}_{i=1}^n$ be a diagonal matrix with a_i being its i -th diagonal entry.

II. PRELIMINARIES: GRADIENT FLOW ON SYMMETRIC MATRIX FACTORIZATION

The problem of our main interest in this paper is the following *symmetric matrix factorization*:

$$\min_{U \in \mathbb{R}^{n \times r}} \mathcal{L}(U) = \frac{1}{4} \|Y - UU^\top\|_F^2, \quad (1)$$

where $Y \in \mathbb{R}^{n \times n}$ with $Y \succeq 0$ is the *target matrix* to be factorized. Although one may find solving (1) easy by simply taking the SVD of Y , the goal of this paper is rather to understand training trajectories when one seeks to solve (1) by gradient flow (GF) on $\mathcal{L}(U)$, i.e.,

$$\dot{U} = -\nabla_U \mathcal{L}(U) = (Y - UU^\top)U, \quad U(0) = U_0, \quad (2)$$

from some initial condition U_0 . More specifically, as stated in the introduction, we are interested in the incremental learning

phenomenon when $U(t)U^\top(t)$ learns the singular components of Y sequentially along the GF trajectory. Therefore, it suffices to study the induced dynamic evolution of the factorized model $U(t)U^\top(t) := W(t)$ from (2), which is

$$\dot{W} = \dot{U}U^\top + U\dot{U}^\top = WY + YW - 2W^2, \quad W(0) = U_0U_0^\top. \quad (3)$$

(3) is a special form of matrix Riccati differential equations, thus can be analyzed from its closed-form solution.

A. Closed-form solution

The closed-form solutions to matrix Riccati differential equations are well-studied, for example, in [23], [24]. We thus have the following result regarding the solution to (3)

Proposition 1. *If Y has K non-zero singular values and the full SVD of $Y \succeq 0$ is $\Phi \begin{bmatrix} \Sigma_Y & 0 \\ 0 & 0 \end{bmatrix} \Phi^\top$, then the matrix Riccati differential equation*

$$\dot{W} = WY + YW - 2W^2, \quad W(0) = W_0 \succeq 0 \quad (4)$$

has unique solution of the following:

$$W(t) = \Phi S(t) \tilde{W}_0 \left(I_n + G(t) \tilde{W}_0 \right)^{-1} S^\top(t) \Phi^\top, \quad (5)$$

where $\tilde{W}_0 = \Phi^\top W_0 \Phi$ and

$$G(t) = \begin{bmatrix} \Sigma_Y^{-1}(e^{2\Sigma_Y t} - I_K) & 0 \\ 0 & 2I_{n-K}t \end{bmatrix}, \quad S(t) = \begin{bmatrix} e^{\Sigma_Y t} & 0 \\ 0 & I_{n-K} \end{bmatrix}. \quad (6)$$

Proof. From [23], [24], one knows that the solution to general matrix Riccati equation

$$\dot{P} = AP + PA^\top - PRP + Q, \quad P(0) = P_0, \quad (7)$$

is given by $P(t) = X_2(t)X_1^{-1}(t)$, where $\{X_1(t), X_2(t)\}$ are solution to an LTI system

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \end{bmatrix} = \begin{bmatrix} -A^\top & R \\ Q & A \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad X_1(0) = I_n, X_2(0) = P_0, \quad (8)$$

To apply this, let $P = \Phi^\top W \Phi$, $A = \Sigma_Y$, $Q = 0$ and $R = I_n$, by solving (8) we have (an equivalent expression for (5))

$$\underbrace{\Phi^\top W(t) \Phi}_{P(t)} = \underbrace{S(t) \Phi^\top W_0 \Phi}_{X_2(t)} \cdot \underbrace{\left(S^{-1}(t) (I_n + G(t) \Phi^\top W_0 \Phi) \right)^{-1}}_{X_1(t)}, \quad (9)$$

where $G(t), S(t)$ are defined in (6). Notice that $\forall t$, all eigenvalues of $G(t)W_0$ are real and non-negative since both $G(t)$ and W_0 are positive semi-definite, then $X_1(t)$ is invertible $\forall t$, ensuring (9) is the unique solution [24, Corollary 1]. \square

B. Additional settings

With the closed-form solution available, one can study this GF under any initial condition $W_0 = U_0U_0^\top \succeq 0$. However, different types of initialization (for example, random v.s. deterministic initialization) may require different analyses. Due to space constraints, we opt to discuss the case of *overparametrized* matrix factorization under deterministic initialization with varying scales.

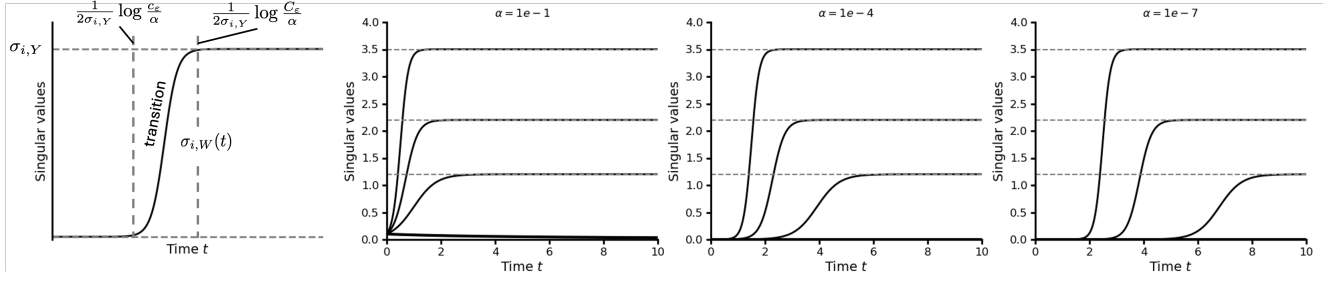


Fig. 1. Incremental learning in symmetric matrix factorization. Under small spectral initialization, (left plot) the dynamic evolution of $\sigma_{i,W}(t)$ for learning singular component $\sigma_{i,Y}$ in target matrix Y has a sharp transition phase where its value grows rapidly from approximately zero to one that is close to the target $\sigma_{i,Y}$. (Right three plots) As the initialization scale α decreases, the transition phases for different target values get further separated, and low-rank approximations of Y emerge along the GF trajectory. Similar dynamic evolutions of $W(t)$ still exist under general initialization, as we show in Theorem 4.

As we described in the introduction, the incremental learning phenomenon generally happens under initialization with a small scale. To isolate this critical effect of the scale/norm of the initialization, it is convenient to write our initial condition as $U(0) = U_0 = \alpha^{1/2} \bar{U}_0$ for some *initialization scale* $\alpha > 0$ and *initialization shape* $\bar{U}_0 \succeq 0$. Our analysis

- assumes a fixed shape \bar{U}_0 and describes how the scale α affects the incremental learning phenomenon;
- In addition, we assume $W_0 = \alpha \bar{U}_0 \bar{U}_0^\top$ is invertible. This necessarily requires that the problem (1) is *over-parametrized* [15], i.e. $r \geq n$, which imply that the optimal value of (1) is always zero (target is exactly factorized) regardless of the target Y .

III. MAIN RESULTS: INCREMENTAL LEARNING IN SYMMETRIC MATRIX FACTORIZATION

In this section, we discuss incremental learning in symmetric factorization problems. First, we revisit the GF under spectral initialization with a small scale and refine the existing asymptotic results [6] to a non-asymptotic one. Then, we show that a similar quantitative characterization of incremental learning holds for general initialization. These main results follow the problem settings stated in the previous section, and afterward, we discuss how to adapt them to other settings.

A. Incremental learning under spectral initialization

Give the GF (2) with the SVD of the target matrix $Y = \Phi \Sigma_Y \Phi^\top$, the spectral initialization of U is any U_0 that has an SVD of the form $U_0 = \alpha^{1/2} \Phi \Sigma_{U_0} V_{U_0}^\top$. Recall that $\alpha > 0$ is the initialization scale. Under a spectral initialization, the induced ode on W (3) becomes:

$$\dot{W} = WY + YW - 2W^2, \quad W(0) = \alpha \Phi \Sigma_{U_0}^2 \Phi^\top, \quad (10)$$

Through a change of variables $\tilde{W} = \Phi^\top W \Phi$, the dynamics on \tilde{W} is given by

$$\dot{\tilde{W}} = \tilde{W} \Sigma_Y + \Sigma_Y \tilde{W} - 2\tilde{W}^2, \quad \tilde{W}(0) = \alpha \Sigma_{U_0}^2, \quad (11)$$

whose solution $\tilde{W}(t)$ is diagonal $\forall t$, and the dynamics of diagonal entries are decoupled from each other. Indeed,

examining the closed-form solution in Proposition 1 under a spectral initialization, we have the following:

Corollary 2. Suppose $\Sigma_Y = \text{diag}\{\sigma_{i,Y}\}_{i=1}^n$. Under some spectral initialization $U(0) = \alpha^{1/2} \Phi \Sigma_{U_0} V_{U_0}^\top$, the solution $U(t)$ to the GF dynamics (2) induces $W(t) = U(t)U^\top(t)$ of the form $W(t) = \Phi \text{diag}\{\sigma_{i,W}(t)\}_{i=1}^n \Phi^\top$ with

$$\sigma_{i,W}(t) = \frac{\alpha \sigma_{i,Y} \sigma_{i,0} e^{2\sigma_{i,Y} t}}{\sigma_{i,Y} + \alpha \sigma_{i,0} (e^{2\sigma_{i,Y} t} - 1)}, \quad \text{if } \sigma_{i,Y} \neq 0; \quad (12)$$

$$\sigma_{i,W}(t) = \frac{\alpha \sigma_{i,0}}{1 + 2\alpha \sigma_{i,0} t}, \quad \text{if } \sigma_{i,Y} = 0, \quad (13)$$

where $\sigma_{i,0} = [\Sigma_{U_0}^2]_{ii} \geq 0, \forall i$.

Spectral initialization offers a simplified analysis of the GF trajectory. Under the induced dynamics (3), we expect the factorized model $W(t)$ to converge to the target matrix Y . Indeed, for each i , (12) is the learning trajectory for a non-zero singular mode $\sigma_{i,Y}$ of Y , and it is easy to see that as long as $\sigma_{i,0} \neq 0$, we have $\lim_{t \rightarrow \infty} \sigma_{i,W}(t) = \sigma_{i,Y}$; (13) is the learning trajectory for a zero singular mode $\sigma_{i,Y}$ of Y , and we always have $\lim_{t \rightarrow \infty} \sigma_{i,W}(t) = 0$. Therefore, if $\sigma_{i,0} \neq 0, \forall i$ with $\sigma_{i,Y} \neq 0$, one has $\lim_{t \rightarrow \infty} W(t) = Y$, thereby learning the target matrix. Note that a similar condition is also necessary in the general case. That is, learning the exact Y requires $\Phi_Y^\top W(0) \Phi_Y$ to have full rank, where columns of Φ_Y form an orthonormal basis of the range space of Y .

As shown in the previous discussion, the asymptotic behavior of these scalar dynamics (12)(13) are easy to study, concerning whether $W(t)$ can successfully learn the target matrix Y . Now, we turn to the questions regarding the *transient behavior* under small initialization with scale $\alpha \ll 1$. Notice that when $0 < \alpha \sigma_{i,0} < \sigma_{i,Y}$, $\sigma_{i,W}(t)$ is strictly monotonically increasing w.r.t. time t . Our following discussions assume α is sufficiently small so that $\sigma_{i,W}(t)$ is monotonically increasing.

For every $\sigma_{i,W}(t)$ following (12) that learns a non-zero mode $\sigma_{i,Y}$ of Y , we have $\sigma_{i,W}(0) = \alpha \sigma_{i,0}$, which is close to zero for sufficiently small α , and we know asymptotically $\lim_{t \rightarrow \infty} \sigma_{i,W}(t) = \sigma_{i,Y}$. Therefore, each $\sigma_{i,W}(t)$ must learn its target $\sigma_{i,Y}$ through some transitional phase during which the value $\sigma_{i,W}(t)$ grow from a small value close to $\alpha \sigma_{i,0}$

to a value close to $\sigma_{i,Y}$. If there is a *time-scale separation* between these transitional phases for different i , then we expect that at some intermediate time along GF, a subset of $\sigma_{i,W}(t)$ has learned their corresponding target $\sigma_{i,Y}$ while the rest stays close to zero, from which the resulting $W(t)$ serves as some low-rank approximation of Y .

The above discussion suggests a closer look at the transient behavior in the solution (12) when the initialization scale is small: For some $c, C > 0$ such that $c\sigma_{i,0} \ll \sigma_{i,Y}$ and $C\sigma_{i,0} \gg \sigma_{i,Y}$, $\forall i$. Then we see that for every i (for any $\alpha \leq c$),

$$\sigma_{i,W} \left(\frac{1}{2\sigma_{i,Y}} \log \frac{c}{\alpha} \right) = \frac{c\sigma_{i,Y}\sigma_{i,0}}{\sigma_{i,Y} + (c - \alpha)\sigma_{i,0}} \simeq c\sigma_{i,0}, \quad (14)$$

$$\sigma_{i,W} \left(\frac{1}{2\sigma_{i,Y}} \log \frac{C}{\alpha} \right) = \frac{C\sigma_{i,Y}\sigma_{i,0}}{\sigma_{i,Y} + (C - \alpha)\sigma_{i,0}} \simeq \sigma_{i,Y}, \quad (15)$$

suggesting a sharp transition in $\sigma_{i,W}(t)$ from a small value $c\sigma_{i,0}$ to one close to its target value $\sigma_{i,Y}$ at a time that scales as $\Theta \left(\frac{1}{\sigma_{i,Y}} \log \frac{1}{\alpha} \right)$. Notably, the transition time depends inverse-proportionally on $\sigma_{i,Y}$ (See Figure 1). Thus, this time-scale separation can result in singular values of Y being learned one by one, with larger ones learned first and smaller ones later, which is exactly an incremental learning phenomenon. Based on these discussions, the conditions to achieve this are:

- Singular values $\sigma_{i,Y}$ of Y are distinct;
- The scale α is sufficiently small such that the transitional phases $\left[\frac{1}{2\sigma_{i,Y}} \log \frac{c}{\alpha}, \frac{1}{2\sigma_{i,Y}} \log \frac{C}{\alpha} \right]$ are non-overlapping.

Formally, we have the following theorem:

Theorem 3 (Incremental learning under small spectral initialization). *Suppose the K non-zero singular values $\sigma_{1,Y}, \sigma_{2,Y}, \dots, \sigma_{K,Y}$ of Y are distinct and ordered in decreasing order. Let the spectral initialization $U(0) = \alpha^{1/2} \Phi \Sigma_{U_0} V_{U_0}^\top$ have a initialization shape $\Sigma_{U_0}^2 = \text{diag}\{\sigma_{i,0}\}_{i=1}^d$ with $\sigma_{i,0} > 0, \forall 1 \leq i \leq K$. Given any error tolerance $0 < \varepsilon \leq \sigma_{K,Y}$, let $c_\varepsilon = \frac{\varepsilon}{\max_i \sigma_{i,0}}$, and $C_\varepsilon = \frac{\sigma_{1,Y}^2}{\varepsilon \min_{1 \leq i \leq K} \sigma_{i,0}}$. Suppose α is sufficiently small such that $\alpha \leq c_\varepsilon$ and that $\frac{-\log \alpha + \log c_\varepsilon}{-\log \alpha + \log C_\varepsilon} > \max_{1 \leq k \leq K-1} \frac{\sigma_{k+1}}{\sigma_k}$, then the time intervals*

$$\mathcal{I}_k := \left[\frac{1}{2\sigma_{k,Y}} \log \frac{C_\varepsilon}{\alpha}, \frac{1}{2\sigma_{k+1,Y}} \log \frac{c_\varepsilon}{\alpha} \right], \quad 1 \leq k \leq K \quad (16)$$

are all non-empty (we have let $\frac{1}{0} := \infty$), and the $W(t)$ in Corollary 2 satisfies that $\forall 1 \leq k \leq K$,

$$\|W(t) - \hat{Y}_k\| \leq \varepsilon, \quad \forall t \in \mathcal{I}_k, \quad (17)$$

where $\hat{Y}_k := \arg \min_{\text{rank}(Z)=k} \|Y - Z\|_F$ is the best rank- k approximation of Y .

Proof. We have assumed small α such that $\frac{-\log \alpha + \log c_\varepsilon}{-\log \alpha + \log C_\varepsilon} > \max_{1 \leq k \leq K-1} \frac{\sigma_{k+1}}{\sigma_k}$, which implies that for every $1 \leq k \leq K$, $\mathcal{I}_k \neq \emptyset$. For a fixed k , choose any t within this interval. Since $\Phi^\top W(t) \Phi$ and $\Phi^\top \hat{Y}_k \Phi$ are diagonal, write (17) as

$$|\sigma_{l,W}(t) - \sigma_{l,Y}| \leq \varepsilon, \forall l \leq k, \text{ and } |\sigma_{l,W}(t)| \leq \varepsilon, \forall l > k. \quad (18)$$

Given the solution in Corollary 2, we have, $\forall 1 \leq l \leq k$,

$$\begin{aligned} \sigma_{l,Y} &\stackrel{(*)}{>} \sigma_{l,W}(t) \stackrel{(*)}{\geq} \sigma_{l,W} \left(\frac{1}{2\sigma_{l,Y}} \log \frac{C_\varepsilon}{\alpha} \right) \\ &\stackrel{(15)}{=} \sigma_{l,Y} - \left(\sigma_{l,Y} - \frac{C_\varepsilon \sigma_{l,Y} \sigma_{l,0}}{\sigma_{l,Y} + (C_\varepsilon - \alpha) \sigma_{l,0}} \right) \\ &= \sigma_{l,Y} - \frac{\sigma_{l,Y}^2}{(\sigma_{l,Y} - \alpha \sigma_{l,0}) + C_\varepsilon \sigma_{l,0}} \\ &\stackrel{(*)}{\geq} \sigma_{l,Y} - \frac{\sigma_{l,Y}^2}{C_\varepsilon \sigma_{l,0}} \geq \sigma_{l,Y} - \varepsilon, \end{aligned} \quad (19)$$

and $\forall k < l \leq K$,

$$\begin{aligned} 0 &\stackrel{(*)}{<} \sigma_{l,W}(t) \stackrel{(*)}{\leq} \sigma_{l,W} \left(\frac{1}{2\sigma_{l,Y}} \log \frac{c_\varepsilon}{\alpha} \right) \\ &\stackrel{(14)}{=} \frac{c_\varepsilon \sigma_{l,Y} \sigma_{l,0}}{(\sigma_{l,Y} - \alpha \sigma_{l,0}) + c_\varepsilon \sigma_{l,0}} \stackrel{(*)}{\leq} c_\varepsilon \sigma_{l,0} \leq \varepsilon, \end{aligned} \quad (20)$$

where $(*)$ uses the monotonicity of $\sigma_{i,W}(t)$, and $(*)$ uses the fact that $\sigma_{l,Y} - \alpha \sigma_{l,0} \geq 0, \forall l \leq K$, derived from $\alpha \leq c_\varepsilon$. Lastly, since we have assumed $\alpha \leq c_\varepsilon$, we have $\forall l > K$,

$$0 < \sigma_{l,W}(t) = \frac{\alpha \sigma_{l,0}}{1 + 2\alpha \sigma_{l,0} t} \leq \alpha \sigma_{l,0} \leq \varepsilon. \quad (21)$$

With (19)(20)(21), the spectral bound in (18) is verified. \square

As shown in Theorem 3 and the preceding discussions, from a dynamical system perspective, the incremental learning phenomenon comes from some time-scale separation among dynamics corresponding to learning different components in the target matrix. By decreasing the initialization scale α , these time-scale separations become more prominent, allowing one to find low-rank (best rank- k) approximations of the target matrix (See Figure 1) after one transitional phase concludes ($t = \frac{1}{2\sigma_{k,Y}} \log \frac{C_\varepsilon}{\alpha}$) and before another kicks in ($t = \frac{1}{2\sigma_{k+1,Y}} \log \frac{c_\varepsilon}{\alpha}$). Our Theorem exactly quantifies the scale needed to ensure the existence of such low-rank approximations (up to error ε), and the time intervals within which one can find these approximations.

B. Incremental learning under general initialization

When one moves to a general initialization $U(0) = \alpha^{1/2} \tilde{U}_0$ for some U_0 , the dynamics

$$\dot{\tilde{W}} = \tilde{W} \Sigma_Y + \Sigma_Y \tilde{W} - 2\tilde{W}^2, \quad \tilde{W}(0) = \alpha \tilde{W}_0, \quad (22)$$

can no longer be decomposed into scalar dynamics, where $\tilde{W}_0 = \Phi^\top \tilde{U}_0 \tilde{U}_0^\top \Phi$. Nonetheless, similar discussions in the previous section still work: Due to the small initialization scale, the dynamic evolution of the part in \tilde{W} that learns top- k singular components of Y has a faster time scale than those that learn the rest of the singular values, and the interval \mathcal{I}_k again separates the two dynamics. By carefully partitioning the closed-form solution obtained in (1) and analyzing the growth of each block component, we have the following theorem suggesting the same incremental learning occur under general initialization:

Theorem 4 (Incremental learning under general small initialization). *Suppose the K non-zero singular values*

$\sigma_{1,Y}, \sigma_{2,Y}, \dots, \sigma_{K,Y}$ of Y are distinct and ordered in decreasing order. Suppose the initialization of $U(0) = \alpha^{1/2} \bar{U}_0$ satisfies that $\bar{W}_0 = \Phi^\top \bar{U}_0 \bar{U}_0^\top \Phi$ has an inverse denoted by V . Given some error tolerance $0 < \varepsilon \leq \min\{\sigma_{K,Y}, 1\}$, let $c_\varepsilon = \frac{\varepsilon}{16M^2}$, $C_\varepsilon = \frac{16\sigma_{1,Y}^2 M^2}{\varepsilon}$, where $M := \max\{\|V\|, \|V^{-1}\|\}$. If the initialization scale α is sufficiently small so that $\frac{-\log \alpha + \log c_\varepsilon}{-\log \alpha + \log C_\varepsilon} > \max_{1 \leq k \leq K-1} \frac{\sigma_{k+1}}{\sigma_k}$ and $\alpha \leq \frac{c_\varepsilon}{M}$, then the time intervals \mathcal{I}_k defined in (16) are all non-empty, and the $W(t)$ in Proposition 1 satisfies that $\forall 1 \leq k \leq K$,

$$\|W(t) - \hat{Y}_k\| \leq \varepsilon, \quad \forall t \in \mathcal{I}_k, \quad (23)$$

where $\hat{Y}_k := \arg \min_{\text{rank}(Z)=k} \|Y - Z\|_F$ is the best rank- k approximation of Y .

Proof. Similar to the proof of Theorem 3, our assumption on α ensures every \mathcal{I}_k is non-empty. To prove (23), we show that for every choice of k , and for any $t_k \in \mathcal{I}_k$, we have $\|W(t_k) - \hat{Y}_k\| \leq \varepsilon$, the following proof proceeds given such a choice of k and t_k .

Rewrite solution in blocks: From (5) in Proposition 1,

$$\begin{aligned} \tilde{W}(t_k) &= S(t_k) \alpha \tilde{W}_0 \left(I_n + \alpha G(t_k) \tilde{W}_0 \right)^{-1} S^\top(t_k) \\ &\stackrel{(\tilde{W}_0^{-1}=V)}{=} \alpha S(t_k) (V + \alpha G(t_k))^{-1} S^\top(t_k), \end{aligned} \quad (24)$$

Let $\Sigma_k = \text{diag}\{\sigma_{l,Y}\}_{l=1}^k$, and $\Sigma_k^c = \text{diag}\{\sigma_{l,Y}\}_{l=k+1}^K$ (note that Σ_k^c can be an empty matrix), write $V, S(t_k), G(t_k)$ with the following block forms:

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{12}^\top & V_{22} \end{bmatrix}, S(t_k) = \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix}, G(t_k) = \begin{bmatrix} G_1 & 0 \\ 0 & G_2 \end{bmatrix},$$

where $S_1 = e^{\Sigma_k t_k}$, $G_1 = \Sigma_k^{-1} (e^{2\Sigma_k t_k} - I_k)$,

$$S_2 = \begin{bmatrix} e^{\Sigma_k^c t_k} & 0 \\ 0 & I_{n-K} \end{bmatrix}, G_2 = \begin{bmatrix} (\Sigma_k^c)^{-1} (e^{2\Sigma_k^c t_k} - I_{K-k}) & 0 \\ 0 & 2I_{n-K} t_k \end{bmatrix}.$$

Then we have

$$\begin{aligned} \tilde{W}(t_k) &= \alpha \begin{bmatrix} S_1^{-1} (V_{11} + \alpha G_1) S_1^{-1} & S_1^{-1} V_{12} S_2^{-1} \\ S_2^{-1} V_{12}^\top S_1^{-1} & S_2^{-1} (V_{22} + \alpha G_2) S_2^{-1} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^\top & H_{22} \end{bmatrix}, \end{aligned} \quad (25)$$

where (by the inverse formula for a block matrix)

$$\begin{aligned} H_{11} &= \alpha S_1 \left((V_{11} - V_{12} (V_{22} + \alpha G_2)^{-1} V_{12}^\top) + \alpha G_1 \right)^{-1} S_1, \\ H_{12} &= -H_{11} S_1^{-1} V_{12} (V_{22} + \alpha G_2)^{-1} S_2, \\ H_{22} &= \alpha S_2 (V_{22} + \alpha G_2)^{-1} S_2 - S_2 (V_{22} + \alpha G_2)^{-1} V_{12}^\top S_1^{-1} H_{12}. \end{aligned}$$

Our goal is to show that

$$\|W(t_k) - \hat{Y}_k\| = \left\| \Phi \begin{bmatrix} H_{11} - \Sigma_k & H_{12} \\ H_{12}^\top & H_{22} \end{bmatrix} \Phi^\top \right\| \leq \varepsilon. \quad (26)$$

Thus it suffices to show the following

$$\|H_{11} - \Sigma_k\| \leq \frac{\varepsilon}{4}, \quad \|H_{12}\| \leq \frac{\varepsilon}{4}, \quad \|H_{22}\| \leq \frac{\varepsilon}{4}. \quad (27)$$

The rest of the proof is to show (27) holds, for which we start by deriving norm bounds on S_1 and S_2 , and then move to $H_{11} - \Sigma_k$, H_{12} and H_{22} .

Norm bounds on S_1 and S_2 : From that $t_k \in \mathcal{I}_k$, we have

$$S_1^{-1} = e^{-\Sigma_k t_k} \preceq e^{-\Sigma_k \frac{1}{2\sigma_k} \log \frac{C_\varepsilon}{\alpha}} \preceq e^{-\frac{1}{2} \log \frac{C_\varepsilon}{\alpha}} I_k = \sqrt{\frac{\alpha}{C_\varepsilon}} I_k,$$

and similarly since $e^{\Sigma_k^c t_k} \preceq e^{\Sigma_k^c \frac{1}{2\sigma_{k+1}} \log \frac{C_\varepsilon}{\alpha}} \preceq e^{\frac{1}{2} \log \frac{C_\varepsilon}{\alpha}} I_k$, we have $S_2 \preceq \begin{bmatrix} \sqrt{\frac{C_\varepsilon}{\alpha}} I_{K-k} & 0 \\ 0 & I_{n-K} \end{bmatrix}$. Together we have

$$\|S_1^{-1}\| \leq \sqrt{\frac{\alpha}{C_\varepsilon}}, \quad \|S_2\| \leq \sqrt{\frac{C_\varepsilon}{\alpha}}. \quad (28)$$

Norm bounds on $H_{11} - \Sigma_k$, H_{12} and H_{22} : The primary focus is the bound on $\|H_{11} - \Sigma_k\|$, for the sake of simplicity, we let $(V_{11} - V_{12}(V_{22} + \alpha G_2)^{-1} V_{12}^\top) := \tilde{V}_1$, then

$$\begin{aligned} H_{11} &= \alpha S_1 \left(\tilde{V}_1 + \alpha G_1 \right)^{-1} S_1 \\ &= \left(\alpha^{-1} S_1^{-1} \tilde{V}_1 S_1^{-1} + S_1^{-1} G_1 S_1^{-1} \right)^{-1} \\ &= \left(\underbrace{\alpha^{-1} S_1^{-1} \tilde{V}_1 S_1^{-1} + S_1^{-1} \Sigma_k^{-1} S_1^{-1}}_{:=\Delta} + \Sigma_k^{-1} \right)^{-1}. \end{aligned}$$

Notice that

$$0 \prec V_{11} - V_{12} V_{22}^{-1} V_{12}^\top \preceq V_{11} - V_{12} (V_{22} + \alpha G_2)^{-1} V_{12}^\top \preceq V_{11},$$

from which we know $\|\tilde{V}_1\| \leq \|V_{11}\| \stackrel{(a)}{\leq} M$. Then by (28),

$$\|\Delta\| \leq \frac{\|S_1^{-1}\|^2 \|\tilde{V}_1\|}{\alpha} + \|S_1^{-1}\|^2 \|\Sigma_k^{-1}\| \leq \frac{M + \sigma_{K,Y}^{-1} \alpha}{C_\varepsilon} \stackrel{(b)}{\leq} \frac{2M}{C_\varepsilon}.$$

As long as $\|\Sigma_k\| \|\Delta\| \stackrel{(c)}{\leq} \frac{1}{2}$, we have, from [25],

$$\begin{aligned} \|H_{11} - \Sigma_k\| &= \|(\Delta + \Sigma_k^{-1})^{-1} - \Sigma_k\| \\ &\leq \frac{\|\Sigma_k\|^2 \|\Delta\|}{1 - \|\Sigma_k\| \|\Delta\|} \leq 2\sigma_{1,Y}^2 \|\Delta\| \leq \frac{4\sigma_{1,Y}^2 M}{C_\varepsilon} \leq \frac{\varepsilon}{4}, \end{aligned}$$

the desired bound on $\|H_{11} - \Sigma_k\|$. Since we have assumed $\varepsilon \leq \sigma_{1,Y}$, this also suggests $\|H_{11}\| \leq \|\Sigma_k\| + \varepsilon \leq 2\sigma_{1,Y}$.

For H_{12}, H_{22} , we use the fact that $\|(V_{22} + \alpha G_2)^{-1}\| \leq \|V_{22}^{-1}\| \stackrel{(a)}{\leq} M$, and $\|V_{12}\| \stackrel{(a)}{\leq} M$, together with (28), to obtain

$$\begin{aligned} \|H_{12}\| &\leq \|H_{11}\| \|S_1^{-1}\| \|V_{12}\| \|V_{22}^{-1}\| \|S_2\| \\ &\leq 2\sigma_{1,Y} M^2 \sqrt{\frac{C_\varepsilon}{\alpha}} = \frac{\varepsilon}{8} \leq \frac{\varepsilon}{4}, \end{aligned} \quad (29)$$

and

$$\begin{aligned} \|H_{22}\| &\leq \alpha \|S_2\|^2 \|V_{22}^{-1}\| + \|H_{12}\| \|S_1^{-1}\| \|V_{12}\| \|V_{22}^{-1}\| \|S_2\| \\ &\leq c_\varepsilon M + \frac{\varepsilon}{8} \sqrt{\frac{C_\varepsilon}{\alpha}} M^2 \leq \frac{\varepsilon}{4}. \end{aligned} \quad (30)$$

This finishes the proof. For the sake of readability, we omitted several arguments (a)(b)(c) used in the last part of the proof, and we explain them here. (a): $\|V_{11}\| \leq M, \|V_{12}\| \leq M$ because the norm of submatrix is bounded by that of the full matrix, and $\|V_{22}^{-1}\| \leq M$ is due to the interlacing theorem [25] for principal submatrix. (b): we have $\sigma_{K,Y}^{-1} \alpha \leq \sigma_{K,Y}^{-1} \varepsilon \leq 1 \leq M$. (c): with $\|\Delta\| \leq \frac{2M}{C_\varepsilon}$, it is easy to verify that $\|\Sigma_k\| \|\Delta\| \leq \sigma_{1,Y} \|\Delta\| \leq \frac{1}{2}$. \square

The incremental learning results in Theorem 4 resemble the one in Theorem 3 for spectral initialization, except for some differences in the definitions of $c_\varepsilon, C_\varepsilon$. It characterizes the quantitative effect of the initialization scale in determining the time-scale separation among the learning dynamics for each target singular component, which provides guidance on how to utilize these results to design training algorithms that can learn low-rank approximations of the ground truth by early stopping. In addition, the proof of the symmetric factorization case might serve as a basis for broader settings.

C. Remarks on other settings

1) *Random initialization:* A more natural setting is arguably the random initialization with small variance, which corresponds to having $U(0) = \alpha^{1/2} \tilde{U}_0$ for some \tilde{U}_0 whose entries are i.i.d. samples from a standard Gaussian (or other sub-Gaussian distributions). Applying Theorem 4 to this random initialization setting requires an extra concentration results on the matrix $\tilde{W}_0 = \tilde{U}_0 \tilde{U}_0^\top$. If one can show that for $\delta \in (0, 1)$, $\exists M_\delta > 0$ such that $\mathbb{P}(\max\{\|\tilde{W}_0\|, \|\tilde{W}_0^{-1}\|\} \leq M_\delta) \geq 1 - \delta$, then one can define $c_\varepsilon, C_\varepsilon$ by M_δ and find initialization scale α accordingly, and Theorem 4 suggests that the incremental learning phenomenon as described by (23) happens with probability at least $1 - \delta$.

2) *Rank-deficient initialization:* Our proof assumes an initialization shape \tilde{U}_0 such that $\tilde{W}_0 = \Phi^\top \tilde{U}_0 \tilde{U}_0^\top \Phi$ is invertible. This is satisfied if $r < n$ when \tilde{W}_0 is always rank-deficient. This requires significant changes in the proof since now we should analyze $\tilde{W}(t_k) = S(t_k) \alpha \tilde{W}_0 (I_n + \alpha G(t_k) \tilde{W}_0)^{-1} S^\top(t_k)$ directly in (24). With Woodbury's matrix identity, one rewrite $\tilde{W}(t_k)$ as

$$\alpha S(t_k) \tilde{U} (I - \tilde{U}^\top G(t_k) \tilde{U} (I + \alpha \tilde{U}^\top G(t_k) \tilde{U})^{-1}) \tilde{U}^\top S(t_k),$$

where $\tilde{U} = \Phi^\top \tilde{U}_0$. Detailed analysis on $\tilde{U}^\top G(t_k) \tilde{U}$ leads to similar results as Theorem 4, which is left to future work.

IV. CONCLUSION

In this paper, we develop a quantitative understanding of the incremental learning phenomenon in GF on symmetric matrix factorization problems through its closed-form solution. From a dynamical system perspective, the incremental learning phenomenon comes from some time-scale separation among dynamics corresponding to learning different components in the target matrix. By decreasing the initialization scale, these time-scale separations become more prominent, allowing one to find low-rank approximations of the target matrix.

Future work includes extending current results to asymmetric matrix factorization. Notably, for asymmetric matrix factorization, [15] shows that if the factors are initialized to satisfy some balancedness conditions, then the learning dynamics can be related to GF on a symmetric factorization problem. Moreover, under a small scale, any initialization is close to one that satisfies such balancedness conditions; therefore, the analysis for general initialization can be viewed

as analyzing perturbed dynamics from a nominal dynamics with closed-form solution.

REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.
- [4] G. Vardi, "On the implicit bias in deep-learning algorithms," *Communications of the ACM*, vol. 66, no. 6, pp. 86–93, 2023.
- [5] S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Implicit regularization in matrix factorization," in *NeurIPS*, 2017.
- [6] G. Gidel, F. Bach, and S. Lacoste-Julien, "Implicit regularization of discrete gradient dynamics in linear neural networks," in *NeurIPS*, 2019.
- [7] J. Jin, Z. Li, K. Lyu, S. S. Du, and J. D. Lee, "Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing," in *ICML*, 2023.
- [8] H. Min, E. Mallada, and R. Vidal, "Early neuron alignment in two-layer relu networks with small initialization," in *ICLR*, 2024.
- [9] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro, "Implicit bias of gradient descent on linear convolutional networks," in *NeurIPS*, 2018.
- [10] E. Moroshko, B. E. Woodworth, S. Gunasekar, J. D. Lee, N. Srebro, and D. Soudry, "Implicit bias in deep linear classification: Initialization scale vs training accuracy," *NeurIPS*, 2020.
- [11] J. Li, T. V. Nguyen, C. Hegde, and R. K. W. Wong, "Implicit sparse regularization: The impact of depth and early stopping," in *NeurIPS*, 2021.
- [12] E. Abbe, S. Bengio, E. Boix-Adsera, E. Littwin, and J. Susskind, "Transformers learn through gradual rank increase," *NeurIPS*, vol. 36, 2023.
- [13] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural network," in *ICLR*, 2014.
- [14] S. Burer and R. D. C. Monteiro, "Local minima and convergence in low-rank semidefinite programming," *Math. Program.*, vol. 103, p. 427–444, July 2005.
- [15] S. Tarmoun, G. França, B. D. Haeffele, and R. Vidal, "Understanding the dynamics of gradient flow in overparameterized linear models," in *ICML*, 2021.
- [16] H. Min, S. Tarmoun, R. Vidal, and E. Mallada, "On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks," in *ICML*, 2021.
- [17] A. C. B. de Oliveira, M. Siami, and E. D. Sontag, "Dynamics and perturbations of overparameterized linear neural networks," in *CDC*, 2023.
- [18] S. Arora, N. Cohen, W. Hu, and Y. Luo, "Implicit regularization in deep matrix factorization," *NeurIPS*, 2019.
- [19] D. Stöger and M. Soltanolkotabi, "Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction," in *NeurIPS*, 2021.
- [20] C. Yaras, P. Wang, L. Balzano, and Q. Qu, "Compressible dynamics in deep overparameterized low-rank learning & adaptation," in *ICML*, 2024.
- [21] Z. Xu, H. Min, L. E. MacDonald, J. Luo, S. Tarmoun, E. Mallada, and R. Vidal, "Understanding the learning dynamics of lora: A gradient flow perspective on low-rank adaptation in matrix factorization," in *AISTATS*, 2025.
- [22] H.-H. Chou, J. Maly, C. M. Verdun, B. F. P. da Costa, and H. Miranda, "Get rid of your constraints and reparametrize: A study in NNLS and implicit bias," in *AISTATS*, 2025.
- [23] V. Kucera, "A review of the matrix riccati equation," *Kybernetika*, vol. 9, no. 1, pp. 42–61, 1973.
- [24] T. Sasagawa, "On the finite escape phenomena for matrix riccati equations," *IEEE Transactions on Automatic Control*, vol. 27, no. 4, pp. 977–979, 1982.
- [25] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York, NY, USA: Cambridge University Press, 2nd ed., 2012.