
Connectivity Shapes Implicit Regularization in Matrix Factorization Models for Matrix Completion

Zhiwei Bai¹, Jiajie Zhao¹, Yaoyu Zhang^{1,2*}

¹ School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.

² Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai 200240, P.R. China
{bai299, zjj0216, zhyy.sjtu}@sjtu.edu.cn.

Abstract

Matrix factorization models have been extensively studied as a valuable test-bed for understanding the implicit biases of overparameterized models. Although both low nuclear norm and low rank regularization have been studied for these models, a unified understanding of when, how, and why they achieve different implicit regularization effects remains elusive. In this work, we systematically investigate the implicit regularization of matrix factorization for solving matrix completion problems. We empirically discover that the connectivity of observed data plays a crucial role in the implicit bias, with a transition from low nuclear norm to low rank as data shifts from disconnected to connected with increased observations. We identify a hierarchy of intrinsic invariant manifolds in the loss landscape that guide the training trajectory to evolve from low-rank to higher-rank solutions. Based on this finding, we theoretically characterize the training trajectory as following the hierarchical invariant manifold traversal process, generalizing the characterization of Li et al. (2020) to include the disconnected case. Furthermore, we establish conditions that guarantee minimum nuclear norm, closely aligning with our experimental findings, and we provide a dynamics characterization condition for ensuring minimum rank. Our work reveals the intricate interplay between data connectivity, training dynamics, and implicit regularization in matrix factorization models.

1 Introduction

Overparameterized models have the capacity to easily fit data with random labels (Zhang et al., 2017, 2021). However, in real-world applications, models with more parameters than training samples still generalize well. This has led researchers to hypothesize that overparameterized models undergo implicit regularization, favoring certain functions as outputs. Overparameterized matrix factorization models, $f_{\theta} = \mathbf{A}\mathbf{B}$ with $\theta = (\mathbf{A}, \mathbf{B})$, $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$, have served as a simplified test-bed for studying this implicit regularization. In the context of matrix completion problems like the Netflix challenge, these models aim to find a low-rank completion of a partially observed matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$. Prior works have offered seemingly conflicting perspectives on the implicit regularization at play, with some claiming it promotes low nuclear norm (Gunasekar et al., 2017) and others arguing for low rank (Arora et al., 2019; Li et al., 2020; Razin and Cohen, 2020). However, a unified understanding of when, how, and why they achieve different implicit regularization effects remains elusive.

Unlike previous works that focus on either low rank or low nuclear norm regularization, we systematically investigate the training dynamics and implicit regularization of matrix factorization for matrix completion. Through extensive experiments, we found that a certain connectivity property of

*Corresponding author: zhyy.sjtu@sjtu.edu.cn.

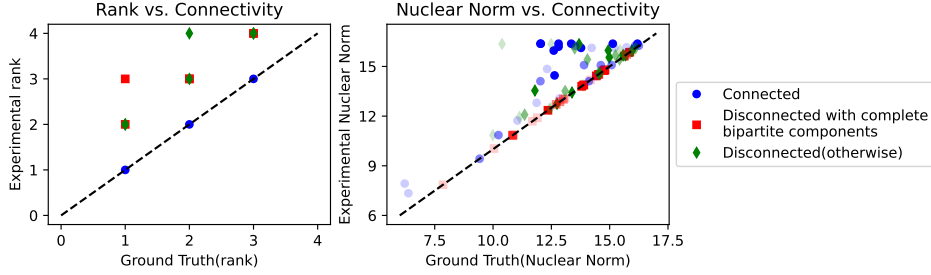


Figure 1: **The connectivity of observed data affects the implicit regularization.** The ground truth matrix $M^* \in \mathbb{R}^{4 \times 4}$ has rank ranging from 1 to 3. The sample size n covers settings where n is equal to, smaller than, and larger than the $2rd - r^2$ threshold required for exact reconstruction. Darker scatter points indicate a greater number of samples, while lighter points indicate fewer samples. The positions of observed entries are randomly chosen, and the experiment is repeated 10 times for each sample size. (Please refer to Appendix B for additional experiments and detailed methodology.)

observed data (see Def. 2 for the definition of connectivity) plays a key role in the implicit regularization effects. As shown in Fig. 1, we sample observations randomly from a ground truth matrix $M^* \in \mathbb{R}^{d \times d}$ with $\text{rank}(M^*) < d$ and train models $f_\theta = AB$, $A, B \in \mathbb{R}^{d \times d}$ from small random initialization without any rank constraints. For each observation set, we calculate the solutions with the minimum nuclear norm and minimum rank, which serve as the ground truth benchmarks. These are then compared with the completion matrix obtained by the model. From Fig. 1, we observe that:

- (i) **Low rank bias in connected case:** When the observed entries are connected, the model consistently learns the lowest-rank solution.
- (ii) **Low nuclear norm bias in certain disconnected case:** When the observed entries are disconnected, the model generally does not find the minimum nuclear norm or lowest-rank solution. However, in the special case where each connected component is a complete bipartite subgraph, the model consistently finds the minimum nuclear norm solution.

To understand how data connectivity modulates the implicit bias, we analyze the loss landscape and optimization dynamics. We find a hierarchy of intrinsic invariant manifolds Ω_k of different ranks in the loss landscape. These manifolds constrain the optimization trajectory, causing the model to learn by incrementally ascending through higher ranks. In the disconnected case, additional sub- Ω_k invariant manifolds emerge within the Ω_k invariant manifold, preventing the model from reaching the global lowest-rank solution. However, we prove that the minimum nuclear norm solution is guaranteed in the disconnected with complete bipartite subgraph case.

The contributions of our work are summarized as follows:

- (i) We systematically investigate the influence of data connectivity on the implicit regularization. Our empirical findings indicate that the connectivity of observed data plays a key role in the implicit bias, leading to a transition from favoring solutions with a low nuclear norm to those with a low rank as the data becomes more connected with an increase in observations (refer to Sec. 4).
- (ii) We characterize the training dynamics of matrix factorization theoretically, showing that the optimization trajectory follows a *Hierarchical Invariant Manifold Traversal (HIMT)* process. This generalizes the characterization of Li et al. (2020), whose proposed *Greedy Low-Rank Learning (GLRL)* algorithm equivalence only corresponding to the connected case (refer to Sec. 5 and Sec. 6.1).
- (iii) Regarding the minimum nuclear norm regularization, we establish conditions that provide guarantees closely aligned with our empirical findings, which complement the results of Gunasekar et al. (2017). For the minimum rank regularization, we present a dynamic characterization condition that assures the attainment of the minimum rank solution (refer to Sec. 6.2).

2 Related works

Norm minimization and rank minimization. Extensive research has been conducted on the implicit regularization of matrix factorization models, focusing on norm minimization and rank mini-

mization. For norm minimization, Gunasekar et al. (2017) proved that gradient flow with infinitesimal initialization converges to the minimum nuclear norm solution in the special case of commutative observations. Ji and Telgarsky (2019); Gunasekar et al. (2018) studied norm minimization regularization in deep linear networks. For rank minimization, numerous works have shown that matrix factorization models favor low-rank solutions. Arora et al. (2019); Gidel et al. (2019); Gissin et al. (2019); Razin and Cohen (2020); Jiang et al. (2023); Belabbas (2020) investigated how infinitesimal initialization of gradient flow encourages low rank in specific settings. Li et al. (2020) showed that under certain assumptions, matrix factorization dynamics are equivalent to a greedy low-rank learning heuristic. Li et al. (2018); Stöger and Soltanolkotabi (2021); Jin et al. (2023) established low-rank recovery guarantees for matrix sensing problems under the Restricted Isometry Property (RIP) condition. Zhang et al. (2022, 2023) studied a broader class of model rank minimization for nonlinear models, of which the matrix factorization model is a special case.

Nonlinear dynamics. The initialization scale can significantly influence the implicit regularization of neural networks. Large initialization typically leads to linear dynamics (Jacot et al., 2018) and poor generalization (Chizat et al., 2019), while small initialization induces nonlinear dynamics (Luo et al., 2021). In this work, we focus on the case of infinitesimal initialization, which corresponds to highly nonlinear dynamics. An important characteristic of nonlinear neural network dynamics is the phenomenon of condensation (Luo et al., 2021; Zhou et al., 2022), where the network’s effective complexity is small. The low-rank Ω_k invariant manifolds we propose are essentially a manifestation of condensation. Zhang et al. (2021, 2022); Bai et al. (2022); Fukumizu et al. (2019); Simsek et al. (2021) established the embedding principle of the loss landscape of neural networks and empirically demonstrated that the training process traverses critical points embedded from smaller subnetworks. Jacot et al. (2021) conjectured a saddle to saddle dynamics for deep linear networks, which is conceptually analogous to the dynamics characterization in this work.

3 Preliminaries

Matrix completion problem. This study focuses on the matrix completion problem, which involves estimating missing entries within a partially observed matrix. Given an incomplete matrix $M \in \mathbb{R}^{d \times d}$, the goal is to predict the entirety of M based on its observed elements. The set of observed entries is represented as $S = \{(i_k, j_k), M_{i_k, j_k}\}_{k=1}^n$, where (i_k, j_k) indicates the row and column indices, and M_{i_k, j_k} is the corresponding value assumed non-zero in the matrix. The set of observed indices is defined as $S_{\mathbf{x}} = \{(i_k, j_k)\}_{k=1}^n$. Entries that are not observed, denoted by \star , are considered missing or unknown. The positions of observed elements in the matrix M are defined by a binary observation matrix P , where $P_{ij} = 1$ indicates that M_{ij} is observed, and $P_{ij} = 0$ indicates that M_{ij} is unobserved.

Matrix factorization model. Matrix factorization is a prevalent approach for addressing the matrix completion problem. It reconstructs the matrix $W \in \mathbb{R}^{d \times d}$ through the product $W = AB$, where $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times d}$. This work studies the overparameterized scenario with $r = d$, aiming to understand the implicit regularization effect in the absence of explicit rank restrictions, paralleling prior research (Gunasekar et al., 2017; Arora et al., 2019; Li et al., 2020; Jin et al., 2023). In this work, we focus on the asymmetric factorization, which can be represented as a parametric model:

$$f_{\theta} = AB, \quad A, B \in \mathbb{R}^{d \times d}. \quad (1)$$

The matrix factorization model parameters are denoted by $\theta = (A, B)$, identified with its vectorized form $\text{vec}(\theta) \in \mathbb{R}^{2d^2}$. The augmented matrix is $W_{\text{aug}}^{\top} = [A^{\top} \quad B^{\top}]^{\top} \in \mathbb{R}^{d \times 2d}$, and $\text{row}(A)$ and $\text{col}(B)$ denote the row and column spaces of A and B , respectively.

Loss function. The learning process for the parameters $\theta = (A, B)$ involves minimizing a loss function that measures the difference between observed and estimated entries. In this work, we focus on the mean squared error, and the empirical risk is thus formulated as

$$R_S(\theta) = \frac{1}{n} \| (AB - M)_{S_{\mathbf{x}}} \|_F^2 := \frac{1}{n} \sum_{k=1}^n (\mathbf{a}_{i_k} \cdot \mathbf{b}_{\cdot, j_k} - M_{i_k, j_k})^2, \quad (2)$$

where \mathbf{a}_i and $\mathbf{b}_{\cdot, j}$ represent the i -th row and j -th column of matrix A and B , respectively. The residual matrix $\delta M = (AB - M)_{S_{\mathbf{x}}}$ has elements $\delta M_{ij} = (AB)_{ij} - M_{ij}$ for $(i, j) \in S_{\mathbf{x}}$ and

$\delta M_{ij} = 0$ for $(i, j) \notin S_x$. The training dynamics follow the gradient flow of $R_S(\theta)$:

$$\frac{d\theta}{dt} = -\nabla_{\theta} R_S(\theta), \quad \theta(0) = \theta_0. \quad (3)$$

4 Connectivity affects implicit regularization

In this section, we define connectivity and present experimental results on implicit regularization for connected and disconnected observational data.

Definition 1 (Associated Observation Graph). Given an incomplete matrix M to be completed and its observation matrix P , the associated observation graph G_M is the bipartite graph with adjacency matrix $\begin{bmatrix} \mathbf{0} & P^{\top} \\ P & \mathbf{0} \end{bmatrix}$, with isolated vertices removed.

Definition 2 (Connectivity). Given an incomplete matrix M to be completed, it is considered connected if its associated observation graph G_M is connected; otherwise, it is disconnected. The connected components of M are defined as the connected components of G_M .

The connectivity of the graph, as defined above, reflects the connectivity of the observed data. Appendix A Sec. A.2 provides a detailed discussion on the equivalent definition of connectivity.

In the case of disconnectivity, there is a special case where each connected component has full observations, characterized by disconnectivity with complete bipartite components.

Definition 3 (Disconnectivity with Complete Bipartite Components). A incomplete matrix M is considered disconnected with complete bipartite components if its associated observation graph G_M is disconnected and each connected component forms a complete bipartite subgraph.

We present examples to demonstrate how connectivity influences the characteristics of the learned solutions. Consider three matrices to be completed, each obtained by adding one more observation to the previous matrix: M_1 (disconnected), M_2 (disconnected with complete bipartite components), and M_3 (connected). Fig. A1 of Appendix B illustrates the associated graphs G_M .

$$M_1 = \begin{bmatrix} 1 & 2 & \star \\ 3 & \star & \star \\ \star & \star & 5 \end{bmatrix}, M_2 = \begin{bmatrix} 1 & 2 & \star \\ 3 & 4 & \star \\ \star & \star & 5 \end{bmatrix}, M_3 = \begin{bmatrix} 1 & 2 & \star \\ 3 & 4 & \star \\ 6 & \star & 5 \end{bmatrix}. \quad (4)$$

Figs. 2(a-b) compare the learned matrices with the ground truth (GT) solutions having the smallest nuclear norm and rank. For disconnected M_1 (blue bars), the learned solution achieves neither the smallest nuclear norm nor rank. For disconnected M_2 with complete bipartite components (green bars), the learned matrix has the smallest nuclear norm but not rank. For connected M_3 (red bars), the lowest rank-2 solution is not unique; the model identifies a particular lowest rank-2 solution, but it does not correspond to the one with the minimum nuclear norm.

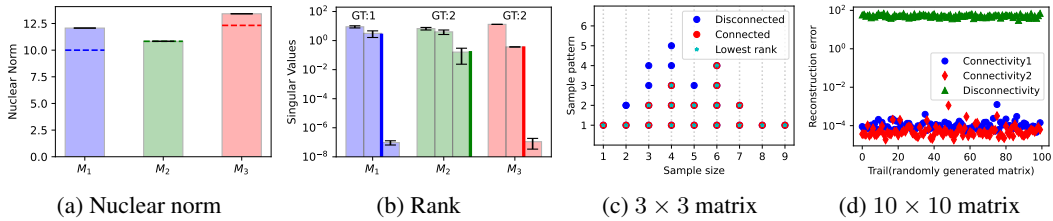


Figure 2: (a) Nuclear norms of the learned solutions for M_1 , M_2 , and M_3 . Dashed lines represent theoretically computed smallest nuclear norms. (b) Singular values of the learned matrices for M_1 , M_2 , and M_3 , with the solid line partitioning significantly nonzero singular values (experimental rank) and the text (GT) showing the ground truth minimum rank. Mean and standard deviation are recorded over 100 repetitions. (c) All equivalent sampling patterns of the 3×3 matrix completion problem (see Appendix B for details). Cyan stars marked the case learning the lowest-rank solution. (d) Reconstruction error of the solutions for a 10×10 matrix reconstruction problem with M^* randomly sampled at rank $r = 1$ and sample size set to the minimum reconstruction setting $n = 2rd - r^2$.

To thoroughly study all possible cases, we examine all sampling patterns of the 3×3 matrix completion. Fig. 2(c) shows that the model consistently learns the lowest-rank solution for connected sampling patterns but fails to do so for disconnected patterns. Fig. 2(d) further verifies the impact of connectivity on low-rank matrix recovery by comparing the reconstruction error for 100 randomly sampled rank-1 matrices using two connected sampling patterns (red and blue dots) and one disconnected sampling pattern (green dots). The model consistently achieves small reconstruction errors under connected sampling patterns, while the error is significantly larger for the disconnected pattern.

These empirical results demonstrate an implicit preference for low rank induced by connectivity and a preference for low nuclear norm in a particular kind of disconnection. In the following section, we will investigate the training dynamics under both connected and disconnected scenarios.

5 Training dynamics in connected and disconnected cases

5.1 Connected case

This section empirically demonstrates the detailed dynamics of connected observed data. Fig. 3(a) shows the connected target matrix M with a single unknown element denoted by \star . The rank of M is at least three and equals three if and only if $\star = 1.2$.

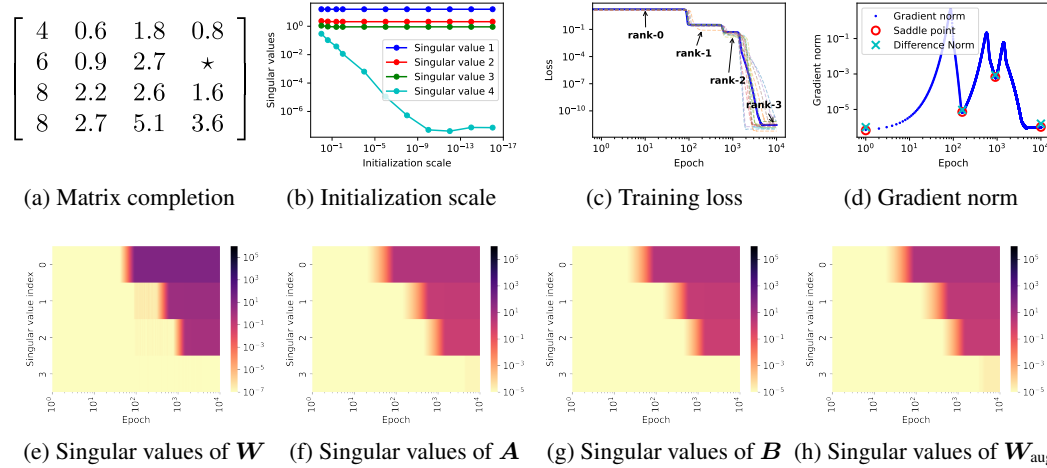


Figure 3: (a) The matrix M to be completed, with the \star position unknown. (b) The four singular values of the learned solution at different initialization scale (Gaussian distribution, mean 0, variance from 10^0 to 10^{-16}). (c) The training loss curves corresponding to different sampling locations. In each sampling, we randomly cover an element and observe the other 15 elements. (d) Evolution of the l_2 -norm of the gradients throughout the training process. The cyan crosses represent the difference between the matrix corresponding to the saddle point and the optimal approximation at each rank. (e-h) Evolution of singular values for matrices W , A , B , and W_{aug} during training.

Learning lowest-rank solution. We initialize A and B with different scales and record the singular values of the learned matrix. As depicted in Fig. 3(b), when starting with larger initialization, the learned solutions are almost always rank-4. Conversely, as the initialization scale decreases, the first three singular values of the learned solution are consistently maintained in magnitude, but the fourth singular value keeps decreasing, resulting in the model learning the lowest rank-3 solution.

Traversing progressive optima at each rank. For a small initialization, the loss curves exhibit a steady, stepwise decline (Fig. 3(c)). The flat periods correspond to small gradient norms, indicating potential saddle points (Fig. 3(d)). We compare the matrices learned at these saddle points with the optimal approximation of each rank and plot their difference in Fig. 3(d), which is very small. These findings suggest that the model starts near 0 (rank-0) and progressively finds optimal approximations within rank-1, rank-2, and higher-rank manifolds until reaching a global minimum.

Alignment of the row space of A and the column space of B . Starting with small initialization, we track the rank (number of significantly non-zero singular values) of $W = AB$, A , B , and the augmented matrix W_{aug} during the training process. We observe that the rank gradually increases, with singular values growing rapidly one after another (Fig. 3(e-h)). Throughout the entire process, we consistently find that $\text{rank}(A) = \text{rank}(B^\top) = \text{rank}(W_{\text{aug}})$, which implies that the row space of A and the column space of B remain aligned at all times. This alignment corresponds to a special structure that we refer to as the ‘‘Hierarchical Intrinsic Invariant Manifold’’ in Sec. 6.1, which plays a crucial role in the overall dynamics of the system.

The dynamics of increasing ranks step by step aligns with the description of *Greedy Low Rank Learning (GLRL)* (Li et al., 2020). However, we will show next that when the observed data are disconnected, the learning process is not equivalent to GLRL.

5.2 Disconnected case

In this section, we present a typical experiment in the disconnected situation. As depicted in Fig. 4(a), the target matrix M contains four unknown elements denoted by \star and is disconnected. The rank of M is at least one, and there are infinitely many rank-1 solutions.

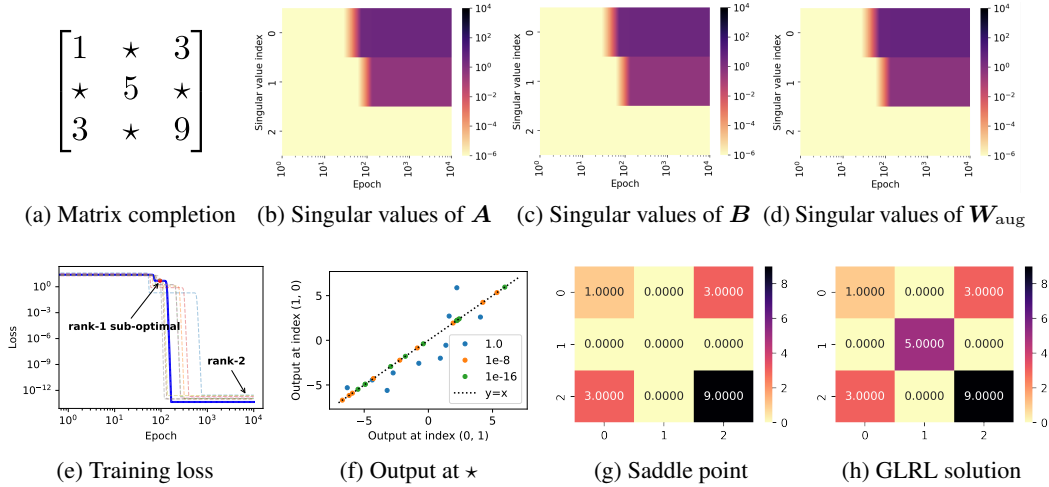


Figure 4: (a) The matrix to be completed, with unknown entries marked by \star . (b-d) Evolution of singular values for A , B , and W_{aug} during training. (e) Training loss for different disconnected sampling patterns, each covering 4 elements and observing the remaining 5. (f) Learned values at symmetric positions (0, 1) and (1, 0) under varying initialization scales (zero mean, varying variance). Each point represents one of ten random experiments per variance; labels show initialization variance. Other symmetric positions exhibit similar behavior. (g) Learned output at the saddle point corresponding to the red dot in (e). (h) Final learned solution of the GLRL algorithm (Li et al., 2020).

Alignment of the row space of A and the column space of B . As shown in Fig. 4(b-e), the learning process in the disconnected case is similar to the previous experiment: the model naturally evolves from low-rank to high-rank, with each step increasing a singular value and satisfying $\text{rank}(A) = \text{rank}(B^\top) = \text{rank}(W_{\text{aug}})$. Fig. 4(f) illustrates that as the initialization scale decreases, the model tends to learn symmetric solutions. However, unlike the connected case, the output does not approach a particular solution as the initialization decreases. For this specific disconnected M , we will show that every symmetric solution learned is a minimal nuclear norm solution (see Sec. 6.2 Thm. 4). For fewer observations, the experimental phenomena are similar (see Appendix B Fig. B5).

Lowest-rank solution is not learned. Despite the adaptive learning behavior, the final learned solution has rank 2, as evidenced by the two significantly non-zero singular values in Fig. 4(b-d). Examining the dynamics (3), we find that they decouple into two independent systems: one for the 1st and 3rd rows of A and columns of B , and another for the 2nd row of A and column of B . Fig. 4(g) shows that the model first learns the surrounding elements 1, 3, 3, 9 (rank-1 saddle point), then learns

the middle element 5 in the next stage. The decoupling of dynamics is equivalent to the definition of disconnection (see Appendix A Prop. A.4 for proof). The loss curves in Fig. 4(e) consistently indicate that in the disconnected case, the model learns a sub-optimal solution in the rank-1 manifold, resulting in a rank-2 solution finally.

Not equivalent to GLRL in disconnected case. We compare the GLRL algorithm (Li et al., 2020) with the matrix factorization model for solving the same matrix completion problem (Fig. 4). Li et al. (2020) claim that the matrix factorization dynamics is mathematically equivalent to the GLRL algorithm under reasonable assumptions. While GLRL learns the same rank-1 saddle point shown in Fig. 4(g) in the first stage, it then fills unobserved elements with 0, resulting in a unique rank-2 solution (Fig. 4(h)). In contrast, the matrix factorization model learns symmetric solutions with some degree of freedom depending on the random seed (Fig. 4(f)). The key difference is that the first critical point (Fig. 4(g)) reached by the trajectory is a sub-optimal and not a second-order stationary point of the rank-1 manifold as assumed by Li et al. (2020). Therefore, the equivalence assumption between GLRL and matrix factorization does not hold in the disconnected case.

6 Theoretical analysis of training dynamics and implicit regularization

6.1 Characterization of training dynamics

Matrix factorization models exhibit a distinctive adaptive learning behavior, progressively evolving from low rank to high rank. Understanding this phenomenon is rooted in grasping the global dynamics of matrix factorization models, where the role of intrinsic invariant manifolds becomes critical.

Proposition 1 (Hierarchical Intrinsic Invariant Manifold (HIIM)). (see Appendix A Prop. A.1 for Proof) Let $\mathbf{f}_\theta = \mathbf{A}\mathbf{B}$ be a matrix factorization model and $\{\alpha_1, \dots, \alpha_k\}$ be k linearly independent vectors. Define the manifold Ω_k as $\Omega_k := \Omega_k(\alpha_1, \dots, \alpha_k) = \{\theta = (\mathbf{A}, \mathbf{B}) \mid \text{row}(\mathbf{A}) = \text{col}(\mathbf{B}) = \text{span}\{\alpha_1, \dots, \alpha_k\}\}$. The manifold Ω_k possesses the following properties:

(i) **Invariance under Gradient Flow:** Given data S and the gradient flow dynamics $\dot{\theta} = -\nabla R_S(\theta)$, if the initial point $\theta_0 \in \Omega_k$, then $\theta(t) \in \Omega_k$ for all $t \geq 0$.

(ii) **Intrinsic Property:** Ω_k is a data-independent invariant manifold, meaning that for any data S , Ω_k remains invariant under the gradient flow dynamics.

(iii) **Hierarchical Structure:** The manifolds Ω_k form a hierarchy: $\Omega_0 \subsetneq \Omega_1 \subsetneq \dots \subsetneq \Omega_{k-1} \subsetneq \Omega_k$.

Figs. 3(f-h) and Figs. 4(b-d) show that the training process with small initialization consistently satisfies $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{B}^\top) = \text{rank}(\mathbf{W}_{\text{aug}})$, aligning with the Ω_k invariant manifold. Since a non-zero initialization in practice, the training trajectory is close to the Ω_k invariant manifold, approaches a critical point, and transitions to the next level invariant manifold without getting trapped.

In both connected and disconnected scenarios, we observe a step-by-step hierarchical Ω_k invariant manifold traversal. In the connected case, at each level model reaches an optimal solution. However, in the disconnected case, we can prove that each connected component induces a sub- Ω_k invariant manifold, leading to the experimentally observed sub-optimal solution (see Fig. 4).

Proposition 2 (Intrinsic Sub- Ω_k Invariant Manifold). (see Appendix A Prop. A.2 for Proof) Let $\mathbf{f}_\theta = \mathbf{A}\mathbf{B}$ be a matrix factorization model, \mathbf{M} be an incomplete matrix and Ω_k be an invariant manifold defined in Prop. 1. If \mathbf{M} is disconnected with m connected components, then there exist m sub- Ω_k manifolds ω_k such that $\omega_k \subsetneq \Omega_k$, each possessing the following properties:

(i) **Invariance under Gradient Flow:** Given data S and the gradient flow dynamics $\dot{\theta} = -\nabla R_S(\theta)$, if the initial point $\theta_0 \in \omega_k$, then $\theta(t) \in \omega_k$ for all $t \geq 0$.

(ii) **Intrinsic Property:** ω_k is a data-value-independent invariant manifold, meaning that for a fixed sampling pattern in \mathbf{M} and any observed values S , ω_k remains invariant under the gradient flow.

(iii) **Strict Subset Relation:** The output set $\{\mathbf{f}_\theta \mid \theta \in \omega_k\}$ is a proper subset of $\{\mathbf{f}_\theta \mid \theta \in \Omega_k\}$, namely, $\{\mathbf{f}_\theta \mid \theta \in \omega_k\} \subsetneq \{\mathbf{f}_\theta \mid \theta \in \Omega_k\}$.

Fig. 5(a) illustrates the trajectory of the experiment in Fig. 4. In the disconnected case, sub- Ω_k invariant manifolds exist and attract the dynamics, leading the model to learn sub-optimal solutions

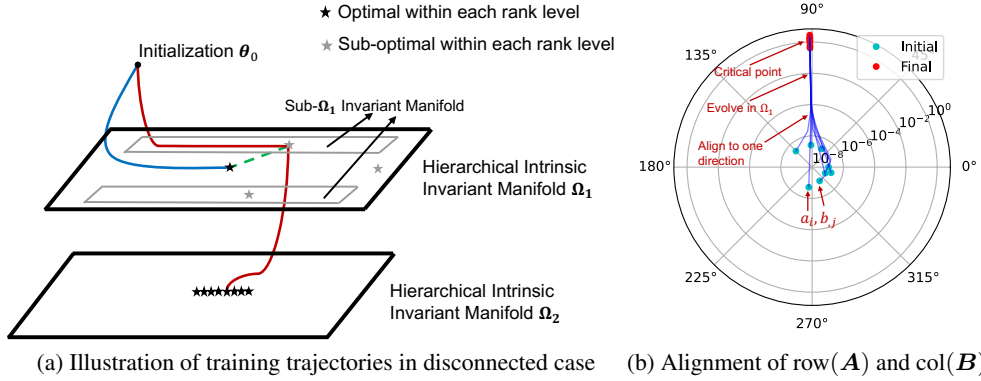


Figure 5: (a) Illustrated trajectories for the experiment in Fig. 4. The blue line represents the trajectory converging to the lowest-rank solution, and the red line represents the actual trajectory experienced by the model. (b) The parameter trajectory escaping from a second-order stationary point to reach the next critical point for the experiment in Fig. 3. The 8 scatter points represent the 4 row vectors of matrix \mathbf{A} and the 4 column vectors of matrix \mathbf{B} . For ease of visualization, we randomly project them onto two dimensions and plot them in polar coordinates.

on the entire Ω_k invariant manifold. In fact, we can prove that these sub-optimal solutions are necessarily strict saddle points. This loss landscape result extends Theorem 5.10 from Li et al. (2020), which established the findings for the specific case of symmetric matrix factorization models (see Appendix A Sec. A.3 for a detailed discussion).

Theorem 1 (Loss Landscape). (see Appendix A Thm. A.3 for Proof) *Given any data S , the critical points of $R_S(\theta)$ are either strict saddle points or global minima.*

Gradient descent easily escapes saddle points (Lee et al., 2016, 2019). Fig. 5(b) shows that when the model escapes a saddle point, the parameters initially appear chaotic but align in one direction after some time, consistent with the “condensation” phenomenon in neural networks (Luo et al., 2021; Zhou et al., 2022). For matrix factorization models, by meticulously analyzing the Hessian matrix structure (see Appendix A.5), we find that this alignment corresponds to an Ω_1 invariant manifold, resulting in a rank increase of one at a time. Under reasonable assumptions, we prove that the training trajectory follows the Ω_k invariant manifold step by step.

Assumption 1 (Top Eigenvalue). Let $\delta\mathbf{M} = (\mathbf{A}_c\mathbf{B}_c - \mathbf{M})_{S_w}$ be the residual matrix at the critical point $\theta_c = (\mathbf{A}_c, \mathbf{B}_c)$. Assume that the top eigenvalue of the matrix $\begin{bmatrix} \mathbf{0} & -\delta\mathbf{M} \\ -\delta\mathbf{M}^\top & \mathbf{0} \end{bmatrix}$ is unique.

Assumption 2 (Second-order Stationary Point). Let Ω be an Ω_k invariant manifold or sub- Ω_k invariant manifold defined in Prop. 1 or 2. Assume θ_c is a second-order stationary point within Ω , i.e., $\nabla R_S(\theta_c) = 0$ and $\theta^\top \nabla^2 R_S(\theta_c) \theta \geq 0$ for all $\theta \in \Omega$.

Theorem 2 (Transition to the Next Rank-level Invariant Manifold). (see Appendix A Thm. A.4 for proof) Consider the dynamics $\dot{\theta} = -\nabla R_S(\theta)$. Let $\varphi(\theta_0, t)$ denote the value of $\theta(t)$ when $\theta(0) = \theta_0$. Let Ω be an Ω_k or sub- Ω_k invariant manifold. Let $\theta_c \in \Omega$ be a critical point satisfying Assump. 1 and 2. Then, for randomly selected θ_0 , with probability 1 with respect to θ_0 , the limit

$$\tilde{\varphi}(\theta_c, t) := \lim_{\alpha \rightarrow 0} \varphi\left(\theta_c + \alpha\theta_0, t + \frac{1}{\lambda_1} \log \frac{1}{\alpha}\right) \quad (5)$$

exists and falls into an invariant manifold Ω_{k+1} . Here λ_1 is the top eigenvalue of $-\nabla^2 R_S(\theta_c)$.

Remark. Assump. 1 ensures that upon departing from a critical point θ_c , the trajectory is constrained to escape along a single dominant eigendirection. Assump. 2 ensures that this escape route aligns with an Ω_1 invariant manifold, leading the trajectory into the Ω_{k+1} invariant manifold. Our Assump. 2 is more general than the conditions posited by Li et al. (2020), as it remains valid across both connected and disconnected configurations, and is consistent with our empirical findings (Figs. 3 and 4). Furthermore, our framework diverges from Li et al. (2020) by emphasizing the role of

hierarchical intrinsic invariant manifolds in directing the training trajectories, a principle potentially extendable to a wider spectrum of models.

Thm. 2 provides a characterization of the escape trajectory. It shows that as the point approaches a second-order stationary point $\theta_c \in \Omega_k$, the trajectory generically converges to a well-defined limit within Ω_{k+1} . Since the origin $\mathbf{0}$ is always a second-order stationary point of Ω_0 , the theorem implies that the trajectory escaping from a small initialization will be close to Ω_1 . This iterative process gives rise to the phenomenon of *Hierarchical Invariant Manifold Traversal (HIMT)*, which involves a sequential progression through these Ω_k manifolds.

6.2 Implicit regularization analysis

Rank minimization is a challenging non-convex optimization problem. Li et al. (2018); Jin et al. (2023) proved that the Restricted Isometry Property (RIP) condition ensures a minimal rank solution. However, the RIP condition is often too stringent for practical matrix completion. For instance, the matrix M_3 in Eq. (4) does not satisfy the RIP criteria, yet the model still finds the minimum rank solution. Our empirical findings (Figs. 1, 2, 3) suggest that a more lenient condition, specifically the connectivity of the observed data, frequently leads to convergence towards the minimal rank solution. Proving this result directly, however, would necessitate a comprehensive examination of the convergence characteristics within each Ω_k invariant manifold, which is an endeavor we leave for future work. Despite this, our insights into the system’s dynamics, i.e., hierarchical invariant manifold traversal, allow us to assert that if a trajectory successfully navigates through the optimal on each rank-level invariant manifold Ω_k , a solution of minimal rank can be achieved naturally.

Theorem 3 (Minimum Rank). *(see Appendix A Thm. A.5 for proof) Consider the dynamics $\dot{\theta} = -\nabla R_S(\theta)$, where $\theta(t) = (\mathbf{A}(t), \mathbf{B}(t))$, and denote $\mathbf{W}_t = \mathbf{A}(t)\mathbf{B}(t)$. Assume \mathbf{W}_t achieves an optimal within each invariant manifold Ω_k . For a full rank initialization \mathbf{W}_0 , if the limit $\widehat{\mathbf{W}} = \lim_{\alpha \rightarrow 0} \mathbf{W}_\infty(\alpha \mathbf{W}_0)$ exists and is a global optimum with $\widehat{\mathbf{W}}_{ij} = M_{ij}$ for all $(i, j) \in S_{\mathbf{x}}$, then*

$$\widehat{\mathbf{W}} \in \operatorname{argmin}_{\mathbf{W}} \operatorname{rank}(\mathbf{W}) \quad \text{s.t.} \quad \mathbf{W}_{ij} = M_{ij}, \forall (i, j) \in S_{\mathbf{x}}. \quad (6)$$

For a disconnected matrix M , our theoretical results (Prop. 2) and experiments (Fig. 4) confirm the existence of sub- Ω_k invariant manifolds. These manifolds attract the training trajectory, leading to sub-optimal solutions and preventing convergence to the lowest-rank solution.

However, in a specific disconnected case, such as disconnection with complete bipartite components, as illustrated in Figs. 1 and 2, the minimum nuclear norm may still serve as a characterization. Gunasekar et al. (2017) proved a special case: if the observations are commutative, then the symmetric model will learn the minimum nuclear norm solution. Intriguingly, for the example M_2 in Eq. (4), even though the observations are not commutative, the model still learns a minimum nuclear norm solution. In fact, we can prove the following result, which aligns well with practical experiments.

Theorem 4 (Minimum Nuclear Norm Guarantee). *(see Appendix A Thm. A.6 for proof) Consider the dynamics $\dot{\theta} = -\nabla R_S(\theta)$, where $\theta(t) = (\mathbf{A}(t), \mathbf{B}(t))$, and let $\mathbf{W}_t = \mathbf{A}(t)\mathbf{B}(t)$. If the observation graph associated with the incomplete matrix M is disconnected with complete bipartite components, and if for a full rank initialization \mathbf{W}_0 , the limit $\widehat{\mathbf{W}} = \lim_{\alpha \rightarrow 0} \mathbf{W}_\infty(\alpha \mathbf{W}_0)$ exists and is a global optimum with $\widehat{\mathbf{W}}_{ij} = M_{ij}$ for all $(i, j) \in S_{\mathbf{x}}$, then*

$$\widehat{\mathbf{W}} \in \operatorname{argmin}_{\mathbf{W}} \|\mathbf{W}\|_* \quad \text{s.t.} \quad \mathbf{W}_{ij} = M_{ij}, \forall (i, j) \in S_{\mathbf{x}}. \quad (7)$$

7 Conclusion and future work

This study presents a comprehensive experimental and theoretical investigation of matrix factorization models. The primary objective was to develop a cohesive framework for understanding the conditions, mechanisms, and reasons behind the diverse implicit regularization effects exhibited by matrix factorization models. A key finding of this research is the pivotal role of the connectivity of observed data in shaping the implicit regularization behavior. To elucidate this phenomenon, we identified the significance of hierarchical invariant manifold traversal within the training dynamics.

Our experiments (Figs. 1, 2, 3) provide strong evidence that connected observed data leads to minimum-rank solutions, as the model learns the optimal of the Ω_k invariant manifold. However,

further investigation is needed to uncover the underlying mechanisms by which connectivity facilitates optimal attainment across different Ω_k invariant manifolds. Additionally, the trade-offs between initialization scale and training efficiency warrant further research, as certain cases may require extremely small initialization, potentially impacting training speed (see Appendix B Sec. B.4).

Generalizing the insights gained from matrix factorization models to other architectures is also an important avenue for future work. Our preliminary experiments indicate that the learning phenomenon from low rank to high rank persists in deep multi-layer matrix factorization and the query-key factorization model in Transformer attention mechanisms (see Appendix B Figs. B7, B9). These findings suggest that the hierarchical invariant manifold traversal process uncovered in our study may have broader implications and merit further exploration.

Acknowledgments and Disclosure of Funding

This work is sponsored by the National Key R&D Program of China Grant No. 2022YFA1008200, the National Natural Science Foundation of China Grant No. 12101402, the Lingang Laboratory Grant No. LG-QS-202202-08, Shanghai Municipal of Science and Technology Major Project No. 2021SHZDZX0102.

References

- Z. Li, Y. Luo, K. Lyu, Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning, in: International Conference on Learning Representations, 2020.
- C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization. *iclr 2017*, arXiv preprint arXiv:1611.03530 (2017).
- C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, *Communications of the ACM* 64 (2021) 107–115.
- S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, N. Srebro, Implicit regularization in matrix factorization, *Advances in neural information processing systems* 30 (2017).
- S. Arora, N. Cohen, W. Hu, Y. Luo, Implicit regularization in deep matrix factorization, *Advances in Neural Information Processing Systems* 32 (2019).
- N. Razin, N. Cohen, Implicit regularization in deep learning may not be explainable by norms, *Advances in neural information processing systems* 33 (2020) 21174–21187.
- Z. Ji, M. Telgarsky, Gradient descent aligns the layers of deep linear networks, in: 7th International Conference on Learning Representations, ICLR 2019, 2019.
- S. Gunasekar, J. D. Lee, D. Soudry, N. Srebro, Implicit bias of gradient descent on linear convolutional networks, *Advances in neural information processing systems* 31 (2018).
- G. Gidel, F. Bach, S. Lacoste-Julien, Implicit regularization of discrete gradient dynamics in linear neural networks, *Advances in Neural Information Processing Systems* 32 (2019).
- D. Gissin, S. Shalev-Shwartz, A. Daniely, The implicit bias of depth: How incremental learning drives generalization, in: International Conference on Learning Representations, 2019.
- L. Jiang, Y. Chen, L. Ding, Algorithmic regularization in model-free overparametrized asymmetric matrix factorization, *SIAM Journal on Mathematics of Data Science* 5 (2023) 723–744.
- M. A. Belabbas, On implicit regularization: Morse functions and applications to matrix factorization, arXiv preprint arXiv:2001.04264 (2020).
- Y. Li, T. Ma, H. Zhang, Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations, in: Conference On Learning Theory, PMLR, 2018, pp. 2–47.
- D. Stöger, M. Soltanolkotabi, Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction, *Advances in Neural Information Processing Systems* 34 (2021) 23831–23843.

- J. Jin, Z. Li, K. Lyu, S. S. Du, J. D. Lee, Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing, in: International Conference on Machine Learning, PMLR, 2023, pp. 15200–15238.
- Y. Zhang, Z. Zhang, L. Zhang, Z. Bai, T. Luo, Z.-Q. J. Xu, Linear stability hypothesis and rank stratification for nonlinear models, arXiv preprint arXiv:2211.11623 (2022).
- Y. Zhang, Z. Zhang, L. Zhang, Z. Bai, T. Luo, Z.-Q. J. Xu, Optimistic estimate uncovers the potential of nonlinear models, arXiv preprint arXiv:2307.08921 (2023).
- A. Jacot, F. Gabriel, C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, Advances in neural information processing systems 31 (2018).
- L. Chizat, E. Oyallon, F. Bach, On lazy training in differentiable programming, Advances in neural information processing systems 32 (2019).
- T. Luo, Z.-Q. J. Xu, Z. Ma, Y. Zhang, Phase diagram for two-layer relu neural networks at infinite-width limit, Journal of Machine Learning Research 22 (2021) 1–47.
- H. Zhou, Z. Qixuan, T. Luo, Y. Zhang, Z.-Q. Xu, Towards understanding the condensation of neural networks at initial training, Advances in Neural Information Processing Systems 35 (2022) 2184–2196.
- Y. Zhang, Z. Zhang, T. Luo, Z. J. Xu, Embedding principle of loss landscape of deep neural networks, Advances in Neural Information Processing Systems 34 (2021) 14848–14859.
- Y. Zhang, Y. Li, Z. Zhang, T. Luo, Z. J. Xu, Embedding principle: A hierarchical structure of loss landscape of deep neural networks, Journal of Machine Learning 1 (2022) 60–113.
- Z. Bai, T. Luo, Z.-Q. J. Xu, Y. Zhang, Embedding principle in depth for the loss landscape analysis of deep neural networks, arXiv preprint arXiv:2205.13283 (2022).
- K. Fukumizu, S. Yamaguchi, Y.-i. Mototake, M. Tanaka, Semi-flat minima and saddle points by embedding neural networks to overparameterization, Advances in Neural Information Processing Systems 32 (2019) 13868–13876.
- B. Simsek, F. Ged, A. Jacot, F. Spadaro, C. Hongler, W. Gerstner, J. Brea, Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 9722–9732.
- A. Jacot, F. Ged, B. Şimşek, C. Hongler, F. Gabriel, Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity, arXiv preprint arXiv:2106.15933 (2021).
- J. D. Lee, M. Simchowitz, M. I. Jordan, B. Recht, Gradient descent only converges to minimizers, in: Conference on learning theory, PMLR, 2016, pp. 1246–1257.
- J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, B. Recht, First-order methods almost always avoid strict saddle points, Mathematical programming 176 (2019) 311–337.
- S. Du, J. Lee, On the power of over-parametrization in neural networks with quadratic activation, in: International conference on machine learning, PMLR, 2018, pp. 1329–1338.

A Proofs of Theoretical Results

In this section, we give all proofs for our theoretical results mentioned in the main text.

A.1 Hierarchical Intrinsic Invariant Manifold and Sub Invariant Manifold

Proposition A.1 (Hierarchical Intrinsic Invariant Manifold (HIIM)). *Let $f_\theta = \mathbf{A}\mathbf{B}$ be a matrix factorization model and $\{\alpha_1, \dots, \alpha_k\}$ be k linearly independent vectors. Define the manifold Ω_k as $\Omega_k := \Omega_k(\alpha_1, \dots, \alpha_k) = \{\theta = (\mathbf{A}, \mathbf{B}) \mid \text{row}(\mathbf{A}) = \text{col}(\mathbf{B}) = \text{span}\{\alpha_1, \dots, \alpha_k\}\}$. The manifold Ω_k possesses the following properties:*

(i) **Invariance under Gradient Flow:** *Given data S and the gradient flow dynamics $\dot{\theta} = -\nabla R_S(\theta)$, if the initial point $\theta_0 \in \Omega_k$, then $\theta(t) \in \Omega_k$ for all $t \geq 0$.*

(ii) **Intrinsic Property:** *Ω_k is a data-independent invariant manifold, meaning that for any data S , Ω_k remains invariant under the gradient flow dynamics.*

(iii) **Hierarchical Structure:** *The manifolds Ω_k form a hierarchy: $\Omega_0 \subsetneq \Omega_1 \subsetneq \dots \subsetneq \Omega_{k-1} \subsetneq \Omega_k$.*

Proof. (i) Invariance under Gradient Flow.

By definition, $\Omega_k := \Omega_k(\alpha_1, \dots, \alpha_k) = \{\theta = (\mathbf{A}, \mathbf{B}) \mid \text{row}(\mathbf{A}) = \text{col}(\mathbf{B}) = \text{span}\{\alpha_1, \dots, \alpha_k\}\}$. Consider the gradient flow dynamics in (8):

$$\begin{cases} \dot{\mathbf{a}}_i = -\frac{2}{n} \sum_{j \in I_i} (\mathbf{a}_i \cdot \mathbf{b}_{\cdot,j} - M_{ij}) \mathbf{b}_{\cdot,j}^\top, \\ \dot{\mathbf{b}}_{\cdot,j} = -\frac{2}{n} \sum_{i \in I_j} (\mathbf{a}_i \cdot \mathbf{b}_{\cdot,j} - M_{ij}) \mathbf{a}_i^\top, \end{cases} \quad (8)$$

where $I_i = \{j \mid \exists i : (i, j) \in S_x\}$, $I_j = \{i \mid \exists j : (i, j) \in S_x\}$, \mathbf{a}_i and $\mathbf{b}_{\cdot,j}$ represent the i -th row and j -th column of \mathbf{A} and \mathbf{B} , respectively.

For any $(i, j) \in S_x$, the evolution of \mathbf{a}_i is coupled with $\mathbf{b}_{\cdot,j}$ for $j \in I_i$. The condition $\text{row}(\mathbf{A}) = \text{col}(\mathbf{B}) = \text{span}\{\alpha_1, \dots, \alpha_k\}$ ensures the existence of k linearly independent vectors $\alpha_1, \dots, \alpha_k \in \mathbb{R}^d$ such that $\mathbf{a}_i, \mathbf{b}_{\cdot,j} \in \text{span}\{\alpha_1, \dots, \alpha_k\}$ for all $1 \leq i, j \leq d$.

Consequently, if \mathbf{a}_i and $\mathbf{b}_{\cdot,j}$ are initially in $\text{span}\{\alpha_1, \alpha_2, \dots, \alpha_k\}$, they will continue to evolve within this subspace under the gradient flow dynamics. Additionally, for $(i, j) \notin S_x$, the gradients for the corresponding \mathbf{a}_i and $\mathbf{b}_{\cdot,j}$ will be zero, provided their initial values are zero, maintaining this state throughout the evolution.

(ii) Intrinsic Property.

As demonstrated in part (i), Ω_k is invariant under gradient flow dynamics for any dataset S , confirming its status as a data-independent invariant manifold.

(iii) Hierarchical Structure.

The invariant manifold Ω_k encompasses matrices of rank up to k , including those of lower ranks. Consequently, the manifolds exhibit the following hierarchical nesting:

$$\Omega_0 \subsetneq \Omega_1 \subsetneq \dots \subsetneq \Omega_{k-1} \subsetneq \Omega_k.$$

□

Proposition A.2 (Intrinsic Sub- Ω_k Invariant Manifold). *Let $f_\theta = \mathbf{A}\mathbf{B}$ be a matrix factorization model, \mathbf{M} be an incomplete matrix and Ω_k be an invariant manifold defined in Prop. 1. If \mathbf{M} is disconnected with m connected components, then there exist m sub- Ω_k manifolds ω_k such that $\omega_k \subsetneq \Omega_k$, each possessing the following properties:*

(i) **Invariance under Gradient Flow:** *Given data S and the gradient flow dynamics $\dot{\theta} = -\nabla R_S(\theta)$, if the initial point $\theta_0 \in \omega_k$, then $\theta(t) \in \omega_k$ for all $t \geq 0$.*

(ii) **Intrinsic Property:** *ω_k is a data-value-independent invariant manifold, meaning that for a fixed sampling pattern in \mathbf{M} and any observed values S , ω_k remains invariant under the gradient flow.*

(iii) **Strict Subset Relation:** The output set $\{\mathbf{f}_\theta \mid \theta \in \omega_k\}$ is a proper subset of $\{\mathbf{f}_\theta \mid \theta \in \Omega_k\}$, namely, $\{\mathbf{f}_\theta \mid \theta \in \omega_k\} \subsetneq \{\mathbf{f}_\theta \mid \theta \in \Omega_k\}$.

Proof. Existence.

Let us consider an incomplete matrix M whose associated observational graph is divided into m connected components, denoted by L_1, L_2, \dots, L_m . For each component L_p , we define $S_x^{L_p}$ as the subset of observed indices within L_p , where $1 \leq p \leq m$ and S_x is the set of all observed indices.

For each L_p , we can identify row indices R_p and column indices C_p corresponding to the observed entries in L_p as follows:

$$R_p = \{i \mid \exists j : (i, j) \in S_x^{L_p}\}, \quad C_p = \{j \mid \exists i : (i, j) \in S_x^{L_p}\}.$$

Here, R_p includes the row indices and C_p includes the column indices of the entries observed in L_p .

Define A^{L_p} and B^{L_p} as the submatrices of A and B corresponding to R_p and C_p , respectively, and let $A_r^{L_p}$ and $B_r^{L_p}$ be the remaining rows not in R_p and C_p .

Let $\Omega_k := \Omega_k(\alpha_1, \dots, \alpha_k) = \{\theta = (A, B) \mid \text{row}(A) = \text{col}(B) = \text{span}\{\alpha_1, \dots, \alpha_k\}\}$ be the given Ω_k invariant manifold.

The sub- Ω_k invariant manifold associated with the connected component L_p can be defined as

$$\omega_k^{L_p} := \{(\theta = (A, B)) \mid \text{row}(A^{L_p}) = \text{col}(B^{L_p}) = \text{span}\{\alpha_1, \dots, \alpha_k\}, A_r^{L_p} = B_r^{L_p} = \mathbf{0}\}. \quad (9)$$

It is easy to check $\omega_k^{L_p}$ is a proper subset of Ω_k .

(i) Invariance under Gradient Flow.

The condition $\text{row}(A^{L_p}) = \text{col}(B^{L_p}) = \text{span}\{\alpha_1, \dots, \alpha_k\}$ along with $A_r^{L_p} = B_r^{L_p} = \mathbf{0}$ guarantees that $\mathbf{a}_i, \mathbf{b}_{\cdot,j} \in \text{span}\{\alpha_1, \dots, \alpha_k\}$ for all $(i, j) \in S_x^{L_p}$, and $\mathbf{a}_i, \mathbf{b}_{\cdot,j} = \mathbf{0}$ for all $(i, j) \notin S_x^{L_p}$.

In other words, the sub- Ω_k invariant manifold $\omega_k^{L_p}$ is the set of all pairs (A, B) where, for each observed position (i, j) in the connected component L_p , the vectors \mathbf{a}_i and $\mathbf{b}_{\cdot,j}$ lie within the span of $\{\alpha_1, \dots, \alpha_k\}$, and for any position not in $S_x^{L_p}$, the vectors are zero.

Considering the dynamics expressed in equation (8), it is evident that the evolution of \mathbf{a}_i is influenced by $\mathbf{b}_{\cdot,j}$ for $(i, j) \in S_x^{L_p}$. Hence, if \mathbf{a}_i and $\mathbf{b}_{\cdot,j}$ are initially in the span of $\{\alpha_1, \alpha_2, \dots, \alpha_k\}$, they will continue to evolve within this span under the gradient flow dynamics. Moreover, for positions $(i, j) \notin S_x^{L_p}$, we consider the following scenarios:

- For $(i, j) \notin S_x$, since the matrix entry M_{ij} does not contribute to the loss $R_S(\theta)$, the gradients for corresponding \mathbf{a}_i and $\mathbf{b}_{\cdot,j}$ will perpetually be zero. Thus, if their initial values are zero, they will remain zero throughout the evolution.
- For $(i, j) \in S_x$ but not in $S_x^{L_p}$, the dynamics corresponding to different connected components are decoupled. Therefore, if the initial values for \mathbf{a}_i and $\mathbf{b}_{\cdot,j}$ are zero, they will stay zero during the evolution.

(ii) Intrinsic Property.

As established in (i), the manifold $\omega_k^{L_p}$ is invariant under gradient flow for any data S with a fixed sampling pattern, qualifying it as a data-value-independent invariant manifold.

(iii) Strict Subset Relation.

The output set $\{\mathbf{f}_\theta \mid \theta \in \Omega_k\}$ encompasses all matrices of rank k , whereas $\{\mathbf{f}_\theta \mid \theta \in \omega_k^{L_p}\}$ is limited to rank- k matrices with specific row and column indices confined to R_p and C_p . Consequently, $\{\mathbf{f}_\theta \mid \theta \in \omega_k^{L_p}\}$ forms a strict subset of $\{\mathbf{f}_\theta \mid \theta \in \Omega_k\}$, as stated by $\{\mathbf{f}_\theta \mid \theta \in \omega_k^{L_p}\} \subsetneq \{\mathbf{f}_\theta \mid \theta \in \Omega_k\}$. \square

A.2 Connectivity

Definition A.1 (Associated Observation Graph). Given an incomplete matrix M to be completed and its observation matrix P , the associated observation graph G_M is the bipartite graph with adjacency matrix $\begin{bmatrix} \mathbf{0} & P^\top \\ P & \mathbf{0} \end{bmatrix}$, with isolated vertices removed.

Definition A.2 (Connectivity). A matrix M to be completed is considered connected if its associated observation graph G_M is connected, otherwise, we call it disconnected. The connected components of M are defined as the connected components of this graph.

Definition A.3 (Disconnectivity with Complete Bipartite Components). A matrix M to be completed is considered disconnected with complete bipartite components if its associated observation graph G_M is disconnected and each connected component forms a complete bipartite subgraph.

Remark. In the bipartite graph representation of the observed data, isolated vertices correspond to entire rows or columns of the matrix M that are not observed. These rows or columns do not contribute to the loss calculation and have no influence on the dynamics of the matrix factorization under infinitesimal initialization. Consequently, when analyzing the connectivity of the observed data and its impact on the learning dynamics, these isolated vertices can be safely disregarded.

Remark. The disconnectivity of the bipartite graph representing the observed data is equivalent to the reducibility of the adjacency matrix $\begin{bmatrix} \mathbf{0} & P^\top \\ P & \mathbf{0} \end{bmatrix}$, where P is the binary observation matrix indicating the positions of the observed entries in M .

In the context of matrix completion problems, such as the Netflix problem, connectivity has a practical interpretation. Connected components in the bipartite graph indicate groups of users and movies that are linked by the users' viewing history. Users within the same connected component are related through the movies they have watched in common. Due to this practical significance, we prefer to use the term "connectivity" instead of "reducibility" when discussing the structure of the observed data in matrix completion problems.

Definition A.4 (Connectivity of Observed Data). Given a matrix M to be completed, an undirected simple graph G can be induced from it: the nodes of the graph are the observed elements in the matrix, and two nodes are adjacent if and only if they are in the same row or column of the matrix M . A matrix M to be completed is considered connected if its induced graph G is connected, otherwise, we call it disconnected.

Lemma A.1. For any simple graph G , if we remove all isolated vertices from G to obtain a new graph G' , then G' is connected if and only if the line graph of G' , denoted as $L(G')$, is connected.

Proof. \implies Assume G' is connected. Consider any two nodes in $L(G')$, which correspond to two edges in G' , say e_1 and e_2 . Since G' is connected, there exists a path connecting the endpoints of e_1 and e_2 . This path corresponds to a sequence of edges in G' , which in turn corresponds to a path connecting the nodes representing e_1 and e_2 in $L(G')$. Therefore, $L(G')$ is connected.

\impliedby Conversely, assume $L(G')$ is connected. Consider any two vertices v_1 and v_2 in G' . Since G' has no isolated vertices, each of v_1 and v_2 is incident to at least one edge. Let these edges be e_1 and e_2 , respectively. Since $L(G')$ is connected, there exists a path connecting the nodes representing e_1 and e_2 in $L(G')$. This path corresponds to a sequence of edges in G' , which in turn corresponds to a path connecting v_1 and v_2 in G' . Therefore, G' is connected.

In conclusion, we have proven that for any simple graph G , if we remove all isolated vertices from G to obtain a new graph G' , then G' is connected if and only if the line graph of G' , denoted as $L(G')$, is connected. \square

Proposition A.3. Given an incomplete matrix M , the connectivity of M defined in Def. A.2 and Def. A.4 is equivalent.

Proof. By definition, each edge of a bipartite graph corresponds to an observed data item, and two edges in a bipartite graph are adjacent if and only if the two corresponding observed data items are in the same row or column. Therefore, the connectivity of the observed data is equivalent to the connectivity of the edges of the bipartite graph, which is, in turn, equivalent to the connectivity of the line graph of the bipartite graph.

According to Lem. A.1, for any graph G , if we remove all isolated vertices from G to obtain a new graph G' , then G' is connected if and only if the line graph of G' , denoted as $L(G')$, is connected.

In the context of the bipartite graph representation of the observed data, removing isolated vertices corresponds to removing rows and columns that contain no observed entries. Thus, the connectivity of the bipartite graph after removing isolated vertices is equivalent to the connectivity of the observed data as defined in Def. A.4.

Consequently, the connectivity of the observed data as defined in Def. A.2 (based on the line graph of the bipartite graph) is equivalent to the connectivity of the observed data as defined in Def. A.4 (based on the connectivity of observed data). \square

Definition A.5 (Decoupling of Dynamics). *Given an incomplete matrix M , consider the gradient flow dynamics of matrix factorization models, $\forall 1 \leq i, j \leq d$,*

$$\begin{cases} \dot{\mathbf{a}}_i = -\frac{2}{n} \sum_{j \in I_i} (\mathbf{a}_i \cdot \mathbf{b}_{\cdot,j} - M_{ij}) \mathbf{b}_{\cdot,j}^\top, \\ \dot{\mathbf{b}}_{\cdot,j} = -\frac{2}{n} \sum_{i \in I_j} (\mathbf{a}_i \cdot \mathbf{b}_{\cdot,j} - M_{ij}) \mathbf{a}_i^\top. \end{cases} \quad (10)$$

The dynamics are said to be decoupled if there exist disjoint subsets of indices $R_1, R_2, \dots, R_k \subseteq \{1, 2, \dots, d\}$ for the rows of \mathbf{A} and $C_1, C_2, \dots, C_k \subseteq \{1, 2, \dots, d\}$ for the columns of \mathbf{B} , such that for each $l \in \{1, 2, \dots, k\}$, the dynamics of $\{\mathbf{a}_i : i \in R_l\}$ and $\{\mathbf{b}_{\cdot,j} : j \in C_l\}$ form an independent system of equations. In other words, the dynamics can be divided into $k(k > 1)$ independent subsystems, each involving a subset of rows of \mathbf{A} and a subset of columns of \mathbf{B} . If such a division is not possible, the dynamics are said to be coupled.

Proposition A.4. *Given an incomplete matrix M , if it is disconnected as defined by Def. A.2, then the dynamics are decoupled as defined by Def. A.5; if it is connected as defined by Def. A.2, then the dynamics are coupled as defined by Def. A.5.*

Proof. Consider a matrix M to be completed, with its associated observation graph comprising m connected components, denoted as L_1, L_2, \dots, L_m . Let $S_x^{L_p} \subseteq S_x$ represent the subset of observed indices corresponding to the connected component L_p , where $1 \leq p \leq m$ and S_x denotes the complete set of observed indices. If the incomplete matrix M is disconnected, then for each connected component L_p , the subset $S_x^{L_p}$ can be partitioned into two subsets R_p and C_p , $1 \leq p \leq m$, such that

$$R_p = \{i | \exists j : (i, j) \in S_x^{L_p}\}, \quad C_p = \{j | \exists i : (i, j) \in S_x^{L_p}\}. \quad (11)$$

In other words, R_p contains the row indices and C_p contains the column indices of the observed entries in the connected component L_p .

It can be easily verified that the dynamics are decoupled in this case, as the subsets $\{R_p, C_p\}_{p=1}^m$ satisfy the conditions in Def. A.5. Each connected component L_p corresponds to an independent subsystem involving the rows of \mathbf{A} indexed by R_p and the columns of \mathbf{B} indexed by C_p .

If M is connected, then its associated observation graph consists of a single connected component, and the entire dynamics are coupled. \square

Examples of connectivity and disconnectivity. Consider three matrices to be completed, each obtained by adding one more observation to the previous matrix: M_1 (disconnected), M_2 (disconnected with complete bipartite components), and M_3 (connected).

$$M_1 = \begin{bmatrix} 1 & 2 & \star \\ 3 & \star & \star \\ \star & \star & 5 \end{bmatrix}, M_2 = \begin{bmatrix} 1 & 2 & \star \\ 3 & 4 & \star \\ \star & \star & 5 \end{bmatrix}, M_3 = \begin{bmatrix} 1 & 2 & \star \\ 3 & 4 & \star \\ 6 & \star & 5 \end{bmatrix}. \quad (12)$$

Fig. A1 illustrates the associated graphs G_M .

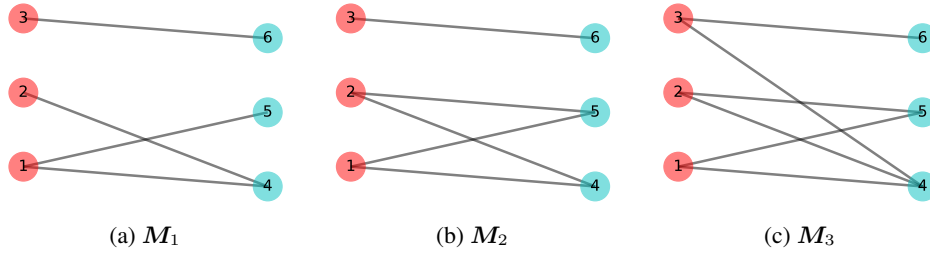


Figure A1: **The associated observation graphs G_M of the incomplete matrices M_1 , M_2 , and M_3 in Eq. 4.** M_1 is disconnected, with its associated observation graph consisting of two connected components. M_2 is also disconnected, but each connected component of its associated observation graph forms a complete bipartite subgraph. In contrast, M_3 is connected, and its associated observation graph consists of a single connected component.

A.3 Loss Landscape

In this paper, we focus on the problem of asymmetric matrix factorization. Previous literature (Gunasekar et al., 2017; Li et al., 2018, 2020; Jin et al., 2023) has predominantly concentrated on symmetric matrix factorization problems. Although asymmetric matrix factorization models can be transformed into symmetric cases, studying symmetric matrix factorization does not necessarily cover all aspects of the asymmetric scenarios.

Generally, an asymmetric matrix factorization model $\mathbf{W} = \mathbf{A}\mathbf{B}$ can be transformed into a symmetric situation by setting

$$\mathbf{U} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B}^\top \end{bmatrix} \in \mathbb{R}^{2d \times d}.$$

We then consider the model $\mathbf{W}' = \mathbf{U}\mathbf{U}^\top$, which corresponds to the following matrix completion problem:

$$\begin{bmatrix} \mathbf{A}\mathbf{A}^\top & \mathbf{A}\mathbf{B} \\ \mathbf{B}^\top\mathbf{A}^\top & \mathbf{B}^\top\mathbf{B} \end{bmatrix}.$$

We define the loss as

$$\mathcal{L}' \left(\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \right) = \frac{1}{2} \mathcal{L}(\mathbf{B}) + \frac{1}{2} \mathcal{L}(\mathbf{C}^\top).$$

Li et al. (2020) established the following results:

Theorem A.1 (Theorem 5.10 in Li et al. (2020)). *Let $f : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ be a convex \mathcal{C}^2 -smooth function. (1). All stationary points of $\mathcal{L} : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$, $\mathcal{L}(\mathbf{U}) = \frac{1}{2} f(\mathbf{U}\mathbf{U}^\top)$ are either strict saddles or global minimizers; (2). For any random initialization, $GF(1)$ converges to strict saddles of $\mathcal{L}(\mathbf{U})$ with probability 0.*

The proof of this theorem relies heavily on Theorem A.2 of Du and Lee (2018), which requires the parameter matrix $\mathbf{U} \in \mathbb{R}^{d \times k}$ to satisfy the condition that $k \geq d$. In the case of symmetric matrix factorization, where $\mathbf{U} \in \mathbb{R}^{d \times d}$, this condition is naturally met. However, for asymmetric matrix factorization, where $\mathbf{U} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B}^\top \end{bmatrix} \in \mathbb{R}^{2d \times d}$, this condition is not satisfied, and thus the proof of Theorem A.1 is only applicable to the symmetric case of matrix factorization.

Theorem A.2 (Theorem 3.1 in Du and Lee (2018)). *Let $f : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ be a \mathcal{C}^2 convex function. Then $\mathcal{L} : \mathbb{R}^{d \times k} \rightarrow \mathbb{R}$, $\mathcal{L}(\mathbf{U}) = f(\mathbf{U}\mathbf{U}^\top)$, $k \geq d$ satisfies that (1). Every local minimizer of \mathcal{L} is also a global minimizer; (2). All saddles are strict. Here saddles denote those stationary points whose hessian are not positive semi-definite (thus including local maximizers).*

Below we give a direct proof of the loss landscape of an asymmetric matrix factorization model.

Theorem A.3 (Loss Landscape). *For any data S , the critical points of $R_S(\boldsymbol{\theta})$ are either strict saddle points or global minima.*

Proof. We start by recalling the definition of the loss function:

$$R_S(\boldsymbol{\theta}) = \mathcal{L}(\mathbf{A}, \mathbf{B}) = \frac{1}{2} \|\mathbf{AB} - \mathbf{M}\|_{S_{\mathbf{x}}}^2 = \frac{1}{2} \sum_{(i,j) \in S_{\mathbf{x}}} ((\mathbf{AB})_{ij} - M_{ij})^2.$$

Let $\boldsymbol{\theta} = (\mathbf{A}, \mathbf{B})$ denote a critical point. We define a new matrix, $\delta\mathbf{M}$, as the difference between the product of \mathbf{A} and \mathbf{B} and the matrix \mathbf{M} , with this difference being computed only over the indices in the set $S_{\mathbf{x}}$. More formally, $\delta\mathbf{M} = (\mathbf{AB} - \mathbf{M})_{S_{\mathbf{x}}}$, where the elements of $\delta\mathbf{M}$ are given by:

- For $(i, j) \in S_{\mathbf{x}}$, we have $\delta\mathbf{M}_{ij} = (\mathbf{AB})_{ij} - M_{ij}$.
- For $(i, j) \notin S_{\mathbf{x}}$, we have $\delta\mathbf{M}_{ij} = 0$.

This definition of $\delta\mathbf{M}$ ensures that we only consider the differences in the entries that belong to the set $S_{\mathbf{x}}$, while all other entries are set to zero.

Consider the function:

$$\begin{aligned} \mathcal{L}(\mathbf{A} + \boldsymbol{\varepsilon}, \mathbf{B} + \boldsymbol{\eta}) &= \frac{1}{2} \|\delta\mathbf{M} + \boldsymbol{\varepsilon}\mathbf{B} + \mathbf{A}\boldsymbol{\eta} + \boldsymbol{\varepsilon}\boldsymbol{\eta}\|_{S_{\mathbf{x}}}^2 \\ &= \frac{1}{2} \|\delta\mathbf{M}\|_{S_{\mathbf{x}}}^2 + \langle \delta\mathbf{M}, \boldsymbol{\varepsilon}\mathbf{B} + \mathbf{A}\boldsymbol{\eta} \rangle_{S_{\mathbf{x}}} + \frac{1}{2} \|\boldsymbol{\varepsilon}\mathbf{B} + \mathbf{A}\boldsymbol{\eta}\|_{S_{\mathbf{x}}}^2 \\ &\quad + \langle \delta\mathbf{M}, \boldsymbol{\varepsilon}\boldsymbol{\eta} \rangle_{S_{\mathbf{x}}} + o(\|\boldsymbol{\varepsilon}\|^2, \|\boldsymbol{\eta}\|^2), \end{aligned} \quad (13)$$

where the inner product of two matrices \mathbf{A}, \mathbf{B} is defined as $\langle \mathbf{A}, \mathbf{B} \rangle := \text{Tr}(\mathbf{AB}^\top)$.

At the critical point, the first order term $\langle \delta\mathbf{M}, \boldsymbol{\varepsilon}\mathbf{B} + \mathbf{A}\boldsymbol{\eta} \rangle_{S_{\mathbf{x}}}$ equals 0. The Hessian operator, representing the second order term, is given by:

$$h_{\mathbf{A}, \mathbf{B}}(\boldsymbol{\varepsilon}, \boldsymbol{\eta}) = \frac{1}{2} \|\boldsymbol{\varepsilon}\mathbf{B} + \mathbf{A}\boldsymbol{\eta}\|_{S_{\mathbf{x}}}^2 + \langle \delta\mathbf{M}, \boldsymbol{\varepsilon}\boldsymbol{\eta} \rangle_{S_{\mathbf{x}}}.$$

Our goal is to demonstrate that if $\delta\mathbf{M} \neq \mathbf{0}$, there always exists $\boldsymbol{\varepsilon}, \boldsymbol{\eta}$ such that $h_{\mathbf{A}, \mathbf{B}}(\boldsymbol{\varepsilon}, \boldsymbol{\eta}) < 0$. To this end, we consider the ranks of matrices \mathbf{A} and \mathbf{B} in two cases:

(i) $\text{rank}(\mathbf{A}) < d$ or $\text{rank}(\mathbf{B}) < d$:

Without loss of generality, we assume $\delta\mathbf{M}_{ij} := \delta\mathbf{M}_{ij} \neq 0$ for some $(i, j) \in S_{\mathbf{x}}$, and $\text{rank}\mathbf{A} < d$. Under these conditions, there exists a non-zero vector \mathbf{v} such that $\mathbf{A}\mathbf{v} = \mathbf{0}$.

We set $\boldsymbol{\eta}_{j}^* = \mathbf{v}$ and $\boldsymbol{\eta}_{s}^* = \mathbf{0}$ for $s \neq j$, where $\boldsymbol{\eta}_{j}^*$ denotes the j -th column of the matrix $\boldsymbol{\eta}$. Let $\boldsymbol{\varepsilon}_i^* = \mathbf{w}^\top \in \mathbb{R}^d$ and $\boldsymbol{\varepsilon}_s = \mathbf{0}$ for $s \neq i$, where $\boldsymbol{\varepsilon}_i^*$ denotes the i -th row of the matrix $\boldsymbol{\varepsilon}$.

We then have:

$$\begin{aligned} h_{\mathbf{A}, \mathbf{B}}(\boldsymbol{\varepsilon}^*, \boldsymbol{\eta}^*) &= \frac{1}{2} \|\boldsymbol{\varepsilon}^* \mathbf{B}\|_{S_{\mathbf{x}}}^2 + \delta\mathbf{M}_{ij} \mathbf{w}^\top \mathbf{v} \\ &\leq \frac{1}{2} \|\mathbf{w}^\top \mathbf{B}\|_{S_{\mathbf{x}}}^2 + \delta\mathbf{M}_{ij} \mathbf{w}^\top \mathbf{v} \\ &= \frac{1}{2} \mathbf{w}^\top \mathbf{B} \mathbf{B}^\top \mathbf{w} + \delta\mathbf{M}_{ij} \mathbf{w}^\top \mathbf{v}. \end{aligned}$$

We define $g(\mathbf{w}, \mathbf{v}) = \frac{1}{2} \mathbf{w}^\top \mathbf{B} \mathbf{B}^\top \mathbf{w} + \delta\mathbf{M}_{ij} \mathbf{w}^\top \mathbf{v}$ and consider:

$$\begin{aligned} g(-\alpha \delta\mathbf{M}_{ij} \mathbf{v}, \mathbf{v}) &= \frac{1}{2} \alpha^2 \delta\mathbf{M}_{ij}^2 \mathbf{v}^\top \mathbf{B} \mathbf{B}^\top \mathbf{v} - \alpha \delta\mathbf{M}_{ij}^2 \mathbf{v}^\top \mathbf{v} \\ &= \frac{1}{2} \alpha^2 \delta\mathbf{M}_{ij}^2 (\mathbf{v}^\top \mathbf{B} \mathbf{B}^\top \mathbf{v} - 2\mathbf{v}^\top \mathbf{v}). \end{aligned}$$

For $0 < \alpha < \frac{2}{\lambda_{\mathbf{B}, \max}}$, where $\lambda_{\mathbf{B}, \max}$ represents the top eigenvalue of $\mathbf{B}\mathbf{B}^\top$, we find $g(-\alpha\delta\mathbf{M}_{ij}\mathbf{v}, \mathbf{v}) < 0$.

Therefore, when $\mathbf{w} = -\alpha\delta\mathbf{M}_{ij}\mathbf{v}$, we obtain $h_{\mathbf{A}, \mathbf{B}}(\boldsymbol{\varepsilon}^*, \boldsymbol{\eta}^*) < 0$. This immediately implies the critical point $\boldsymbol{\theta} = (\mathbf{A}, \mathbf{B})$ is a strict saddle point.

(ii) $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{B}) = d$:

Let $\boldsymbol{\varepsilon} = \alpha\delta\mathbf{M}\mathbf{B}^{-1}$ and $\boldsymbol{\eta} = 0$. In this scenario, the first order term $\langle \delta\mathbf{M}, \boldsymbol{\varepsilon}\mathbf{B} + \mathbf{A}\boldsymbol{\eta} \rangle_{S_{\mathbf{x}}}$ in Eq. (13) simplifies to $\alpha\|\delta\mathbf{M}\|_{S_{\mathbf{x}}}^2$. At a critical point, this quantity equals zero, it implies that $\delta\mathbf{M} = 0$. This in turn implies that the critical point $\boldsymbol{\theta} = (\mathbf{A}, \mathbf{B})$ is a global minimum.

This concludes the proof, establishing that the critical points of $R_S(\boldsymbol{\theta})$ are either strict saddle points or global minima. \square

A.4 Escaping from Top Eigendirection

In this section, we focus on the dynamics of escaping from a critical point. According to Prop. A.3, the loss landscape consists solely of strict saddle points and a global minimum. Consequently, gradient-based methods can readily escape from a critical point that is not a global minimum.

In the following, we will demonstrate that the escaping dynamics near a critical point can be approximated by a linearized version of these dynamics. For this, consider the following:

$$\dot{\boldsymbol{\theta}} = -\nabla R_S(\boldsymbol{\theta}). \quad (14)$$

Assume $\boldsymbol{\theta}_c$ is a saddle point for which $\nabla R_S(\boldsymbol{\theta}_c) = 0$. We can apply a first-order Taylor expansion to the right-hand side of Eq. (14), yielding:

$$-\nabla R_S(\boldsymbol{\theta}) = -\nabla R_S(\boldsymbol{\theta}_c) - \nabla^2 R_S(\boldsymbol{\theta}_c)(\boldsymbol{\theta} - \boldsymbol{\theta}_c) + \mathcal{O}(\|\boldsymbol{\theta} - \boldsymbol{\theta}_c\|^2), \quad (15)$$

where $\nabla^2 R_S(\boldsymbol{\theta}_c)$ represents the Hessian matrix. Given that $\nabla R_S(\boldsymbol{\theta}_c) = 0$, the gradient flow dynamics around $\boldsymbol{\theta}_c$ can be approximated as:

$$\dot{\boldsymbol{\theta}} = \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_c), \quad (16)$$

where $\mathbf{H} := -\nabla^2 R_S(\boldsymbol{\theta}_c)$. Eq. (16) is a classic linear ordinary differential equation, with the solution:

$$\boldsymbol{\theta}(t) = e^{t\mathbf{H}}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_c) + \boldsymbol{\theta}_c. \quad (17)$$

The dynamics near a critical point can be approximated by a linearized version. Hence, in the vicinity of a critical point, we can analyze the linearized dynamics to understand the escape mechanism. In the following, we will show that during this escape process, the dynamics follow a pattern referred to as *dominant eigenvalue dynamics*.

Eq. (17) elucidates that the dynamics near the critical point $\boldsymbol{\theta}_c$ are predominantly dictated by the properties of \mathbf{H} , a real symmetric matrix in $\mathbb{R}^{2d^2 \times 2d^2}$. Its eigendecomposition is given by:

$$\mathbf{H} := -\nabla^2 R_S(\boldsymbol{\theta}_c) = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top, \quad (18)$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix and \mathbf{Q} is an orthogonal matrix. Let $\lambda_1 > \lambda_2 > \dots > \lambda_s \in \mathbb{R}$ denote the eigenvalues of \mathbf{H} , and let $\mathbf{q}_{i1}, \mathbf{q}_{i2}, \dots, \mathbf{q}_{id_i}$ represent the eigenvectors corresponding to λ_i .

Given that $\lambda_1 > \lambda_2$, the ratio $e^{\lambda_1 t} / e^{\lambda_i t}$ for $i > 1$ grows exponentially fast. Consequently, near $\boldsymbol{\theta}_c$, the evolution of the system is primarily driven by the eigenvectors $\mathbf{q}_{11}, \mathbf{q}_{12}, \dots, \mathbf{q}_{1l_1}$ associated with the largest eigenvalue λ_1 .

This following proposition formalizes the intuitive idea that in the vicinity of a saddle point, the dynamics primarily follow the direction associated with the largest eigenvalue. This leading eigendirection becomes increasingly dominant as time evolves, allowing for an escape from the saddle point and facilitating a specific structured transition.

Let's consider $\boldsymbol{\theta}_c$ to be a saddle point. Consider:

$$\dot{\boldsymbol{\theta}} = -\nabla^2 R_S(\boldsymbol{\theta}_c)(\boldsymbol{\theta} - \boldsymbol{\theta}_c), \quad \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \quad (19)$$

Here, θ_0 and θ_c are in close enough proximity for the linearized dynamics to be valid over a sufficiently long period. We can then establish the following proposition:

Proposition A.5 (Escape from Saddle Points Following a Dominant Eigenvalue Dynamics).

Consider the linearized dynamics given by $\dot{\theta} = -\nabla^2 R_S(\theta_c)(\theta - \theta_c)$, we denote $\mathbf{H} := -\nabla^2 R_S(\theta_c)$. Let $\lambda_1 \in \mathbb{R}$ be the largest eigenvalue of \mathbf{H} , with corresponding eigenvectors $\mathbf{q}_{11}, \mathbf{q}_{12}, \dots, \mathbf{q}_{1l_1}$. Denote $c_j = \langle \theta_0 - \theta_c, \mathbf{q}_{1j} \rangle, \forall 1 \leq j \leq l_1$. Assume there exists j such that $c_j \neq 0$, then, given any $\varepsilon > 0$, there exists a $t_0 > 0$ such that for all $t \geq t_0$, the following holds:

$$\left\| \frac{\theta(t) - \theta_c}{e^{\lambda_1 t} \sum_{j=1}^{l_1} c_j \mathbf{q}_{1j}} - \mathbf{1} \right\| < \varepsilon. \quad (20)$$

This means that, as time progresses, the direction of the parameter evolution increasingly aligns with the dominant eigenvectors of \mathbf{H} .

This proposition is a consequence of the fact that the solution of the differential equation is given by $\theta(t) = e^{\mathbf{H}t}\theta(0)$, and as t tends to infinity, the term corresponding to the dominant eigenvalue in the matrix exponential $e^{\mathbf{H}t}$ becomes dominant. Therefore, near the saddle point, we have $\theta(t) \approx$

$$e^{\lambda_1 t} \sum_{j=1}^{l_1} c_j \mathbf{q}_{1j} + \theta_c.$$

Proof. The solution of the ordinary differential equation (19) is $\theta(t) = e^{t\mathbf{H}}(\theta_0 - \theta_c) + \theta_c$. Here, $\mathbf{H} = -\nabla^2 R_S(\theta_c)$ is a real symmetric matrix, which can be diagonalized. Let $\lambda_1 > \lambda_2 > \dots > \lambda_s$ be the eigenvalues of \mathbf{H} , and let $\mathbf{q}_{i1}, \mathbf{q}_{i2}, \dots, \mathbf{q}_{il_i}$ be the eigenvectors corresponding to λ_i . We can then express $\theta(t)$ as:

$$\theta(t) = \sum_{i=1}^s \sum_{j=1}^{l_i} e^{\lambda_i t} \langle \theta_0 - \theta_c, \mathbf{q}_{ij} \rangle \mathbf{q}_{ij} + \theta_c. \quad (21)$$

Next, we analyze the norm of the relative difference between $\theta(t)$ and a term dominating its growth:

$$\begin{aligned} & \left\| \frac{\theta(t) - \theta_c}{\sum_{j=1}^{l_1} e^{\lambda_1 t} \langle \theta_0 - \theta_c, \mathbf{q}_{1j} \rangle \mathbf{q}_{1j}} - \mathbf{1} \right\| = \left\| \frac{\sum_{i=2}^s \sum_{j=1}^{l_i} e^{\lambda_i t} \langle \theta_0 - \theta_c, \mathbf{q}_{ij} \rangle \mathbf{q}_{ij}}{\sum_{j=1}^{l_1} e^{\lambda_1 t} \langle \theta_0 - \theta_c, \mathbf{q}_{1j} \rangle \mathbf{q}_{1j}} \right\| \\ & \leq \sum_{i=2}^s \sum_{j=1}^{l_i} e^{-(\lambda_1 - \lambda_i)t} \left\| \frac{\langle \theta_0 - \theta_c, \mathbf{q}_{ij} \rangle \mathbf{q}_{ij}}{\sum_{j=1}^{l_1} \langle \theta_0 - \theta_c, \mathbf{q}_{1j} \rangle \mathbf{q}_{1j}} \right\| \\ & \leq e^{-(\lambda_1 - \lambda_2)t} \sum_{i=2}^s \sum_{j=1}^{l_i} \left\| \frac{\langle \theta_0 - \theta_c, \mathbf{q}_{ij} \rangle \mathbf{q}_{ij}}{\sum_{j=1}^{l_1} \langle \theta_0 - \theta_c, \mathbf{q}_{1j} \rangle \mathbf{q}_{1j}} \right\|. \end{aligned} \quad (22)$$

We define $C = \sum_{i=2}^s \sum_{j=1}^{l_i} \left\| \frac{\langle \theta_0 - \theta_c, \mathbf{q}_{ij} \rangle \mathbf{q}_{ij}}{\sum_{j=1}^{l_1} \langle \theta_0 - \theta_c, \mathbf{q}_{1j} \rangle \mathbf{q}_{1j}} \right\|$. By choosing $t_0 = \frac{\log \frac{C}{\varepsilon}}{\lambda_1 - \lambda_2}$, we ensure that for all

$t > t_0$, the following condition is met:

$$\left\| \frac{\theta(t) - \theta_c}{\sum_{j=1}^{l_1} e^{\lambda_1 t} \langle \theta_0 - \theta_c, \mathbf{q}_{1j} \rangle \mathbf{q}_{1j}} - \mathbf{1} \right\| < \varepsilon. \quad (23)$$

□

Prop. A.5 describes that under the linearized dynamics, the parameters will escape from the saddle point along a specific direction. However, when considering the original nonlinear dynamics $\dot{\boldsymbol{\theta}}(t) = -\nabla R_S(\boldsymbol{\theta})$, we encounter a trade-off: we should choose t_0 sufficiently large so that the trajectory can align well with the dominant eigendirection while escaping the saddle point, but if t_0 is too large, the linearization approximation will fail as $\boldsymbol{\theta}(t_0)$ moves away from $\boldsymbol{\theta}_c$. Li et al. (2020) (Theorem 5.3) proved a general dynamical result through careful analysis and error control: assuming the eigenvector corresponding to the maximum eigenvalue is unique and the initialization is sufficiently close to the saddle point, there always exists a suitable t_0 such that the linear dynamics can align with the dominant eigendirection before the linearization breaks down. We can generalize this result to the case where the eigenvector corresponding to the largest eigenvalue is not unique:

Proposition A.6. *Consider the dynamics given by $\dot{\boldsymbol{\theta}}(t) = -\nabla R_S(\boldsymbol{\theta})$, we use $\varphi(\boldsymbol{\theta}_0, t)$ to denote the value of $\boldsymbol{\theta}(t)$ in the case of $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$. At a critical point $\boldsymbol{\theta}_c$, we denote the negative Hessian as $\mathbf{H} := -\nabla^2 R_S(\boldsymbol{\theta}_c)$. Let $\lambda_1 \in \mathbb{R}$ be the largest eigenvalue of \mathbf{H} , with corresponding eigenvectors $\mathbf{q}_{11}, \mathbf{q}_{12}, \dots, \mathbf{q}_{1l_1}$. Denote $c_j = \langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}_c, \mathbf{q}_{1j} \rangle, \forall 1 \leq j \leq l_1$, and $\mathbf{v}_1 = \sum_{j=1}^{l_1} c_j \mathbf{q}_{1j}$. Assume there exists j such that $c_j \neq 0$. Let $\mathbf{z}_\alpha(t) := \varphi\left(\boldsymbol{\theta}_c + \alpha \mathbf{v}_1, t + \frac{1}{\lambda_1} \log \frac{1}{\alpha}\right)$ for every $\alpha > 0$, then $\mathbf{z}(t) := \lim_{\alpha \rightarrow 0} \mathbf{z}_\alpha(t)$ exists and is also a solution of the given dynamics, i.e., $\dot{\mathbf{z}}(t) = \varphi(\mathbf{z}(0), t)$. Furthermore, $\forall t \in \mathbb{R}$, there exists a constant $C > 0$ such that*

$$\left\| \varphi\left(\boldsymbol{\theta}_c + \alpha \mathbf{v}_1, t + \frac{1}{\lambda_1} \log \frac{1}{\alpha}\right) - \mathbf{z}(t) \right\|_2 \leq C \alpha^{\frac{\lambda_1 - \lambda_2}{2\lambda_1 - \lambda_2}}$$

for every sufficiently small α , where $\lambda_1 - \lambda_2 > 0$ is the eigenvalue gap.

Proof. By Theorem 5.3 in Section 5.1 of Li et al. (2020), we know that if the eigenspace corresponding to λ_1 is one-dimensional, i.e., $l_1 = 1$, then the escaping direction will be the top eigenvector direction, and the convergence rate is $O(\alpha^{\frac{\lambda_1 - \lambda_2}{2\lambda_1 - \lambda_2}})$. Therefore, Proposition A.6 holds in this case.

Now, if the eigenspace corresponding to λ_1 is not one-dimensional, we denote $c_j = \langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}_c, \mathbf{q}_{1j} \rangle, \forall 1 \leq j \leq l_1$, and $\mathbf{v}_1 = \sum_{j=1}^{l_1} c_j \mathbf{q}_{1j}$ will be the escaping direction. Following the same technique as in Li et al. (2020), we can easily verify that the convergence rate remains $O(\alpha^{\frac{\lambda_1 - \lambda_2}{2\lambda_1 - \lambda_2}})$. Therefore, Proposition A.6 holds in this case as well.

□

A.5 Eigenvalues and Eigenvectors of Hessian

Suppose the dominant directions fulfill specific conditions, such as any combination $c_1 \mathbf{q}_{11} + c_2 \mathbf{q}_{12} + \dots + c_{l_1} \mathbf{q}_{1l_1}$, leading to rank 1 model parameters (\mathbf{A}, \mathbf{B}) . In such scenarios, we may observe a phenomenon where the rank of the matrix increases incrementally.

Firstly, we analyze the eigenvector structure of the Hessian matrix at the critical point $\boldsymbol{\theta}_c = (\mathbf{A}_c, \mathbf{B}_c)$ to understand why the parameter will enter the rank-1 invariant manifold.

Computation of the Hessian Matrix at a Critical Point. To compute the Hessian matrix, we first consider the gradient:

$$\begin{aligned} R_S(\boldsymbol{\theta}) &= \mathbb{E}_S \ell(f(\mathbf{x}, \boldsymbol{\theta}), f^*(\mathbf{x})), \\ \nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}) &= \mathbb{E}_S \nabla \ell(f(\mathbf{x}, \boldsymbol{\theta}), f^*(\mathbf{x}))^\top \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{x}), \\ &= \sum_{i=1}^{d^2} \mathbb{E}_S \partial_i \ell(f_{\boldsymbol{\theta}}, f^*) \nabla_{\boldsymbol{\theta}} (f_{\boldsymbol{\theta}})_i, \\ &= \sum_{i=1}^{d^2} \mathbb{E}_S (f_{\boldsymbol{\theta}} - f^*)_i \nabla_{\boldsymbol{\theta}} (f_{\boldsymbol{\theta}})_i, \end{aligned}$$

where $\partial_i \ell(\mathbf{f}_\theta, \mathbf{f}^*)$ is the i -th element of $\nabla \ell(\mathbf{f}(\mathbf{x}, \theta), \mathbf{f}^*(\mathbf{x}))$, and $(\mathbf{f}_\theta)_i$ is the i -th element of the vectorization of \mathbf{f}_θ .

For the Hessian matrix $\mathbf{H}_S(\theta)$, we have

$$\begin{aligned} \mathbf{H}(\theta) &:= \nabla_\theta \nabla_\theta R_S(\theta) = \sum_{i=1}^{d^2} \mathbb{E}_S \nabla_\theta (\partial_i \ell(\mathbf{f}_\theta, \mathbf{f}^*)) \nabla_\theta (\mathbf{f}_\theta)_i + \sum_{i=1}^{d^2} \mathbb{E}_S \partial_i \ell(\mathbf{f}_\theta, \mathbf{f}^*) \nabla_\theta \nabla_\theta ((\mathbf{f}_\theta)_i) \\ &= \sum_{i,j=1}^{d^2} \mathbb{E}_S \partial_{ij} \ell(\mathbf{f}_\theta, \mathbf{f}^*) \nabla_\theta (\mathbf{f}_\theta)_i \left(\nabla_\theta (\mathbf{f}_\theta)_j \right)^\top + \sum_{i=1}^{d^2} \mathbb{E}_S \partial_i \ell(\mathbf{f}_\theta, \mathbf{f}^*) \nabla_\theta \nabla_\theta ((\mathbf{f}_\theta)_i), \\ &= \sum_{i,j=1}^{d^2} \nabla_\theta (\mathbf{f}_\theta)_i \left(\nabla_\theta (\mathbf{f}_\theta)_j \right)^\top + \sum_{i=1}^{d^2} \mathbb{E}_S (\mathbf{f}_\theta - \mathbf{f}^*)_i \nabla_\theta \nabla_\theta ((\mathbf{f}_\theta)_i), \end{aligned}$$

where $\partial_{ij} \ell(\mathbf{f}_\theta, \mathbf{f}^*)$ is the (i, j) -th element of $\nabla \nabla \ell(\mathbf{f}(\mathbf{x}, \theta), \mathbf{f}^*(\mathbf{x}))$.

We define matrices $\mathbf{H}^{(1)}(\theta)$ and $\mathbf{H}^{(2)}(\theta)$ as follows:

$$\begin{aligned} \mathbf{H}^{(1)}(\theta) &:= \sum_{i,j=1}^{d^2} \nabla_\theta (\mathbf{f}_\theta)_i \left(\nabla_\theta (\mathbf{f}_\theta)_j \right)^\top, \\ \mathbf{H}^{(2)}(\theta) &:= \sum_{i=1}^{d^2} \mathbb{E}_S (\mathbf{f}_\theta - \mathbf{f}^*)_i \nabla_\theta \nabla_\theta ((\mathbf{f}_\theta)_i), \end{aligned}$$

We further denote that $\mathbf{H}(\theta) := \mathbf{H}^{(1)}(\theta) + \mathbf{H}^{(2)}(\theta)$.

For matrix factorization model, the eigenvectors of $\mathbf{H}^{(2)}$ has a special structure, as characterized by Lem. A.2.

Lemma A.2 (Data-Independent Interleaved Structure of Eigenvectors of $\mathbf{H}^{(2)}$). *Let $\theta_c = (\mathbf{A}_c, \mathbf{B}_c)$ be any critical point of the matrix factorization model. If λ is an eigenvalue of $\mathbf{H}^{(2)}(\theta_c) \in \mathbb{R}^{2d^2 \times 2d^2}$, then there exist at least d eigenvectors associated with λ . These d eigenvectors take the form $\mathbf{v} \otimes \mathbf{e}_1, \mathbf{v} \otimes \mathbf{e}_2, \dots, \mathbf{v} \otimes \mathbf{e}_d \in \mathbb{R}^{2d^2}$, where $\mathbf{v} \in \mathbb{R}^{2d}$ is a vector to be determined and \mathbf{e}_i is the unit vector representing the i -th column of the identity matrix $\mathbf{I}_d \in \mathbb{R}^{d \times d}$.*

Proof. Let's denote the residual matrix at the critical point as $\delta \mathbf{M} = (\mathbf{A}_c \mathbf{B}_c - \mathbf{M})_{S_{\mathbf{x}}}$, where $(\mathbf{A}_c, \mathbf{B}_c)$ is a critical point. For the vectorized parameter θ_c , by direct calculation the matrix $\mathbf{H}^{(2)} := -\nabla^2 R_S(\theta_c)$ can be formulated as a block matrix, with the diagonal blocks being 0. The specific format is as follows:

$$\mathbf{H}^{(2)} = \begin{bmatrix} \mathbf{0} & -\delta \mathbf{M} \otimes \mathbf{I}_d \\ -\delta \mathbf{M}^\top \otimes \mathbf{I}_d & \mathbf{0} \end{bmatrix}. \quad (24)$$

Next, we compute the eigenvectors of $\mathbf{H}^{(2)}$. Let λ be an eigenvalue of $\mathbf{H}^{(2)}$. We need to verify that $\mathbf{v} \otimes \mathbf{e}_1, \mathbf{v} \otimes \mathbf{e}_2, \dots, \mathbf{v} \otimes \mathbf{e}_d \in \mathbb{R}^{2d^2}$ are the eigenvectors of $\mathbf{H}^{(2)}$ corresponding to λ , for a particular $\mathbf{v} \in \mathbb{R}^{2d}$ yet to be determined. That is, we need to ensure that for all $1 \leq i \leq d$, the equation $(\mathbf{H}^{(2)} - \lambda \mathbf{I}_{2d^2})(\mathbf{v} \otimes \mathbf{e}_i) = \mathbf{0}$ has a non-zero solution for \mathbf{v} . Notice that

$$\begin{aligned} (\mathbf{H}^{(2)} - \lambda \mathbf{I}_{2d^2})(\mathbf{v} \otimes \mathbf{e}_i) &= \begin{bmatrix} -\lambda \mathbf{I}_d \otimes \mathbf{I}_d & -\delta \mathbf{M} \otimes \mathbf{I}_d \\ -\delta \mathbf{M}^\top \otimes \mathbf{I}_d & -\lambda \mathbf{I}_d \otimes \mathbf{I}_d \end{bmatrix} (\mathbf{v} \otimes \mathbf{e}_i) \\ &= \left(\begin{bmatrix} -\lambda \mathbf{I}_d & -\delta \mathbf{M} \\ -\delta \mathbf{M}^\top & -\lambda \mathbf{I}_d \end{bmatrix} \otimes \mathbf{I}_d \right) (\mathbf{v} \otimes \mathbf{e}_i) \\ &= \left(\begin{bmatrix} -\lambda \mathbf{I}_d & -\delta \mathbf{M} \\ -\delta \mathbf{M}^\top & -\lambda \mathbf{I}_d \end{bmatrix} \mathbf{v} \right) \otimes \mathbf{e}_i. \end{aligned} \quad (25)$$

Since λ is an eigenvalue of $\mathbf{H}^{(2)}$, the determinant of the matrix $\mathbf{H}^{(2)} - \lambda \mathbf{I}_{2d^2}$ equals zero. Hence

$$\det \left(\begin{bmatrix} -\lambda \mathbf{I}_d & -\delta \mathbf{M} \\ -\delta \mathbf{M}^\top & -\lambda \mathbf{I}_d \end{bmatrix} \right)^{2d} = \det \left(\begin{bmatrix} -\lambda \mathbf{I}_d & -\delta \mathbf{M} \\ -\delta \mathbf{M}^\top & -\lambda \mathbf{I}_d \end{bmatrix} \otimes \mathbf{I}_d \right) = 0. \quad (26)$$

Consequently, from Eq. (25), we conclude that there always exists a non-zero vector $\mathbf{v} \in \mathbb{R}^{2d}$ such that $(\mathbf{H}^{(2)} - \lambda \mathbf{I}_{2d^2})(\mathbf{v} \otimes \mathbf{e}_i) = 0$. Since $\mathbf{v} \neq \mathbf{0}$, it is evident that $\mathbf{v} \otimes \mathbf{e}_1, \mathbf{v} \otimes \mathbf{e}_2, \dots, \mathbf{v} \otimes \mathbf{e}_d \in \mathbb{R}^{2d^2}$ are linearly independent, and thus they represent d eigenvectors corresponding to λ . \square

Proposition A.7 (Eigenvectors Structure of H at the Origin). *Consider the dynamics given by Eq. (19), where we denote $\mathbf{H} := -\nabla^2 R_S(\mathbf{0})$. If λ is an eigenvalue of $\mathbf{H} \in \mathbb{R}^{2d^2 \times 2d^2}$, then there exist at least d eigenvectors associated with λ in \mathbf{H} . These d eigenvectors take the form $\mathbf{v} \otimes \mathbf{e}_1, \mathbf{v} \otimes \mathbf{e}_2, \dots, \mathbf{v} \otimes \mathbf{e}_d \in \mathbb{R}^{2d^2}$, where $\mathbf{v} \in \mathbb{R}^{2d}$ is a vector to be determined and \mathbf{e}_i is the unit vector representing the i -th column of the identity matrix $\mathbf{I}_d \in \mathbb{R}^{d \times d}$.*

Proof. In the matrix factorization model, at the origin the gradient $\nabla_{\theta}(\mathbf{f}_{\theta}) = \mathbf{0}$ and thus $\mathbf{H}^{(1)}(\mathbf{0}) = 0$ and the Hessian matrix reduces to $\mathbf{H}^{(2)}(\mathbf{0})$, making $\mathbf{H} = -\nabla^2 R_S(\mathbf{0}) = -\mathbf{H}^{(2)}(\mathbf{0})$.

Let's denote the residual matrix at the origin as $\delta \mathbf{M} = (\mathbf{A}_c \mathbf{B}_c - \mathbf{M})_{S_a}$, where at the origin $(\mathbf{A}_c, \mathbf{B}_c) = (\mathbf{0}, \mathbf{0})$. For the vectorized parameter θ_c , the matrix $\mathbf{H} := -\nabla^2 R_S(\mathbf{0})$ can be formulated as a block matrix, with the diagonal blocks being 0. The specific format is as follows:

$$\mathbf{H} = \mathbf{H}^{(2)} = \begin{bmatrix} \mathbf{0} & -\delta \mathbf{M} \otimes \mathbf{I}_d \\ -\delta \mathbf{M}^\top \otimes \mathbf{I}_d & \mathbf{0} \end{bmatrix}. \quad (27)$$

By Lem. A.2, the proof is completed. \square

Lemma A.3 (Eigenvectors Structure of H at Second-order Stationary Point). *Let Ω denote an Ω_k invariant manifold or sub- Ω_k invariant manifold defined in Prop. A.1 and A.2, and consider a second-order stationary point θ_c within Ω , i.e., $\nabla R_S(\theta_c) = 0$ and $\theta^\top \nabla^2 R_S(\theta_c) \theta \geq 0$ for all $\theta \in \Omega$. Then, the eigenvectors corresponding to the negative eigenvalues of the Hessian matrix $\mathbf{H}(\theta_c)$ are contained within the span of the eigenvectors corresponding to the negative eigenvalues of $\mathbf{H}^{(2)}(\theta_c)$.*

Proof. Recall the definitions of $\mathbf{H}^{(1)}(\theta)$ and $\mathbf{H}^{(2)}(\theta)$ given by:

$$\begin{aligned} \mathbf{H}^{(1)}(\theta) &:= \sum_{i,j=1}^{d^2} \nabla_{\theta}(\mathbf{f}_{\theta})_i (\nabla_{\theta}(\mathbf{f}_{\theta})_j)^\top, \\ \mathbf{H}^{(2)}(\theta) &:= \sum_{i=1}^{d^2} \mathbb{E}_S [(\mathbf{f}_{\theta} - \mathbf{f}^*)_i \nabla_{\theta}^2(\mathbf{f}_{\theta})_i]. \end{aligned}$$

The Hessian matrix $\mathbf{H}(\theta)$ at θ is $\mathbf{H}(\theta) := \mathbf{H}^{(1)}(\theta) + \mathbf{H}^{(2)}(\theta)$.

The manifold Ω is an affine subspace with orthogonal complement denoted by Ω^\perp . Let \mathbf{H} have the following block representation in the bases of Ω and Ω^\perp :

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix}. \quad (28)$$

Since Ω is an invariant subspace under the gradient flow, we have $\mathbf{H}\theta \in \Omega$ for all $\theta \in \Omega$, which implies that $\mathbf{H}_{12} = \mathbf{0}$. Since \mathbf{H} is symmetry, we have $\mathbf{H}_{21} = \mathbf{0}$.

Let $\lambda < 0$ be a negative eigenvalue of $\mathbf{H} := \mathbf{H}(\theta_c)$ with \mathbf{v} as the corresponding eigenvector. Since \mathbf{H}_{11} is positive semi-definite, \mathbf{v} must lie in Ω^\perp .

At a critical point $\theta_c = (\mathbf{A}_c, \mathbf{B}_c)$, by direct calculation, the gradient $\nabla_{\theta} \mathbf{f}_{\theta_c}$ can be structured as:

$$\nabla_{\theta} \mathbf{f}_{\theta_c} = \begin{bmatrix} \mathbf{B}_c & & & \\ & \mathbf{B}_c & & \\ & & \ddots & \\ a_{11} \mathbf{I} & a_{21} \mathbf{I} & \cdots & a_{d1} \mathbf{I} \\ a_{12} \mathbf{I} & a_{22} \mathbf{I} & \cdots & a_{d2} \mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1d} \mathbf{I} & a_{2d} \mathbf{I} & \cdots & a_{dd} \mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{I} \otimes \mathbf{B}_c \\ \mathbf{A}_c^\top \otimes \mathbf{I} \end{bmatrix}_{2d^2 \times d^2}, \quad (29)$$

where \otimes denotes the Kronecker product.

Note that $\nabla_{\theta}(\mathbf{f}_{\theta^*})_j$ is the j -th column of matrix $\nabla_{\theta}f_{\theta_c}$ and it falls precisely within the defined Ω_k invariant manifold or sub- Ω_k invariant manifold Ω . Therefore, we have:

$$\left(\nabla_{\theta}(\mathbf{f}_{\theta_c})_j\right)^{\top} \mathbf{v} = 0 \quad \forall 1 \leq j \leq d^2, \quad (30)$$

which implies that \mathbf{v} is orthogonal to the image of $\nabla_{\theta}(\mathbf{f}_{\theta_c})$, placing it in the null space of $\mathbf{H}^{(1)}(\theta_c)$.

As a result, we have:

$$\mathbf{H}(\theta_c)\mathbf{v} = \left(\mathbf{H}^{(1)}(\theta_c) + \mathbf{H}^{(2)}(\theta_c)\right)\mathbf{v} = \mathbf{H}^{(2)}(\theta_c)\mathbf{v} = \lambda\mathbf{v}.$$

Thus, the eigenvector \mathbf{v} of the Hessian $\mathbf{H}(\theta_c)$, corresponding to the negative eigenvalue λ , is also an eigenvector of $\mathbf{H}^{(2)}(\theta_c)$, confirming that \mathbf{v} is within the span of the eigenvectors of $\mathbf{H}^{(2)}(\theta_c)$. \square

A.6 Transition to the Next Rank-level Invariant Manifold

Proposition A.8. *The linear combination of the eigenvectors of $\mathbf{H}^{(2)}$: $c_1(\mathbf{v} \otimes \mathbf{e}_1) + c_2(\mathbf{v} \otimes \mathbf{e}_2) + \dots + c_d(\mathbf{v} \otimes \mathbf{e}_d)$ falls within the invariant manifold $\Omega_1(\mathbf{c})$, where $\mathbf{c} = (c_1, c_2, \dots, c_d)^{\top}$.*

Proof. Notice that

$$\begin{aligned} & c_1(\mathbf{v} \otimes \mathbf{e}_1) + c_2(\mathbf{v} \otimes \mathbf{e}_2) + \dots + c_d(\mathbf{v} \otimes \mathbf{e}_d) \\ &= \mathbf{v} \otimes [c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \dots + c_d\mathbf{e}_d] \\ &= \mathbf{v} \otimes \mathbf{c}. \end{aligned} \quad (31)$$

By Definition A.1, the data-independent invariant manifold generated by \mathbf{c} is $\Omega_1(\mathbf{c}) = \{(\mathbf{A}, \mathbf{B}) | \mathbf{a}_i, \mathbf{b}_j \in \text{span}\{\mathbf{c}\}, \forall 1 \leq i, j \leq d\}$. If $\theta = (\mathbf{A}, \mathbf{B}) \in \Omega_1(\mathbf{c})$, then \mathbf{A}, \mathbf{B} must take the form

$$\mathbf{A} = \begin{bmatrix} \beta_1\mathbf{c} \\ \beta_2\mathbf{c} \\ \vdots \\ \beta_d\mathbf{c} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \beta_{d+1}\mathbf{c} \\ \beta_{d+2}\mathbf{c} \\ \vdots \\ \beta_{2d}\mathbf{c} \end{bmatrix}, \quad (32)$$

for some $\beta = [\beta_1, \beta_2, \dots, \beta_{2d}]^{\top} \in \mathbb{R}^{2d}$, and the vectorized parameter $\theta = \text{vec}((\mathbf{A}, \mathbf{B})) \in \mathbb{R}^{2d}$ takes the form $\beta \otimes \mathbf{c}$. Let $\beta = \mathbf{v}$, and the proof is complete. \square

Lemma A.4. *Suppose $\alpha_1, \alpha_2, \dots, \alpha_{k+1} \in \mathbb{R}^d$ are linearly independent, the data-independent invariant manifold exhibits the property $\Omega_k(\alpha_1, \alpha_2, \dots, \alpha_k) + \Omega_1(\alpha_{k+1}) = \Omega_{k+1}(\alpha_1, \alpha_2, \dots, \alpha_{k+1})$.*

Proof. Assume that $\theta = (\mathbf{A}, \mathbf{B}) \in \Omega_{k+1}(\alpha_1, \alpha_2, \dots, \alpha_{k+1})$. Then \mathbf{A}, \mathbf{B} should adopt the form:

$$\mathbf{A} = \sum_{i=1}^k \beta_i \alpha_i^{\top} + \beta_{k+1} \alpha_{k+1}^{\top}, \quad \mathbf{B} = \sum_{i=1}^k \gamma_i \alpha_i^{\top} + \gamma_{k+1} \alpha_{k+1}^{\top}. \quad (33)$$

Denote $\mathbf{c}_i = \begin{bmatrix} \beta_i \\ \gamma_i \end{bmatrix} \in \mathbb{R}^{2d}$, then the vectorized parameters $\theta := \text{vec}(\theta)$ can be expressed as:

$$\theta = \sum_{i=1}^k \mathbf{c}_i \otimes \alpha_i + \mathbf{c}_{k+1} \otimes \alpha_{k+1}. \quad (34)$$

Since we know that $\sum_{i=1}^k \mathbf{c}_i \otimes \alpha_i \in \Omega_k(\alpha_1, \alpha_2, \dots, \alpha_k)$ and $\mathbf{c}_{k+1} \otimes \alpha_{k+1} \in \Omega_1(\alpha)$, it is straightforward to validate that $\Omega_{k+1}(\alpha_1, \alpha_2, \dots, \alpha_{k+1}) = \Omega_k(\alpha_1, \alpha_2, \dots, \alpha_k) + \Omega_1(\alpha_{k+1})$. \square

Assumption A.1 (Unique Top Eigenvalue). *Let $\delta\mathbf{M} = (\mathbf{A}_c\mathbf{B}_c - \mathbf{M})_{S_x}$ be the residual matrix at the critical point $\theta_c = (\mathbf{A}_c, \mathbf{B}_c)$. Assume that the top eigenvalue of the matrix $\begin{bmatrix} \mathbf{0} & -\delta\mathbf{M} \\ -\delta\mathbf{M}^{\top} & \mathbf{0} \end{bmatrix}$ is unique.*

Assumption A.2 (Second-order Stationary Point). Let Ω be an Ω_k invariant manifold or sub- Ω_k invariant manifold defined in Prop. A.1 or A.2. Assume θ_c is a second-order stationary point within Ω , i.e., $\nabla R_S(\theta_c) = 0$ and $\theta^\top \nabla^2 R_S(\theta_c) \theta \geq 0$ for all $\theta \in \Omega$.

Theorem A.4 (Transition to the Next Rank-level Invariant Manifold). Consider the dynamics $\dot{\theta} = -\nabla R_S(\theta)$. Let $\varphi(\theta_0, t)$ denote the value of $\theta(t)$ when $\theta(0) = \theta_0$. Let Ω be a Ω_k invariant manifold or sub- Ω_k invariant manifold. Let $\theta_c \in \Omega$ be a critical point satisfying Assump. A.1 and A.2. Then, for randomly selected θ_0 , with probability 1 with respect to θ_0 , the limit

$$\tilde{\varphi}(\theta_c, t) := \lim_{\alpha \rightarrow 0} \varphi \left(\theta_c + \alpha \theta_0, t + \frac{1}{\lambda_1} \log \frac{1}{\alpha} \right) \quad (35)$$

exists and falls into an invariant manifold Ω_{k+1} . Here λ_1 is the top eigenvalue of negative Hessian $-\nabla^2 R_S(\theta_c)$.

Proof. At the critical point θ_c , we denote the negative Hessian as $\mathbf{H} := -\nabla^2 R_S(\theta_c)$. Let $\lambda_1 \in \mathbb{R}$ be the largest eigenvalue of \mathbf{H} , with corresponding eigenvectors $\mathbf{q}_{11}, \mathbf{q}_{12}, \dots, \mathbf{q}_{1l_1}$.

Denote $c_j = \langle \theta_0 - \theta_c, \mathbf{q}_{1j} \rangle, \forall 1 \leq j \leq l_1$, and $\mathbf{v}_1 = \sum_{j=1}^{l_1} c_j \mathbf{q}_{1j}$. For a randomly selected θ_0 , with probability 1, there exists at least one j such that $c_j \neq 0$.

Consider the path $\mathbf{z}_\alpha(t) := \varphi \left(\theta_c + \alpha \mathbf{v}_1, t + \frac{1}{\lambda_1} \log \frac{1}{\alpha} \right)$ for every $\alpha > 0$. By Prop. A.6, the limit $\mathbf{z}(t) := \lim_{\alpha \rightarrow 0} \mathbf{z}_\alpha(t)$ exists and satisfies the dynamics $\dot{\mathbf{z}}(t) = \varphi(\mathbf{z}(0), t)$.

Furthermore, $\forall t \in \mathbb{R}$, there exists a constant $C > 0$ such that

$$\left\| \varphi \left(\theta_c + \alpha \theta_0, t + \frac{1}{\lambda_1} \log \frac{1}{\alpha} \right) - \mathbf{z}(t) \right\|_2 \leq C \alpha^{\frac{\lambda_1 - \lambda_2}{2\lambda_1 - \lambda_2}}$$

for every sufficiently small α , where $\lambda_1 - \lambda_2 > 0$ is the eigenvalue gap.

This implies that the limit $\lim_{\alpha \rightarrow 0} \varphi \left(\theta_c + \alpha \theta_0, t + \frac{1}{\lambda_1} \log \frac{1}{\alpha} \right)$ exists and

$$\tilde{\varphi}(\theta_c, t) := \lim_{\alpha \rightarrow 0} \varphi \left(\theta_c + \alpha \theta_0, t + \frac{1}{\lambda_1} \log \frac{1}{\alpha} \right) = \lim_{\alpha \rightarrow 0} \varphi \left(\theta_c + \alpha \mathbf{v}_1, t + \frac{1}{\lambda_1} \log \frac{1}{\alpha} \right).$$

Assuming $\theta_c \in \Omega_k$ and satisfies Assumps. A.1 and A.2, we aim to show the existence of a rank- $(k+1)$ invariant manifold Ω_{k+1} containing $\theta_c + \alpha \mathbf{v}_1$.

Define the following matrices:

$$\begin{aligned} \mathbf{H}^{(1)}(\theta_c) &:= \sum_{i,j=1}^{d^2} \nabla_{\theta_c} (f_{\theta_c})_i \left(\nabla_{\theta_c} (f_{\theta_c})_j \right)^\top, \\ \mathbf{H}^{(2)}(\theta_c) &:= \sum_{i=1}^{d^2} \mathbb{E}_S [(f_{\theta_c} - f^*)_i \nabla_{\theta_c}^2 (f_{\theta_c})_i]. \end{aligned}$$

The Hessian matrix $\mathbf{H}(\theta_c)$ at θ_c can be expressed as $\mathbf{H}(\theta_c) := \mathbf{H}^{(1)}(\theta_c) + \mathbf{H}^{(2)}(\theta_c)$, and we have $\mathbf{H} = -\mathbf{H}(\theta_c)$.

Assump. A.1 and Lem. A.2 imply that there exist exactly d eigenvectors associated with the top eigenvalue λ_1 of $-\mathbf{H}^{(2)}(\theta_c)$. These eigenvectors are of the form $\mathbf{v} \otimes \mathbf{e}_i$ for $i = 1, \dots, d$, where $\mathbf{v} \in \mathbb{R}^{2d}$ is a vector to be determined and \mathbf{e}_i is the i -th standard basis vector in \mathbb{R}^d . By Assump. A.2 and Lem. A.3, the eigenvectors corresponding to λ_1 of \mathbf{H} are contained within the span of the eigenvectors associated with the negative eigenvalues of $\mathbf{H}^{(2)}(\theta_c)$.

Prop. A.8 ensures that the escaping direction \mathbf{v}_1 lies within a rank-1 invariant manifold Ω_1 . Lem. A.4 then guarantees the existence of an invariant manifold Ω_{k+1} that includes $\theta_c + \alpha \mathbf{v}_1$. Since Ω_{k+1} is invariant under the gradient flow, the trajectory $\varphi \left(\theta_c + \alpha \mathbf{v}_1, t + \frac{1}{\lambda_1} \log \frac{1}{\alpha} \right)$ remains within Ω_{k+1} .

Finally, since Ω_{k+1} is a closed subspace, the limit $\tilde{\varphi}(\theta_c, t)$ lies in $\bar{\Omega}_{k+1}$, concluding the proof. \square

A.7 Minimum Rank

Theorem A.5 (Minimum Rank). *Let Ω denote an invariant as defined previously. Assume \mathbf{W}_t achieves a global minimum within each invariant manifold Ω_k . If the limit $\widehat{\mathbf{W}} = \lim_{\alpha \rightarrow 0} \mathbf{W}_\infty(\alpha I)$ exists and is a global optimum with $\widehat{\mathbf{W}}(i, j) = \mathbf{M}(i, j)$ for all $(i, j) \in S_{\mathbf{x}}$, then*

$$\widehat{\mathbf{W}} \in \operatorname{argmin}_{\mathbf{W}} \operatorname{rank}(\mathbf{W}) \quad \text{s.t.} \quad \mathbf{W}(i, j) = \mathbf{M}(i, j), \forall (i, j) \in S_{\mathbf{x}}. \quad (36)$$

Proof. Consider the invariant manifold Ω_k , which is defined as follows:

$$\Omega_k := \Omega_k(\alpha_1, \dots, \alpha_k) = \{(\mathbf{A}, \mathbf{B}) \mid \mathbf{a}_i, \mathbf{b}_{\cdot, j} \in \operatorname{span}\{\alpha_1, \dots, \alpha_k\}, \forall 1 \leq i, j \leq d\},$$

where \mathbf{a}_i denotes the i -th row of \mathbf{A} , $\mathbf{b}_{\cdot, j}$ denotes the j -th column of \mathbf{B} , and $\alpha_1, \dots, \alpha_k$ are independent vectors that span the invariant subspace associated with Ω_k .

According to Thm. A.4, the training trajectory adheres to a hierarchical traversal across invariant manifolds. For any matrix \mathbf{C} with $\operatorname{rank}(\mathbf{C}) \leq k$, we will show that there always exists $\theta = (\mathbf{A}, \mathbf{B}) \in \Omega_k$ such that $\mathbf{A}\mathbf{B} = \mathbf{C}$. Therefore, Ω_k contains all matrices of rank k .

In fact, Since $\operatorname{rank}(\mathbf{C}) \leq k$, we can express \mathbf{C} as a sum of k rank-one matrices: $\mathbf{C} = \sum_{i=1}^k \mathbf{u}_i \mathbf{v}_i^\top$ where \mathbf{u}_i and \mathbf{v}_i are column vectors. By the definition of Ω_k , for any $\theta = (\mathbf{A}, \mathbf{B}) \in \Omega_k$, each row of \mathbf{A} and each column of \mathbf{B} can be expressed as a linear combination of $\{\alpha_1, \dots, \alpha_k\}$: $\mathbf{a}_i = \sum_{j=1}^k c_{ij} \alpha_j$, $\mathbf{b}_{\cdot, j} = \sum_{i=1}^k d_{ij} \alpha_i$ where c_{ij} and d_{ij} are scalars. We can write:

$$\mathbf{A} = \begin{bmatrix} \sum_{j=1}^k c_{1j} \alpha_j \\ \vdots \\ \sum_{j=1}^k c_{dj} \alpha_j \end{bmatrix}, \mathbf{B} = [\sum_{i=1}^k d_{i1} \alpha_i \quad \dots \quad \sum_{i=1}^k d_{id} \alpha_i].$$

Now, we can express the product $\mathbf{A}\mathbf{B}$ as: $\mathbf{A}\mathbf{B} = \sum_{i=1}^k \sum_{j=1}^k \left(\sum_{l=1}^d c_{li} d_{jl} \right) \alpha_i \alpha_j^\top$. By choosing appropriate values for c_{ij} and d_{ij} , we can make $\mathbf{A}\mathbf{B} = \mathbf{C}$. This is possible because the outer products $\alpha_i \alpha_j^\top$ span the same subspace as the rank-one matrices $\mathbf{u}_i \mathbf{v}_i^\top$ in the expression of \mathbf{C} .

Therefore, for any matrix \mathbf{C} with $\operatorname{rank}(\mathbf{C}) \leq k$, there always exists $\theta = (\mathbf{A}, \mathbf{B}) \in \Omega_k$ such that $\mathbf{A}\mathbf{B} = \mathbf{C}$.

If the output matrix \mathbf{W}_t attains optimums within each Ω_k , it suggests that the optimization process is selecting the best approximation from the set of all possible rank- k matrices. Provided that each step in the optimization is optimal, the resulting solution will naturally be the matrix with the lowest feasible rank that satisfies the matrix completion criteria, thereby completing the proof. \square

A.8 Minimum Nuclear Norm Guarantee

Lemma A.5 (Minimal Nuclear Norm Computation). *Given a matrix \mathbf{M} to be completed with observed diagonal entries, i.e., $\operatorname{diag}(\mathbf{M}) = \mathbf{v}$, the minimal nuclear norm solution among all possible completions is $\|\mathbf{v}\|_1$.*

Proof. The nuclear norm of a matrix is the dual of the spectral norm $\|\cdot\|_2$, defined as:

$$\|\mathbf{A}\|_* = \max_{\|\mathbf{X}\|_2 \leq 1} \langle \mathbf{A}, \mathbf{X} \rangle.$$

Given that $\|\operatorname{diag}(\operatorname{sign}(\mathbf{v}))\|_2 \leq 1$, for any matrix \mathbf{A} with $\operatorname{diag}(\mathbf{A}) = \mathbf{v}$, it follows that:

$$\|\mathbf{A}\|_* \geq \langle \mathbf{A}, \operatorname{diag}(\operatorname{sign}(\mathbf{v})) \rangle = \langle \mathbf{v}, \operatorname{sign}(\mathbf{v}) \rangle = \|\mathbf{v}\|_1.$$

Specifically, the nuclear norm of the diagonal matrix with \mathbf{v} on its diagonal is $\|\operatorname{diag}(\mathbf{v})\|_* = \|\mathbf{v}\|_1$, which establishes that the diagonal matrix with \mathbf{v} is indeed a minimizer for the nuclear norm. \square

Diagonal Observations

Proposition A.9 (Minimum Nuclear Norm Guarantee in Diagonal Case). *Consider the dynamics $\dot{\boldsymbol{\theta}} = -\nabla R_S(\boldsymbol{\theta})$, where $\boldsymbol{\theta}(t) = (\mathbf{A}(t), \mathbf{B}(t))$ and denote $\mathbf{W}_t = \mathbf{A}(t)\mathbf{B}(t)$. If the observation data is diagonal, and if for a full rank initialization \mathbf{W}_0 , the limit $\widehat{\mathbf{W}} = \lim_{\alpha \rightarrow 0} \mathbf{W}_\infty(\alpha \mathbf{W}_0)$ exists and is a global optimum with $\widehat{\mathbf{W}}_{ij} = M_{ij}$ for all $(i, j) \in S_{\mathbf{x}}$, then*

$$\widehat{\mathbf{W}} \in \operatorname{argmin}_{\mathbf{W}} \|\mathbf{W}\|_* \quad \text{s.t.} \quad \mathbf{W}_{ij} = M_{ij}, \forall (i, j) \in S_{\mathbf{x}}. \quad (37)$$

Proof. Without loss of generality, assume that M is a diagonal matrix given by:

$$\mathbf{M} = \begin{bmatrix} \mu_1 & & & \\ & \mu_2 & & \\ & & \ddots & \\ & & & \mu_d \end{bmatrix}.$$

By Lem. A.5, the minimal nuclear norm among all possible completions is $|\mu_1| + |\mu_2| + \dots + |\mu_d|$. When the matrix to be completed is diagonal, the evolution of the i -th row of \mathbf{A} is influenced only by the i -th column of \mathbf{B} . Hence, the dynamics decouple into d independent parts, each equivalent to learning a scalar μ_i . The learning process thus unfolds in d stages, with each stage passing through a critical point to learn a respective μ_i .

By Lem. A.2, the second term of the Hessian matrix can be expressed as:

$$\mathbf{H}^{(2)} = \begin{bmatrix} \mathbf{0} & -\delta \mathbf{M} \otimes \mathbf{I}_d \\ -\delta \mathbf{M}^\top \otimes \mathbf{I}_d & \mathbf{0} \end{bmatrix}.$$

According to Lem. A.2, the d eigenvectors of $\mathbf{H}^{(2)}$ take the form $\mathbf{v} \otimes \mathbf{e}_1, \mathbf{v} \otimes \mathbf{e}_2, \dots, \mathbf{v} \otimes \mathbf{e}_d \in \mathbb{R}^{2d^2}$, where $\mathbf{v} \in \mathbb{R}^{2d}$ is a vector to be determined and \mathbf{e}_i is the unit vector corresponding to the i -th column of the identity matrix $\mathbf{I}_d \in \mathbb{R}^{d \times d}$.

(i) Suppose $\mu_1 > \mu_2 > \dots > \mu_d > 0$.

With an infinitesimal initialization, the training dynamics first focus on the element with the largest singular value, then proceed sequentially to blocks with smaller singular values. This pattern of learning is consistent with the concept of ‘‘sequential learning’’ as reported in the literature (Gidel et al., 2019; Gissin et al., 2019; Jiang et al., 2023).

For a diagonal observation matrix, the residual matrix $\delta \mathbf{M}$ at any critical point remains a diagonal matrix. While starting to learn μ_i from a critical point, direct calculation confirms that vector $\mathbf{v} = [\mathbf{e}_i, \mathbf{e}_i]^\top \in \mathbb{R}^{2d}$. Escaping from each saddle point $\boldsymbol{\theta}_c$, the trajectory $\boldsymbol{\theta}(t) - \boldsymbol{\theta}_c$ approximates $\sum_{i=1}^d c_i(\mathbf{v} \otimes \mathbf{e}_i)$, which satisfies $\mathbf{B}_{,i} = (\mathbf{A}^\top)_{,i}$. Thus, learning a diagonal matrix M using the asymmetric model $\mathbf{A}\mathbf{B}$ is equivalent to using a symmetric model $\mathbf{A}\mathbf{A}^\top$. The final outcome ensures that $\operatorname{diag}(\mathbf{A}\mathbf{B}) = \operatorname{diag}(\mathbf{A}\mathbf{A}^\top) = \operatorname{diag}(M)$.

The nuclear norm of $\mathbf{A}\mathbf{A}^\top$ equals the sum of its eigenvalues, which is precisely the trace of the matrix, and $\operatorname{tr}(\mathbf{A}\mathbf{A}^\top) = \operatorname{tr}(M) = \mu_1 + \mu_2 + \dots + \mu_d$. Therefore, the nuclear norm of the learned matrix $\mathbf{W} = \mathbf{A}\mathbf{B} = \mathbf{A}\mathbf{A}^\top$ remains $|\mu_1| + |\mu_2| + \dots + |\mu_d|$.

(ii) If some $\mu_i < 0$, assume without loss of generality that $|\mu_1| > |\mu_2| > \dots > |\mu_n| > 0$.

While starting to learn μ_i from a critical point, direct calculation confirms that $\mathbf{v} = [\mathbf{e}_i, \operatorname{sign}(\mu_i)\mathbf{e}_i]^\top \in \mathbb{R}^{2d}$. Escaping from each saddle point $\boldsymbol{\theta}_c$, the trajectory $\boldsymbol{\theta}(t) - \boldsymbol{\theta}_c$ approximates $\sum_{i=1}^d c_i(\mathbf{v} \otimes \mathbf{e}_i)$, satisfying $\mathbf{B}_{,i} = \operatorname{sign}(\mu_i)(\mathbf{A}^\top)_{,i}$. Hence, $\mathbf{A}\mathbf{B} = \mathbf{A}\mathbf{A}^\top \mathbf{Q}$, where \mathbf{Q} is an orthogonal matrix given by:

$$\mathbf{Q} = \begin{bmatrix} \operatorname{sign}(\mu_1) & & & \\ & \operatorname{sign}(\mu_2) & & \\ & & \ddots & \\ & & & \operatorname{sign}(\mu_d) \end{bmatrix}.$$

The final result ensures that $\text{diag}(\mathbf{AB}) = \text{diag}(\mathbf{AA}^\top \mathbf{Q}) = \text{diag}(\mathbf{M})$, meaning $\text{diag}(\mathbf{AA}^\top) = \text{diag}(\mathbf{QM})$. The nuclear norm of \mathbf{AA}^\top equals the sum of its eigenvalues, which is the trace of the matrix, and $\text{tr}(\mathbf{AA}^\top) = \text{tr}(\mathbf{QM}) = |\mu_1| + |\mu_2| + \dots + |\mu_d|$.

Since an orthogonal transformation does not change the nuclear norm of a matrix, the nuclear norm of the final learned matrix $\mathbf{W} = \mathbf{AB} = \mathbf{AA}^\top \mathbf{Q}$ is still $|\mu_1| + |\mu_2| + \dots + |\mu_d|$. \square

Disconnected with Complete Bipartite Components

Theorem A.6 (Minimum Nuclear Norm Guarantee). *Consider the dynamics $\dot{\theta} = -\nabla R_S(\theta)$, where $\theta(t) = (\mathbf{A}(t), \mathbf{B}(t))$, and let $\mathbf{W}_t = \mathbf{A}(t)\mathbf{B}(t)$. If the observation graph associated with the matrix \mathbf{M} to be completed is disconnected with complete bipartite components, and if for a full rank initialization \mathbf{W}_0 , the limit $\widehat{\mathbf{W}} = \lim_{\alpha \rightarrow 0} \mathbf{W}_\infty(\alpha \mathbf{W}_0)$ exists and is a global optimum with $\widehat{\mathbf{W}}_{ij} = \mathbf{M}_{ij}$ for all $(i, j) \in S_{\mathbf{x}}$, then*

$$\widehat{\mathbf{W}} \in \text{argmin}_{\mathbf{W}} \|\mathbf{W}\|_* \quad \text{s.t.} \quad \mathbf{W}_{ij} = \mathbf{M}_{ij}, \forall (i, j) \in S_{\mathbf{x}}. \quad (38)$$

Proof. Consider a matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ composed of m connected components, with each component forming a complete bipartite subgraph. Since \mathbf{M} is disconnected, it can be represented in a block diagonal form without loss of generality:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & & & \\ & \mathbf{M}_2 & & \\ & & \ddots & \\ & & & \mathbf{M}_m \end{bmatrix},$$

where each block $\mathbf{M}_i \in \mathbb{R}^{d_i \times d'_i}$, and $\sum_{i=1}^m d_i = d, \sum_{i=1}^m d'_i = d$, representing the sum of the dimensions of the blocks.

Each block \mathbf{M}_i corresponds some singular values of the corresponding Hessian matrix at a critical point. With a infinitesimal initialization, the training dynamics first focus on the block with the largest singular value, then proceed sequentially to blocks with smaller singular values. This pattern of learning is consistent with the concept of ‘‘sequential learning’’ as reported in the literature (Gidel et al., 2019; Gissin et al., 2019; Jiang et al., 2023).

Since each connected component of \mathbf{M} forms a complete bipartite subgraph, the block \mathbf{M}_i is fully observed. We can do singular value decomposition (SVD) on each sub-block \mathbf{M}_i as $\mathbf{M}_i = \mathbf{U}_i \Sigma_i \mathbf{V}_i^\top$, where \mathbf{U}_i and \mathbf{V}_i are orthogonal matrices, and Σ_i is a diagonal matrix with the singular values of \mathbf{M}_i .

Construct block diagonal matrices \mathbf{U} and \mathbf{V} as follows:

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 & & & \\ & \mathbf{U}_2 & & \\ & & \ddots & \\ & & & \mathbf{U}_m \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & & & \\ & \mathbf{V}_2 & & \\ & & \ddots & \\ & & & \mathbf{V}_m \end{bmatrix}.$$

This leads to the diagonal matrix:

$$\mathbf{UMV}^\top = \begin{bmatrix} \Sigma_1 & & & \\ & \Sigma_2 & & \\ & & \ddots & \\ & & & \Sigma_m \end{bmatrix} = \begin{bmatrix} \mu_1 & & & \\ & \mu_2 & & \\ & & \ddots & \\ & & & \mu_d \end{bmatrix}.$$

Orthogonal transformations preserve the nuclear norm, so by Lem. A.5, the minimal nuclear norm among all possible completions is the sum of the absolute values of the diagonal entries, i.e., $|\mu_1| + |\mu_2| + \dots + |\mu_d|$.

Consider an incomplete matrix \mathbf{M} whose associated observational graph is divided into m connected components, denoted by L_1, L_2, \dots, L_m . For each component L_p , we define $S_{\mathbf{x}}^{L_p}$ as the subset of observed indices within L_p , where $1 \leq p \leq m$ and $S_{\mathbf{x}}$ is the set of all observed indices.

For each L_p , we can identify row indices R_p and column indices C_p corresponding to the observed entries in L_p as follows:

$$R_p = \{i|\exists j : (i, j) \in S_{\mathbf{x}^{L_p}}\}, \quad C_p = \{j|\exists i : (i, j) \in S_{\mathbf{x}^{L_p}}\}.$$

Here, R_p includes the row indices and C_p includes the column indices of the entries observed in L_p .

Define \mathbf{A}^{L_p} and \mathbf{B}^{L_p} as the submatrices of \mathbf{A} and \mathbf{B} corresponding to R_p and C_p , respectively. The evolution of \mathbf{A}^{L_p} is influenced only by \mathbf{B}^{L_p} . Thus, the dynamics decouple into m independent parts, each equivalent to learning a fully observed matrix \mathbf{M}_i .

Accordingly, we partition \mathbf{A} and \mathbf{B} into m blocks:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_m \end{bmatrix}, \quad \mathbf{B} = [\mathbf{B}_1 \quad \cdots \quad \mathbf{B}_m],$$

where $\mathbf{A}_i = \mathbf{A}^{L_i}$ and $\mathbf{B}_j = \mathbf{B}^{L_j}$.

Denote $L_i = \|\mathbf{A}_i \mathbf{B}_i - \mathbf{M}_i\|_2^2$. The overall loss function $L = \sum_{i=1}^m L_i$ can be decomposed into m independent parts. Performing orthogonal transformations $\tilde{\mathbf{A}}_i = \mathbf{U}_i^\top \mathbf{A}_i$ and $\tilde{\mathbf{B}}_i = \mathbf{B}_i \mathbf{V}_i$, we obtain a diagonal loss $\tilde{L}_i = \|\tilde{\mathbf{A}}_i \tilde{\mathbf{B}}_i - \Sigma_i\|_2^2$ for $1 \leq i \leq C$.

Since gradient descent is the steepest descent in the l_2 norm and orthogonal transformations preserve this norm, the dynamics of optimizing \tilde{L}_i are equivalent to those of optimizing L_i .

Without loss of generality, assume $\mu_1 > \mu_2 > \cdots > \mu_d > 0$. Otherwise, as with Prop. A.9, a sign orthogonal transformation \mathbf{Q} can be applied without changing the nuclear norm.

By Prop. A.9, the learning result for a diagonal matrix implies $\tilde{\mathbf{B}}_i = \tilde{\mathbf{A}}_i^\top$, which means $\mathbf{B}_i \mathbf{V}_i = \mathbf{U}_i^\top \mathbf{A}_i^\top$.

The final learning result

$$\tilde{\mathbf{A}} = \begin{bmatrix} \tilde{\mathbf{A}}_1 \\ \vdots \\ \tilde{\mathbf{A}}_m \end{bmatrix}, \quad \tilde{\mathbf{B}} = [\tilde{\mathbf{B}}_1 \quad \cdots \quad \tilde{\mathbf{B}}_m],$$

satisfies $\tilde{\mathbf{B}} = \tilde{\mathbf{A}}^\top$.

The result ensures that $\text{diag}(\tilde{\mathbf{A}} \tilde{\mathbf{B}}) = \text{diag}(\tilde{\mathbf{A}} \tilde{\mathbf{A}}^\top) = \text{diag}(\Sigma)$. The nuclear norm of $\tilde{\mathbf{A}} \tilde{\mathbf{A}}^\top$ equals the sum of its eigenvalues, which is the trace of the matrix, and $\text{Tr}(\tilde{\mathbf{A}} \tilde{\mathbf{A}}^\top) = \text{Tr}(\Sigma) = |\mu_1| + |\mu_2| + \cdots + |\mu_d|$.

Since orthogonal transformations do not alter the nuclear norm of a matrix, the nuclear norm of $\mathbf{W} = \mathbf{A} \mathbf{B}$ is also $|\mu_1| + |\mu_2| + \cdots + |\mu_d|$, concluding the proof. \square

B Experimental Setup and Supplementary Experiments

In this section, we present the supplementary experiments mentioned in the main text and the details of experiments.

B.1 Experimental Setup

For all our experiments, we employ gradient descent with a carefully chosen small learning rate. A learning rate is deemed suitable when it yields a smooth, monotonically decreasing training trajectory for the loss function, free from any abrupt fluctuations or oscillations. We initialize all model parameters using a Gaussian distribution with a mean of zero and a variance that is detailed for each specific experiment. Because of the small size of the experiment, the experiment can be completed on a single CPU.

The criterion for the sufficiency of training in all cases is a training loss that falls below 10^{-10} . To ascertain the rank of the matrix produced by the learning process, we utilize a technique of

extrapolation with an infinitesimally small initialization. As depicted in Fig. 3(b), if a singular value persistently diminishes in response to decreasing initialization magnitudes, it is then inferred that such a singular value will not contribute to the rank in the context of an infinitesimal initialization.

We have included code in the Supplementary Material that determines the connectivity of a partially observed matrix and provides specific examples illustrating the implicit regularization effects. This code can be used to reproduce our results and explore the relationship between data connectivity and the implicit biases of matrix factorization models in various matrix completion scenarios.

B.2 Connectivity Experiments

In the connectivity experiments corresponding to Fig. 1, we explore the behavior of randomly generated 4×4 matrices with intrinsic ranks of 1, 2, and 3. To investigate the impact of sampling density on matrix reconstruction, we sample matrices at three different levels: $2rd - r^2$, which meets the threshold for exact reconstruction, $2rd - r^2 - 1$, which is just below the threshold, and $2rd - r^2 + 1$, which exceeds the threshold.

For each sampling size, we randomly generate 10 sets of sampling positions. We then assess the connectivity of the sampled positions and compute both the rank and the nuclear norm of the solutions obtained through gradient descent. As an illustration, in Fig. B1, panel (a) presents a scenario with connected sampling positions, panel (b) shows disconnected sampling positions, and panel (c) depicts disconnected sampling with each disconnected component forming a complete bipartite graph.

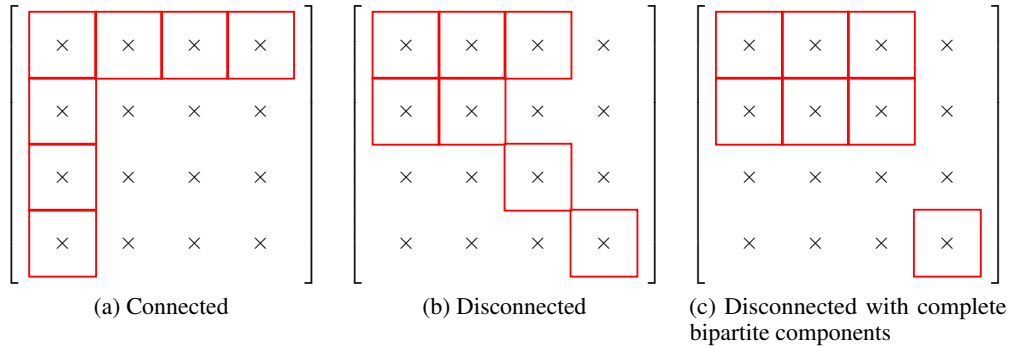


Figure B1: Examples of connected sampling and disconnected sampling patterns in Fig. 1.

In the connectivity experiments depicted in Figs. 2(c-d), we examine the behavior of randomly generated matrices of size 4×4 and 10×10 with a rank of 1. The matrices are sampled at a size of $2rd - r^2$, which corresponds to the threshold for exact reconstruction. We evaluate two connected and one disconnected sampling patterns.

Fig. B2(a) displays the first connected sampling pattern, where all entries in the first row and the first column are sampled. Fig. B2(b) illustrates the second connected sampling pattern, which forms a “Z” shape across the matrix. Fig. B2(c) shows the disconnected sampling pattern, where the samples are split into two unconnected blocks, one in the top-left and the other in the bottom-right of the matrix. A similar approach is taken for the 10×10 matrices.

For Figs. 2 (a-b), we performed 100 random initializations for each initialization scale, recorded the mean and standard deviation, and plotted them on the figure. For a sequence x_1, x_2, \dots, x_n , the mean is calculated by $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, and the standard deviation is calculated by:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Figs. 2(c-d) demonstrate that when the target matrix has a rank of 1 and the number of samples meets the minimum requirement for reconstruction with connected sampling positions, the matrix factorization model is capable of accurately reconstructing the original target matrix.

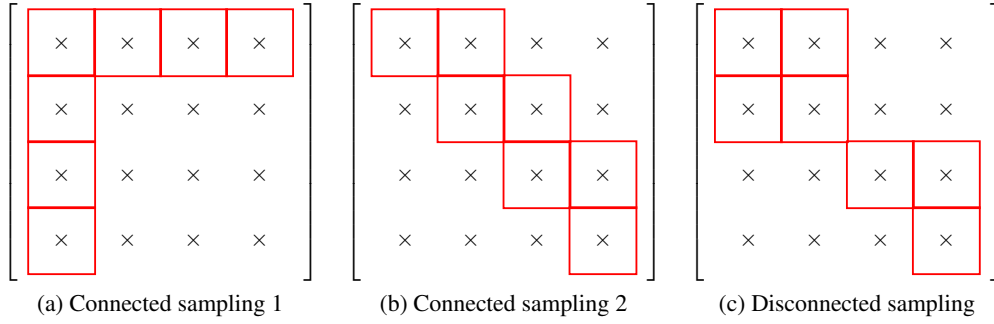


Figure B2: Examples of connected sampling and disconnected sampling patterns in Fig. 2(c).

In scenarios where the target matrix has a higher rank, we have extended our experiments accordingly. For a randomly chosen 4×4 matrix with rank 2, we selected a sample count less than the threshold of $2rd - r^2 = 12$, specifically 10 samples, while ensuring that the sampling pattern is connected, as shown in Fig. B3. The resulting solution from the matrix completion has a rank of 2, which is the minimal rank that fits the sampled data.

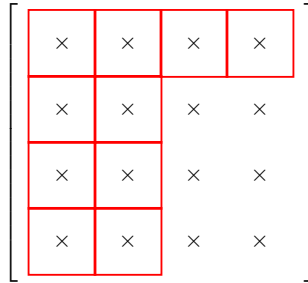


Figure B3: The sample pattern in Fig. B4.

Fig. B4 reveals distinct behaviors of the matrix completion depending on the scale of initialization. With a larger initialization, the third and fourth singular values of the completed matrix remain relatively significant, suggesting that the model does not converge to the lowest rank solution. On the other hand, with a smaller initialization, the third and fourth singular values are uniformly small, indicating that the model successfully converges to the lowest rank solution.

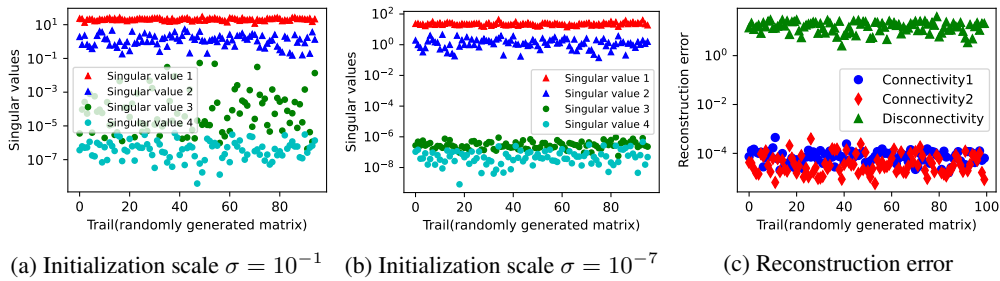


Figure B4: For a randomly selected 4×4 matrix with rank 2, we chose 10 samples, fewer than the threshold $2rd - r^2 = 12$, ensuring connected sampling positions, as shown in Fig. B3. The figures show the experimental results for the four singular values of the matrix learned under Gaussian initialization with mean 0 and standard deviations of 10^{-1} (a) and 10^{-7} (b), respectively. (c) Reconstruction error of the solutions for a 4×4 matrix reconstruction problem with M^* randomly sampled at rank $r = 1$ and sample size set to the minimum reconstruction setting $n = 2rd - r^2$. Red and blue scatter points represent two connected sampling patterns, while green points represent a disconnected pattern.

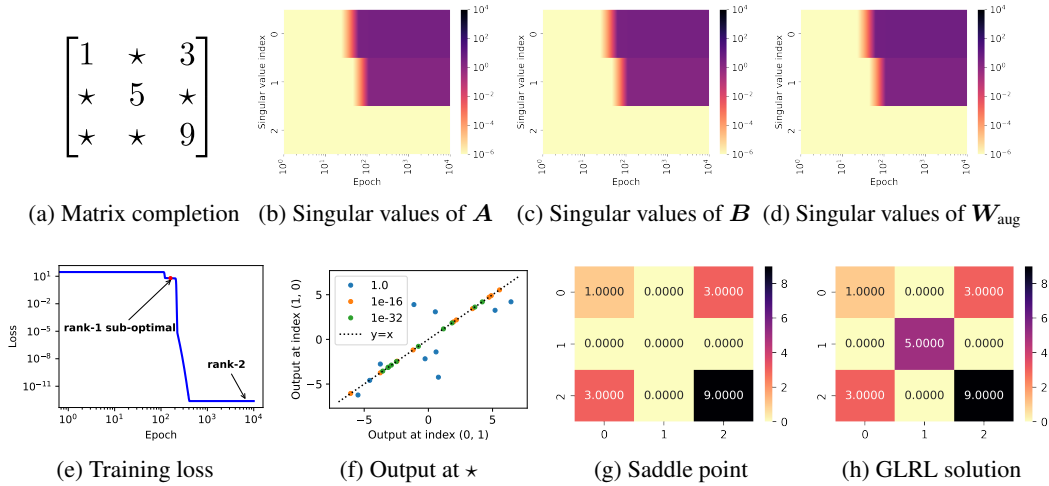


Figure B5: (a) The matrix to be completed, with unknown entries marked by \star . (b-d) Evolution of singular values for \mathbf{A} , \mathbf{B} , and \mathbf{W}_{aug} during training. (e) Training loss for the disconnected sampling pattern. (f) Learned values at symmetric positions $(0, 1)$ and $(1, 0)$ under varying initialization scales (zero mean, varying variance). Each point represents one of ten random experiments per variance; labels show initialization variance. Other symmetric positions exhibit similar behavior. (g) Learned output at the saddle point corresponding to the red dot in (e). (h) Final learned solution of the GLRL algorithm (Li et al., 2020).

B.3 Equivalent Sampling Patterns

For a given sample size, there are different sampling models corresponding to connected or disconnected sampling. As shown in Fig. B1, 7 observations are sampled, but different sampling positions affect connectivity or disconnection. To thoroughly study all possible cases, we examine all sampling cases of a 3×3 matrix completion, as illustrated in Figure 2(c).

For a 3×3 matrix, the sample size varies from 1 to 9. When using the matrix decomposition model $\mathbf{f}_\theta = \mathbf{A}\mathbf{B}$ for matrix completion, the dynamics obtained by exchanging rows or columns or transposing the matrix to be completed are equivalent. These three operations allow us to divide all sampling patterns equally.

For sample size = 1, there is only one sampling pattern in the equivalent sense, and the observation matrix \mathbf{P} is:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

where 1 indicates that the position is observed and non-zero, and 0 means that the position is not observed or is 0.

For sample size = 2, there are only 2 sampling patterns in the equivalent sense:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{P}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

For sample size = 3, there are only 4 sampling patterns in the equivalent sense:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{P}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{P}_3 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{P}_4 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

For sample size = 4, there are only 5 sampling patterns in the equivalent sense:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{P}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{P}_3 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{P}_4 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{P}_5 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

For sample size = 5, there are only 3 sampling patterns in the equivalent sense:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{P}_2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \mathbf{P}_3 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

For sample size = 6, there are only 4 sampling patterns in the equivalent sense:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{P}_2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{P}_3 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \mathbf{P}_4 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

For sample size = 7, there are only 2 sampling patterns in the equivalent sense:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \mathbf{P}_2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

For sample size = 8, there is only 1 sampling pattern in the equivalent sense:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

For sample size = 9, there is only 1 sampling pattern in the equivalent sense:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

B.4 Initialization Scale Analysis

Our experimental findings indicate that when the observational data is connected, matrix factorization models often learn the lowest-rank solution starting from a small initialization. However, the required scale of initialization is not constant across different instances. We empirically observed that if the magnitude of the numerical values in the matrix to be completed varies significantly, an extremely small initialization is necessary, which, in some cases, can exceed machine precision.

Consider the following two simple 2×2 matrix completion problems, with the only difference being that the number 3 in the first row is replaced by 20. When training begins from a small initialization, for \mathbf{M}_4 , the fourth element only needs to learn the value 6 to be a rank-1 solution. However, for \mathbf{M}_5 , the fourth element needs to learn the value 40 to achieve rank-1.

$$\mathbf{M}_4 = \begin{bmatrix} 1 & 2 \\ 3 & \times \end{bmatrix}, \quad \mathbf{M}_5 = \begin{bmatrix} 1 & 2 \\ 20 & \times \end{bmatrix}.$$

Fig. B6 illustrates the difficulty in learning these two examples. For \mathbf{M}_4 , an initialization variance of approximately 10^{-7} is sufficient to learn the lowest-rank solution. In contrast, for \mathbf{M}_5 , an extremely small initialization variance is required, making it challenging to learn a rank-1 solution. Yet, with an exceedingly small initialization variance of 10^{-83} , we can still observe the second singular value plummeting to zero. The origin is a saddle point. The smaller the initialization, the longer it will stay at the origin. If initialization continues to decrease, training will stagnate. Therefore, if the magnitude difference is even greater, such as replacing 20 with 100 in \mathbf{M}_5 , then with the small initialization allowed by machine precision, it is nearly impossible to learn the value 200 completely.

B.5 Dynamics of Deep Matrix Factorization

In the context of depth-3 matrix factorization models, we consider the functional form:

$$\mathbf{f}_\theta = \mathbf{ABC}, \quad \text{where } \mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{d \times d}.$$

Figs. B7 and B8 suggest that even for a depth-3 model, the learning process exhibits a progression from low rank to high rank structures.

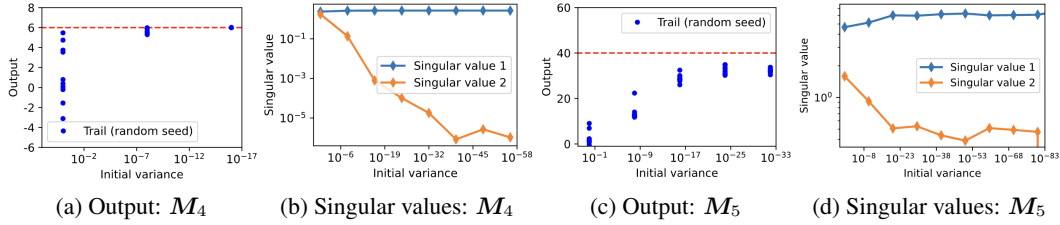


Figure B6: (a, c) The value of the (2, 2) element learned by the model as the initialization decreases. (b, d) The singular values of the matrix learned by the model with decreasing initialization.

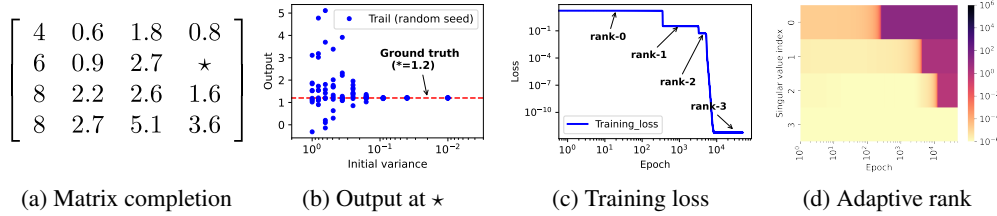


Figure B7: **Deep Matrix factorization models learn adaptively from low rank to high rank with small initialization.** (a) The matrix to be completed, the \star position is unknown. (b) The value at \star learned under different initialization scales (mean is zero, variance changes), 10 random experiments were done under each variance, and each blue point represents an experiment. (c) The training loss curve with an initial variance of 10^{-14} . (d) Evolution of singular values for $f_\theta = ABC$ during training. The count of significantly non-zero singular values is indicative of the rank.

B.6 Incorporating Attention Mechanisms

Within the Transformer architecture, the matrix factorization component retains its significance. The attention mechanism is formalized as follows:

$$f_\theta(X) = \sum_{i=1}^h \text{softmax}_{\text{row}} \left(\frac{XW_{Q_i}W_{K_i}^\top X^\top}{\sqrt{d_k}} \right) XW_{V_i}W_{O_i},$$

where the row-wise softmax operation is applied to the attention scores, and the sum is over the h different attention heads, with $W_{Q_i}, W_{K_i}, W_{V_i}, W_{O_i}$ representing the learnable weight matrices for queries, keys, values, and output transformations, respectively, and d_k is the dimensionality of the key vectors.

The attention module’s ability to capture low-rank representations is reflected in the depth-2 matrix factorization model. As illustrated in Fig. B9, the attention models consistently learns representations that evolve from lower to higher ranks.

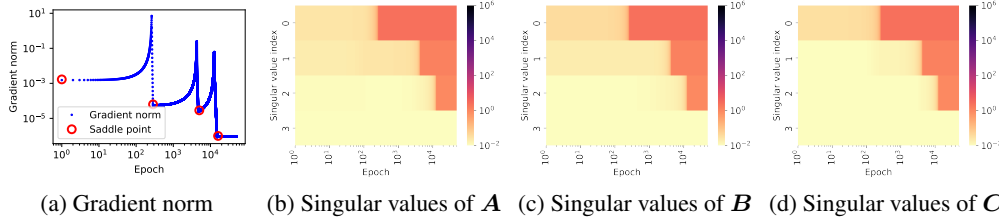


Figure B8: **Deep Matrix factorization models learn adaptively from low rank to high rank with small initialization.** (a) Evolution of the l_2 -norm of the gradients for all parameters throughout the training process. (b-d) Evolution of singular values for matrices $A, B,$ and C during training. The count of non-zero singular values is indicative of the rank.

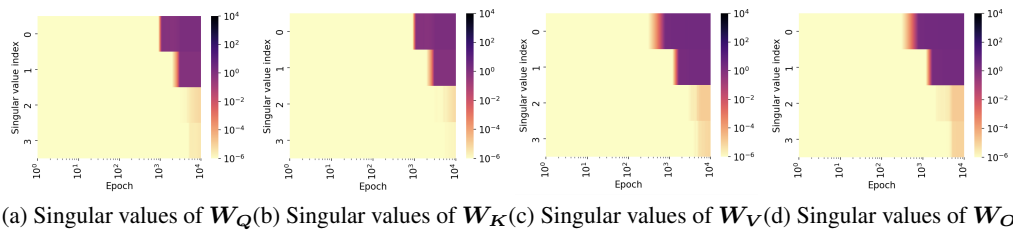


Figure B9: The attention modules in Transformer learn adaptively from low rank to high rank with small initialization. (a-d) The evolution of singular values for the matrices W_Q , W_K , W_V , and W_O throughout the training process. The number of significantly non-zero singular values suggests the effective rank of each matrix.