上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Connectivity Shapes Implicit Regularization in Matrix Factorization Models for Matrix Completion

Zhiwei Bai
Shanghai Jiao Tong University
School of Mathematical Sciences & Institute of Natural Sciences

September 26, 2024

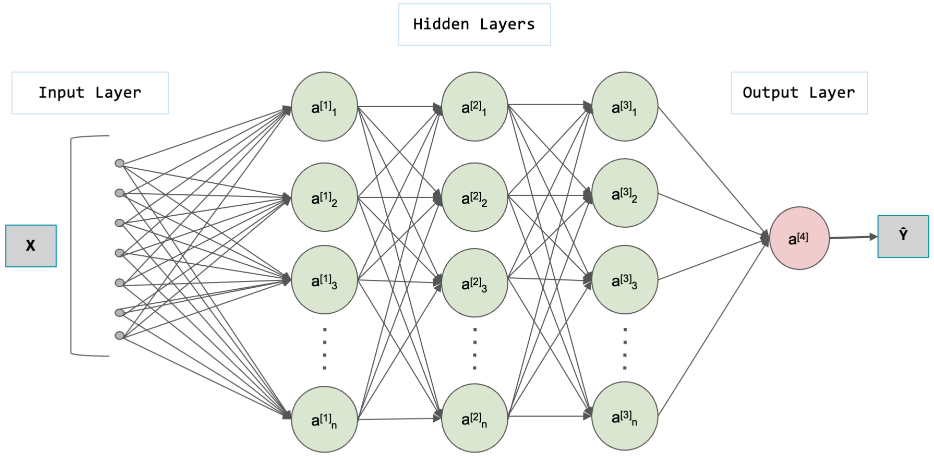
目录

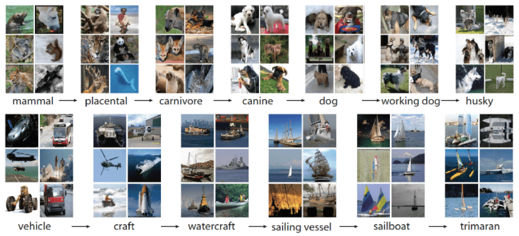
CONTENTS

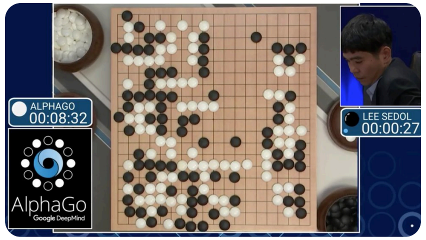
1. Introduction and Motivation

Background: DNNs as Function Approximator

- **Deep Neural Networks (DNNs)** have achieved remarkable success in various fields.

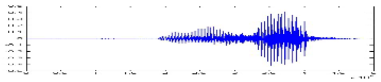






- **DNNs as Function Approximator**

$f($  $)$ = "Cat"

$f($  $)$ = "How are you"

$f($  $)$ = "5-5" (next move)

- **Key Structure: Composition of Functions Layer by Layer**

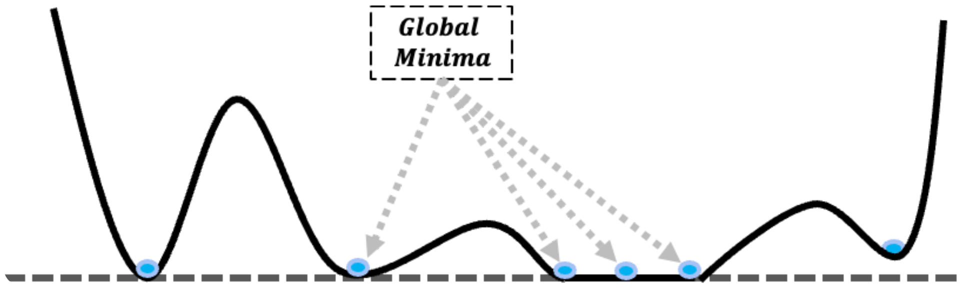
$$\boldsymbol{f}_{\boldsymbol{\theta}}^{[l]}(\boldsymbol{x}) = \sigma(\boldsymbol{W}^{[l]}\boldsymbol{f}_{\boldsymbol{\theta}}^{[l-1]}(\boldsymbol{x}) + \boldsymbol{b}^{[l]}), l = 1, 2, \cdots, L-1.$$

$$\boldsymbol{f}_{\boldsymbol{\theta}}^{[l]}(\boldsymbol{X}) = \sum_{i=1}^{h} \mathrm{softmax}_{\mathrm{row}}\left(\frac{\boldsymbol{X}\boldsymbol{W}_{Q_i}\boldsymbol{W}_{K_i}^{\top}\boldsymbol{X}^{\top}}{\sqrt{d_k}}\right)\boldsymbol{X}\boldsymbol{W}_{V_i}\boldsymbol{W}_{O_i}$$

Background: How to Understand the Learning Behavior?

- Theory: Understanding the learning behavior
- DNNs: Overparameterization

$$\#param >> \#data$$



Q: Which Global Minimum is learned?

- Mathematical Formulation
  - Empirical risk:

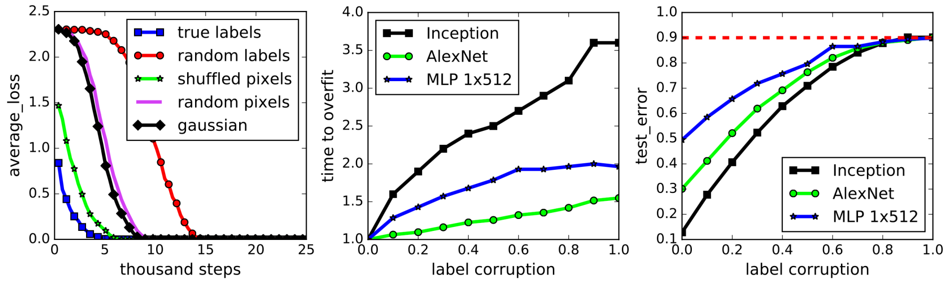
$$R_S(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{f}(\boldsymbol{x}_i;\boldsymbol{\theta}), \boldsymbol{y}_i)$$

  - Model: $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})$
  - Data: $S = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$
  - Loss function: $\ell(\cdot,\cdot)$
  - Learning dynamics: $\dot{\boldsymbol{\theta}} = -\nabla R_S(\boldsymbol{\theta})$ with $\boldsymbol{\theta}_0 \sim N(\mathbf{0}, \sigma^2)$

How to analyze the learning dynamics?

Background: the Generalization Mystery

- DNNs' capacity is very large



(a) learning curves (b) convergence slowdown (c) generalization error growth

Sufficiently large for memorizing the entire random dataset

Q: Is explicit regularization necessary?

[Zhang et al.] Understanding deep learning requires rethinking generalization. ICLR 2017 (Best Paper)

- DNNs generalize well without explicit regularization

model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75
(fitting random labels)		no	no	100.0	9.78
Inception w/o BatchNorm	1,649,402	no	yes	100.0	83.00
		no	no	100.0	82.00
(fitting random labels)		no	no	100.0	10.12
Alexnet	1,387,786	yes	yes	99.90	81.22
		yes	no	99.82	79.66
		no	yes	100.0	77.36
		no	no	100.0	76.07
(fitting random labels)		no	no	99.82	9.86
MLP 3x512	1,735,178	no	yes	100.0	53.35
		no	no	100.0	52.39
(fitting random labels)		no	no	100.0	10.48
MLP 1x512	1,209,866	no	yes	99.80	50.39
		no	no	100.0	50.51
(fitting random labels)		no	no	99.34	10.61

Not Sufficient

Not Necessary

Explicit regularization may improve generalization performance, but is neither necessary nor sufficient

The Generalization Mystery $\Longrightarrow$ Implicit Regularization

- Matrix Completion

$$\begin{bmatrix} 1 & 2 & 3 \\ \star & 4 & \star \\ \star & \star & 9 \end{bmatrix} \Longrightarrow \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

- Non–Factorization Model (Overparameterization):

$$\boldsymbol{f_\theta} = \boldsymbol{W} \in \mathbb{R}^{d\times d}, \boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{W}) \in \mathbb{R}^{d^2}$$

- Linear w.r.t. $\boldsymbol{\theta}$
- Convex Optimization

$R_S(\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^{n}((\boldsymbol{f_\theta})_{i_k j_k} - \boldsymbol{M}_{i_k j_k})^2$

- Implicit Regularization ($\dot{\boldsymbol{\theta}} = -\nabla R_S(\boldsymbol{\theta})$)

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 = \|\boldsymbol{W} - \boldsymbol{W}_0\|_F$$

- Matrix Completion

$$\begin{bmatrix} 1 & 2 & 3 \\ \star & 4 & \star \\ \star & \star & 9 \end{bmatrix} \Longrightarrow \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$$

- Matrix Factorization Model (Composition Structure, Overparameterization)

$$\boldsymbol{f_\theta} = \boldsymbol{AB} \in \mathbb{R}^{d\times d}, \boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$$

- Non–Linear w.r.t. $\boldsymbol{\theta}$
- Non–Convex Optimization

$R_S(\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^{n}((\boldsymbol{f_\theta})_{i_k j_k} - \boldsymbol{M}_{i_k j_k})^2$

- Implicit Regularization ($\dot{\boldsymbol{\theta}} = -\nabla R_S(\boldsymbol{\theta})$)

?????????????

Recent Works on Implicit Regularization in Matrix Factorization

[Gunasekar et al.] 2017 NeurIPS

[Arora et al.] 2019 NeurIPS

[Jin et al.] 2023 ICML

Nuclear Norm Minimization

Theorem 1. In the case where the observation matrices $\{\mathcal{A}_i\}_{i=1}^m$ commute, the symmetrical matrix factorization model $f_\theta = UU^\top$ finds the minimal nuclear norm solution.

Nuclear Norm Minimization

Theorem 2. In the case where the observation matrices $\{\mathcal{A}_i\}_{i=1}^m$ commute, the asymmetrical matrix factorization model $f_\theta = AB$ finds the minimal nuclear norm solution.

Rank Minimization

Theorem 3. In the case where the observation matrices $\{\mathcal{A}_i\}_{i=1}^m$ satisfy the RIP condition, the symmetrical matrix factorization model $f_\theta = UU^\top$ finds the minimal rank solution.

Are these characterizations sufficient? Do they describe the whole picture of matrix factorization models?

Examples

- Observation Matrices Commute:
  $E_{ij}E_{mn} = \delta_{jm}E_{in} = E_{mn}E_{ij} = \delta_{ni}E_{mj}$

$$\Longrightarrow \begin{bmatrix} \times & \star & \star & \checkmark \\ \times & \star & \star & \star \\ \times & \star & \star & \star \\ \times & \times & \times & \times \end{bmatrix}$$

- Counterexample:

$$\begin{bmatrix} 0 & 1 \\ 2 & \star \end{bmatrix} \overset{GD}{\Longrightarrow} \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ \star & 3 \end{bmatrix} \overset{GD}{\Longrightarrow} \begin{bmatrix} 1 & 2 \\ 1.5 & 3 \end{bmatrix}$$

- GD still learned the minimal nuclear norm solution although the observation matrices do not commute

- Restricted Isometry Property (RIP):
  The measurement operator $\mathcal{A}$ satisfies the (δ, r) RIP if
  $(1-\delta)\|\boldsymbol{Z}\|_{\mathrm{F}}^2 \le \|\mathcal{A}(\boldsymbol{Z})\|_2^2 \le (1+\delta)\|\boldsymbol{Z}\|_{\mathrm{F}}^2$
  for all $\boldsymbol{Z} \in \mathbb{R}^{d\times d}$ with $\mathrm{rank}(\boldsymbol{Z}) \le r$

- Counterexample:

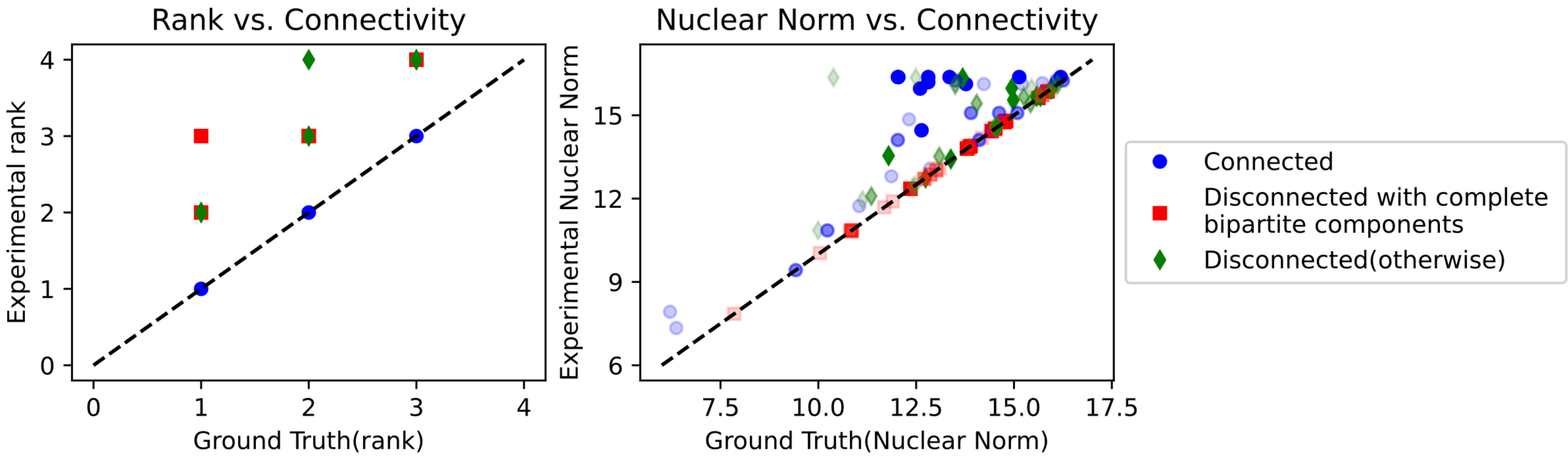
$$\begin{bmatrix} 1 & 2 \\ 3 & \star \end{bmatrix} \overset{GD}{\Longrightarrow} \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 10 & \star \end{bmatrix} \overset{GD}{\Longrightarrow} \begin{bmatrix} 1 & 2 \\ 10 & 20 \end{bmatrix}$$

- GD still learned the minimal rank solution although the observation matrices do not satisfy the (δ, r) RIP condition

How to construct a unified understanding of when, how, and why they achieve different implicit regularization effects?

Empirical Observations

- The connectivity of observed data affects the implicit regularization



- Low rank bias in connected case
- Low nuclear norm bias in certain disconnected case

2. Connectivity Affects Implicit Regularization

Definition of Connectivity

Observation matrix $\boldsymbol{P}$

$$\boldsymbol{P}_{ij} = \begin{cases} 1, & \boldsymbol{M}_{ij} \text{ is observed and non-zero} \\ 0, & \text{otherwise} \end{cases}$$

Associated Observation Graph G_M

Definition 1 (Associated Observation Graph). The associated observation graph G_M is the bipartite graph with adjacency matrix $\begin{bmatrix} \mathbf{0} & \boldsymbol{P}^\top \\ \boldsymbol{P} & \mathbf{0} \end{bmatrix}$, with isolated vertices removed.

Connectivity

Definition 2. Connected: G_M is connected; **Disconnected:** G_M is disconnected
The connected components of M are defined as the connected components of G_M.

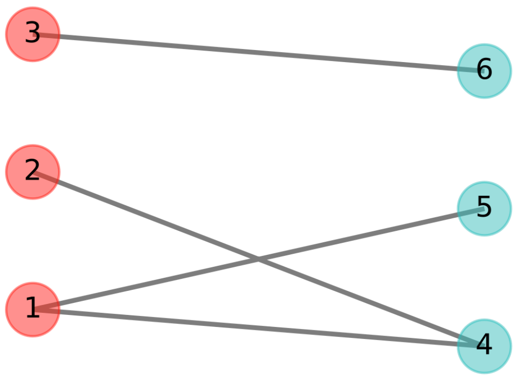
Examples of Connectivity

Disconnectivity with Complete Bipartite Components

Definition 3. Disconnectivity with Complete Bipartite Components: Graph G_M is disconnected and each connected component forms a complete bipartite subgraph.

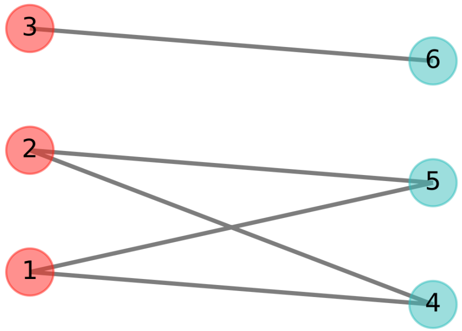
- Disconnected

$$M_1 = \begin{bmatrix} 1 & 2 & \star \\ 3 & \star & \star \\ \star & \star & 5 \end{bmatrix}$$



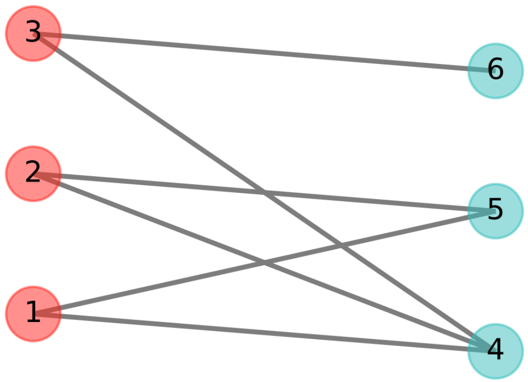
- Disconnected (complete bipartite components)

$$M_2 = \begin{bmatrix} 1 & 2 & \star \\ 3 & 4 & \star \\ \star & \star & 5 \end{bmatrix}$$



- Connected

$$M_3 = \begin{bmatrix} 1 & 2 & \star \\ 3 & 4 & \star \\ 6 & \star & 5 \end{bmatrix}$$



Connectivity Affects Implicit Regularization

- Disconnected

$$M_1 = \begin{bmatrix} 1 & 2 & \star \\ 3 & \star & \star \\ \star & \star & 5 \end{bmatrix}$$

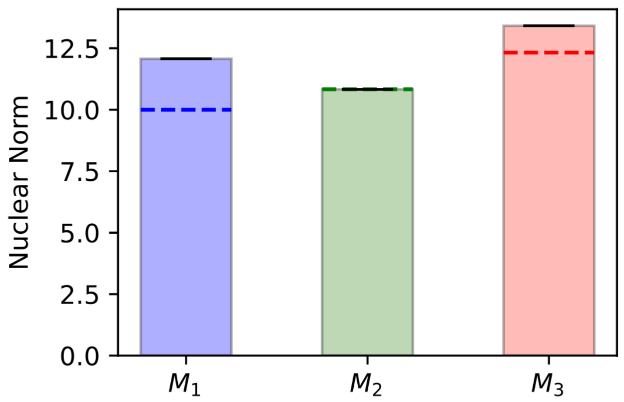
- Disconnected (complete bipartite components)

$$M_2 = \begin{bmatrix} 1 & 2 & \star \\ 3 & 4 & \star \\ \star & \star & 5 \end{bmatrix}$$

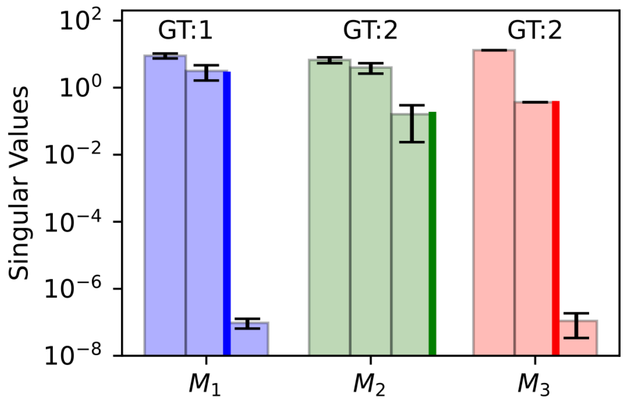
- Connected

$$M_3 = \begin{bmatrix} 1 & 2 & \star \\ 3 & 4 & \star \\ 6 & \star & 5 \end{bmatrix}$$

-



(a) Nuclear norm



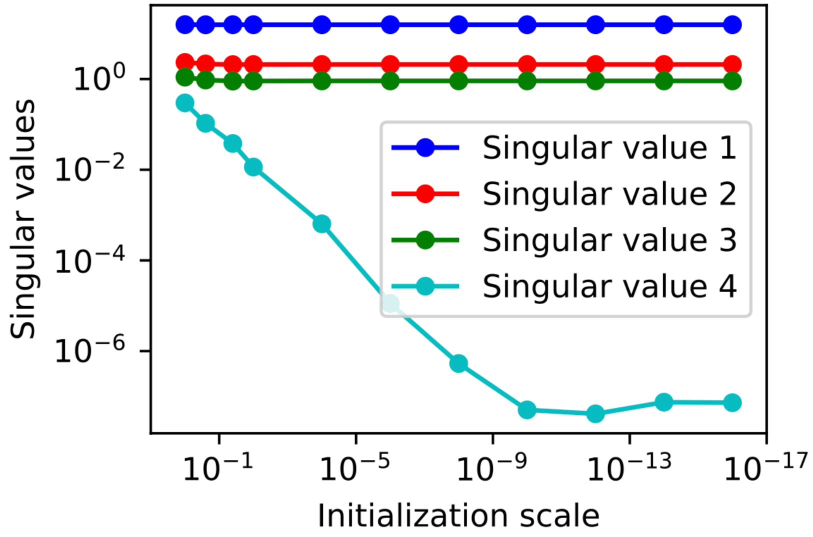
(b) Rank

Connected Case—Initialization Matters

- Matrix Completion

$$\begin{bmatrix} 4 & 0.6 & 1.8 & 0.8 \\ 6 & 0.9 & 2.7 & \star \\ 8 & 2.2 & 2.6 & 1.6 \\ 8 & 2.7 & 5.1 & 3.6 \end{bmatrix}$$

- Matrix Factorization Model $f_\theta = AB$

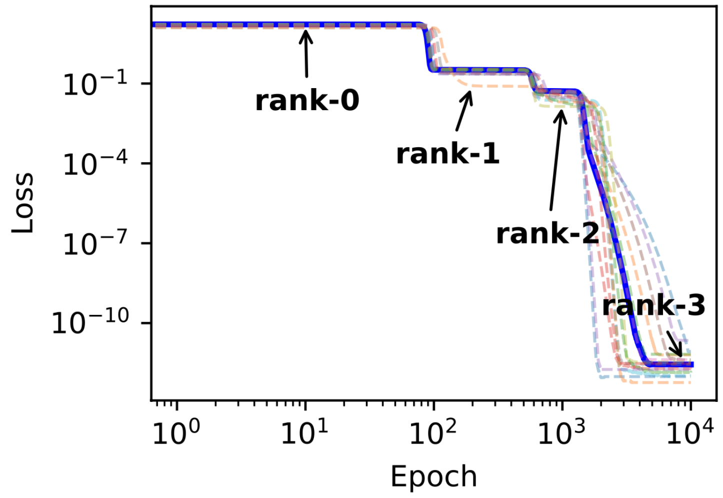


- Large initialization: rank–4
- Small initialization: rank–3

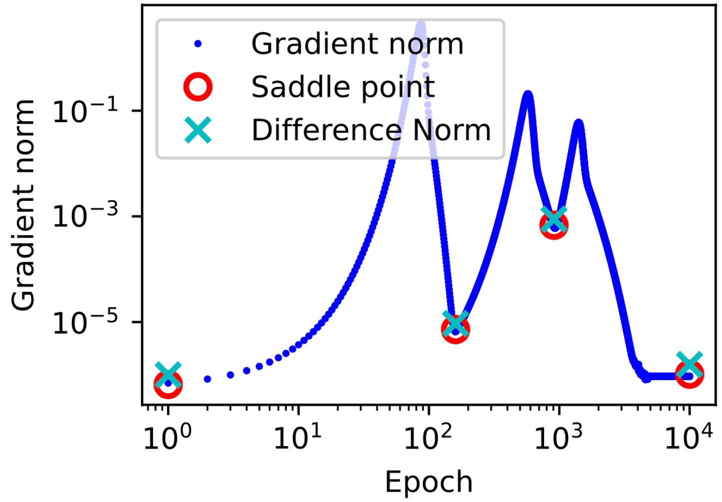
Learning lowest–rank solution in infinitesimal initialization

Connected Case—Traversing Progressive Optima

- Training Loss at Small Initialization



- Gradient Norm during Training

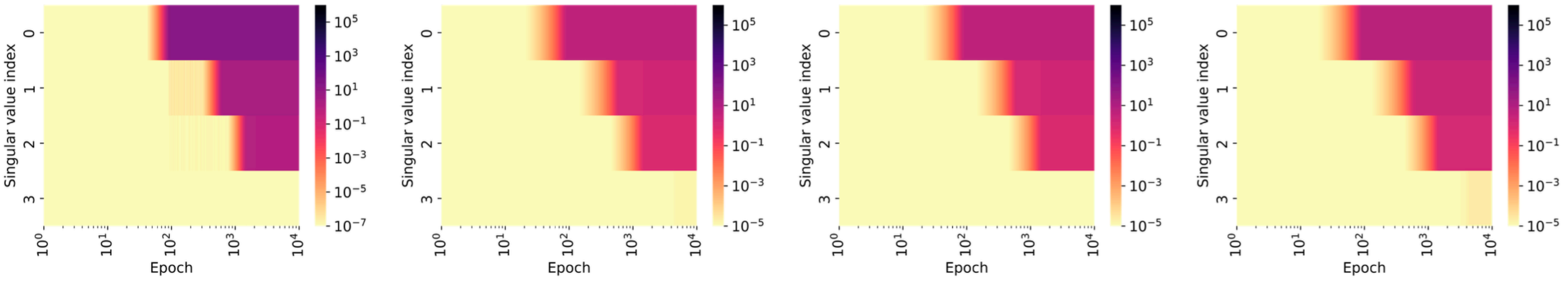


- Training Loss: stepwise decline
- Saddle Points: Experience optimal approximation of each rank

Traversing progressive optima at each rank

Connected Case—Alignment of $\mathrm{row}(\boldsymbol{A})$ and $\mathrm{col}(\boldsymbol{B})$

- Evolution of Singular Values



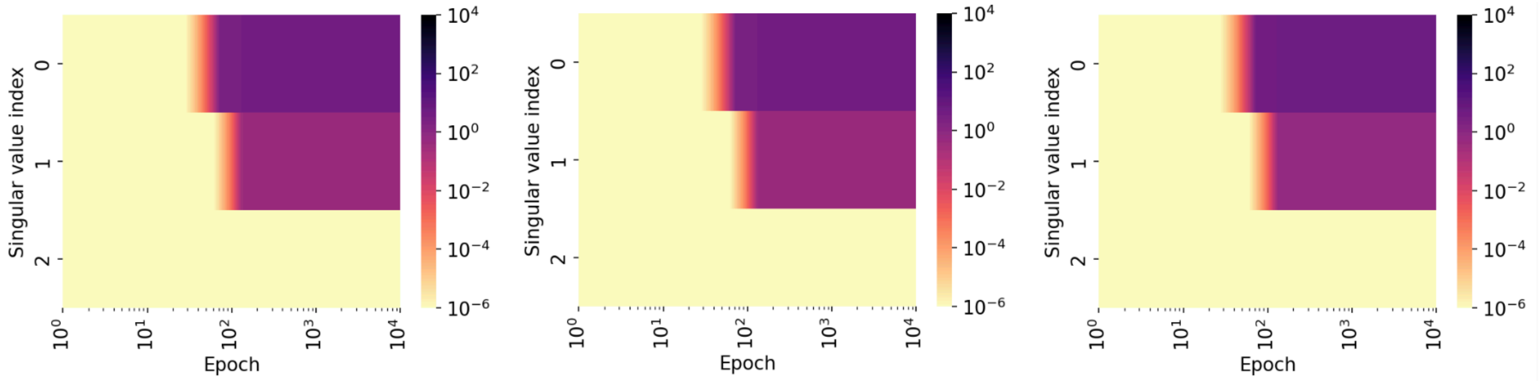
(e) Singular values of $\boldsymbol{W}$ (f) Singular values of $\boldsymbol{A}$ (g) Singular values of $\boldsymbol{B}$ (h) Singular values of $\boldsymbol{W}_{\mathrm{aug}}$

- Rank increases step by step
- $\mathrm{rank}(\boldsymbol{A}) = \mathrm{rank}\left(\boldsymbol{B}^\top\right) = \mathrm{rank}\left(\boldsymbol{W}_{\mathrm{aug}}\right)$, where $\boldsymbol{W}_{\mathrm{aug}} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{B}^\top \end{bmatrix}$
- $\Longrightarrow$ $\mathrm{row}(\boldsymbol{A}) = \mathrm{col}(\boldsymbol{B})$, which induces an **invariant manifold** in theoretical analysis

Disconnected Case—Alignment of $\mathrm{row}(\boldsymbol{A})$ and $\mathrm{col}(\boldsymbol{B})$

- Evolution of Singular Values

$$\begin{bmatrix} 1 & \star & 3 \\ \star & 5 & \star \\ 3 & \star & 9 \end{bmatrix}$$



(a) Matrix completion (b) Singular values of $\boldsymbol{A}$ (c) Singular values of $\boldsymbol{B}$ (d) Singular values of $\boldsymbol{W}_{\mathrm{aug}}$

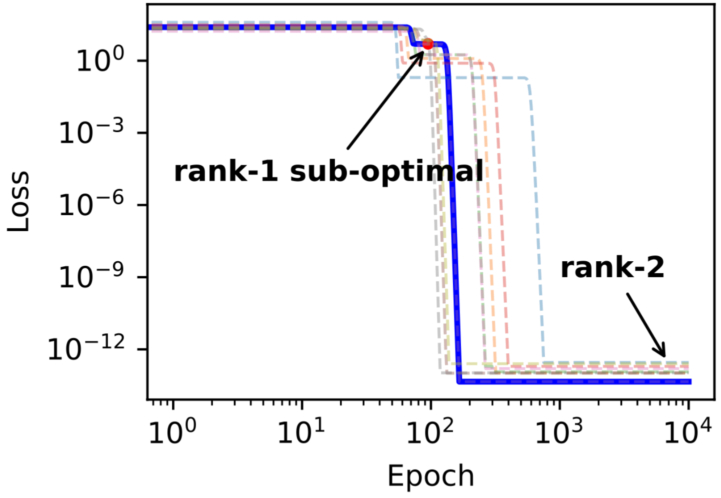
- Alignment of the row space of $\boldsymbol{A}$ and the column space of $\boldsymbol{B}$:

$$\mathrm{row}(\boldsymbol{A}) = \mathrm{col}(\boldsymbol{B})$$

- Lowest–rank solution is not learned $(\mathrm{rank}\text{-}2)$ in disconnected case!
- Lowest nuclear norm solution is learned in this disconnected case

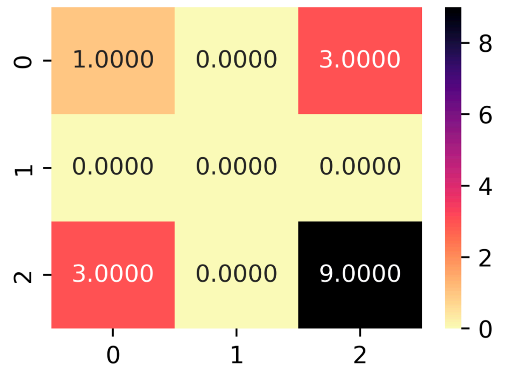
Disconnected Case—Learn Sub-optimal Saddle Point

- Training Loss at Small Initialization



- Rank–1 Sub–optimal

$$\begin{bmatrix} 1 & \star & 3 \\ \star & 5 & \star \\ 3 & \star & 9 \end{bmatrix}$$



- Dynamics: decouple into two independent systems in the disconnected case

$$\begin{cases} \dot{\boldsymbol{a}}_i = -\frac{2}{n}\sum_{j\in I_i}(\boldsymbol{a}_i\cdot\boldsymbol{b}_{\cdot,j}-\boldsymbol{M}_{ij})\boldsymbol{b}_{\cdot,j}^\top, i\in\{1,3\} \\ \dot{\boldsymbol{b}}_{\cdot,j} = -\frac{2}{n}\sum_{i\in I_j}(\boldsymbol{a}_i\cdot\boldsymbol{b}_{\cdot,j}-\boldsymbol{M}_{ij})\boldsymbol{a}_i^\top, j\in\{1,3\} \end{cases}$$

$$\begin{cases} \dot{\boldsymbol{a}}_2 = -\frac{2}{n}(\boldsymbol{a}_2\cdot\boldsymbol{b}_{\cdot,2}-\boldsymbol{M}_{22})\boldsymbol{b}_{\cdot,2}^\top \\ \dot{\boldsymbol{b}}_{\cdot,2} = -\frac{2}{n}(\boldsymbol{a}_2\cdot\boldsymbol{b}_{\cdot,2}-\boldsymbol{M}_{ij})\boldsymbol{a}_2^\top \end{cases}$$

3. Training Dynamics Analysis

Hierarchical Intrinsic Invariant Manifold

Hierarchical Intrinsic Invariant Manifold (HIIM)

Proposition 1 (Hierarchical Intrinsic Invariant Manifold (HIIM)). Let $f_{\theta} = AB$ be a matrix factorization model and $\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_k\}$ be k linearly independent vectors. Define the manifold $\boldsymbol{\Omega}_k$ as

$$\boldsymbol{\Omega}_k := \boldsymbol{\Omega}_k(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_k) = \{\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{B}) \mid \mathrm{row}(\boldsymbol{A}) = \mathrm{col}(\boldsymbol{B}) = \mathrm{span}\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_k\}\}$$

The manifold $\boldsymbol{\Omega}_k$ possesses the following properties:

(1) Invariance under Gradient Flow: Given data S and the gradient flow dynamics $\dot{\boldsymbol{\theta}} = -\nabla R_S(\boldsymbol{\theta})$, if the initial point $\boldsymbol{\theta}_0 \in \boldsymbol{\Omega}_k$, then $\boldsymbol{\theta}(t) \in \boldsymbol{\Omega}_k$ for all $t \geq 0$.

(2) Intrinsic Property: $\boldsymbol{\Omega}_k$ is a data-independent invariant manifold, meaning that for any data S, $\boldsymbol{\Omega}_k$ remains invariant under the gradient flow dynamics.

(3) Hierarchical Structure: The manifolds $\boldsymbol{\Omega}_k$ form a hierarchy: $\boldsymbol{\Omega}_0 \subsetneqq \boldsymbol{\Omega}_1 \subsetneqq \cdots \subsetneqq \boldsymbol{\Omega}_{k-1} \subsetneqq \boldsymbol{\Omega}_k$.

Disconnected Case: Intrinsic Sub-Ω_k Invariant Manifold

Intrinsic Sub–$\boldsymbol{\Omega}_k$ Invariant Manifold

Proposition 2 (Intrinsic Sub–$\boldsymbol{\Omega}_k$ Invariant Manifold). Let $\boldsymbol{f_\theta} = \boldsymbol{AB}$ be a matrix factorization model, $\boldsymbol{M}$ be an incomplete matrix and $\boldsymbol{\Omega}_k$ be an invariant manifold defined in Prop. 1. **If $\boldsymbol{M}$ is disconnected with m connected components, then there exist m sub– $\boldsymbol{\Omega}_k$ manifolds $\boldsymbol{\omega}_k$ such that $\boldsymbol{\omega}_k \subsetneqq \boldsymbol{\Omega}_k$,** each possessing the following properties:

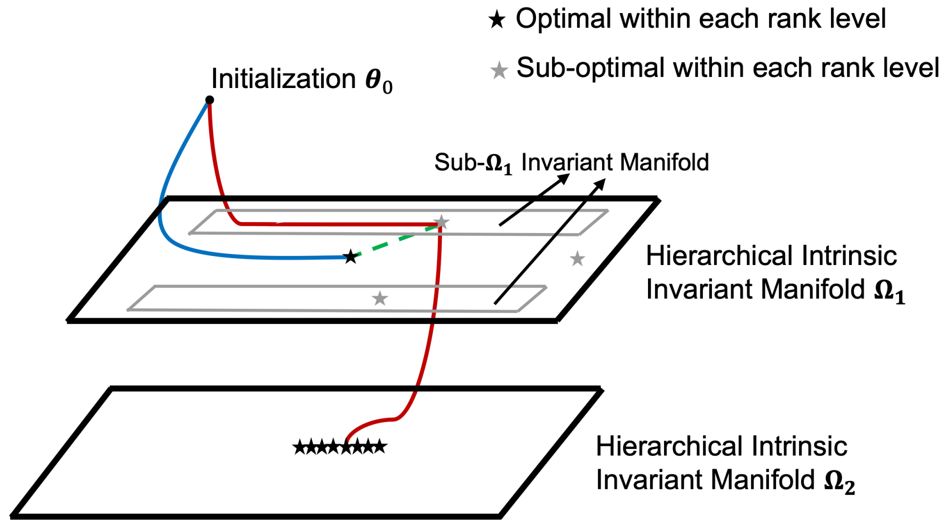
(1) Invariance under Gradient Flow: Given data S and the gradient flow dynamics $\dot{\boldsymbol{\theta}} = -\nabla R_S(\boldsymbol{\theta})$, if the initial point $\boldsymbol{\theta}_0 \in \boldsymbol{\omega}_k$, then $\boldsymbol{\theta}(t) \in \boldsymbol{\omega}_k$ for all $t \geq 0$.

(2) Intrinsic Property: ω_k is a data–value–independent invariant manifold, meaning that for a fixed sampling pattern in $\boldsymbol{M}$ and any observed values $S, \boldsymbol{\omega}_k$ remains invariant under the gradient flow.

(3) Strict Subset Relation: The output set $\{\boldsymbol{f_\theta} \mid \boldsymbol{\theta} \in \boldsymbol{\omega}_k\}$ is a proper subset of $\{\boldsymbol{f_\theta} \mid \boldsymbol{\theta} \in \boldsymbol{\Omega}_k\}$, namely, $\{\boldsymbol{f_\theta} \mid \boldsymbol{\theta} \in \boldsymbol{\omega}_k\} \subsetneq \{\boldsymbol{f_\theta} \mid \boldsymbol{\theta} \in \boldsymbol{\Omega}_k\}$

Intuitive Illustration

- Illustration of Training Trajectories



Blue line represents the trajectory converging to the lowest–rank solution. Red line represents the actual trajectory experienced by the model

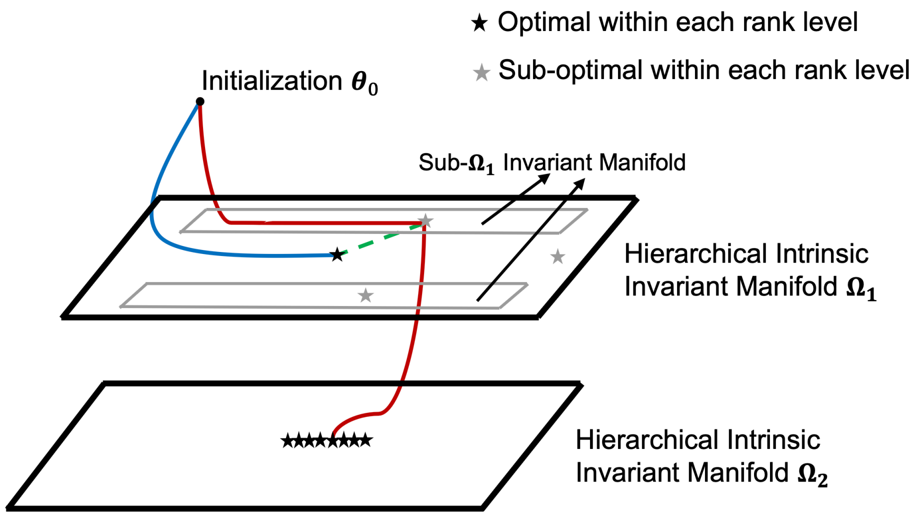
- Connected case: Model traverses with invariant manifold $\boldsymbol{\Omega}_k$
- Disconnected case:
  - Sub–$\boldsymbol{\Omega}_k$ invariant manifold emerges
  - Each sub–$\boldsymbol{\Omega}_k$ induces a sub–optimal saddle point
  - Sub–optima prevent the model from learning the lowest–rank solution

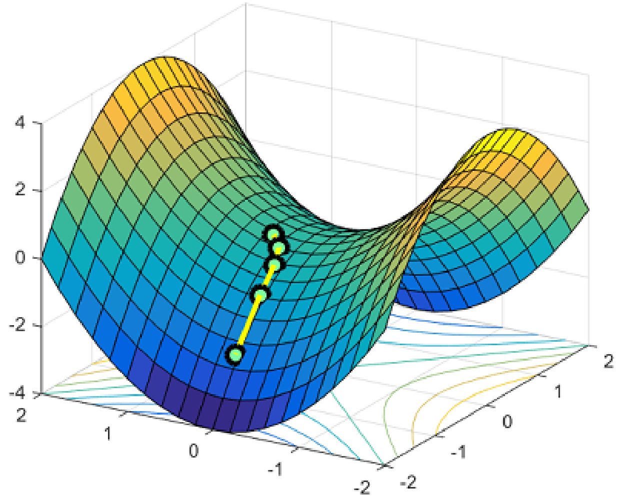
Loss Landscape does not Contain any Local Minima

Loss Landscape

Theorem 1 (Loss Landscape). Given any data S, the critical points of $R_S(\boldsymbol{\theta})$ are either strict saddle points or global minima.

- Gradient descent easily escapes saddle points





Assumptions for Encountered Critical Points

Assumption 1 Top Eigenvalue

Assumption 1 (Top Eigenvalue). Let $\delta \boldsymbol{M} = (\boldsymbol{A}_c \boldsymbol{B}_c - \boldsymbol{M})_{S_{\boldsymbol{x}}}$ be the residual matrix at the critical point $\boldsymbol{\theta}_c = (\boldsymbol{A}_c, \boldsymbol{B}_c)$. Assume that the top singular value of the matrix $\delta \boldsymbol{M}$ is unique.

Assumption 2 Second-order Stationary Point

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

Characterization of Training Dynamics

Transition to the Next Rank–level Invariant Manifold

Theorem 2 (Transition to the Next Rank–level Invariant Manifold). Consider the dynamics $\dot{\boldsymbol{\theta}} = -\nabla R_S(\boldsymbol{\theta})$. Let $\varphi\left(\boldsymbol{\theta}_0, t\right)$ denote the value of $\boldsymbol{\theta}(t)$ when $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$. Let $\boldsymbol{\Omega}$ be an $\boldsymbol{\Omega}_k$ or sub– $\boldsymbol{\Omega}_k$ invariant manifold. Let $\boldsymbol{\theta}_c \in \boldsymbol{\Omega}$ be a critical point satisfying Assump. 1 and 2. Then, for randomly selected $\boldsymbol{\theta}_0$, with probability 1 with respect to $\boldsymbol{\theta}_0$, the limit

$$\tilde{\varphi}\left(\boldsymbol{\theta}_c, t\right) := \lim_{\alpha \to 0} \varphi\left(\boldsymbol{\theta}_c + \alpha \boldsymbol{\theta}_0, t + \frac{1}{\lambda_1} \log \frac{1}{\alpha}\right)$$

exists and falls into an invariant manifold $\boldsymbol{\Omega}_{k+1}$. Here λ_1 is the top eigenvalue of $-\nabla^2 R_S\left(\boldsymbol{\theta}_c\right)$.

Proof Sketch

- **Linear Approximation** near critical point $\boldsymbol{\theta}_c$: $\frac{d\boldsymbol{\theta}}{dt} \approx H(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_c)$.
- **Solution** $\boldsymbol{\theta}(t) = e^{tH}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_c) + \boldsymbol{\theta}_c$, specifically

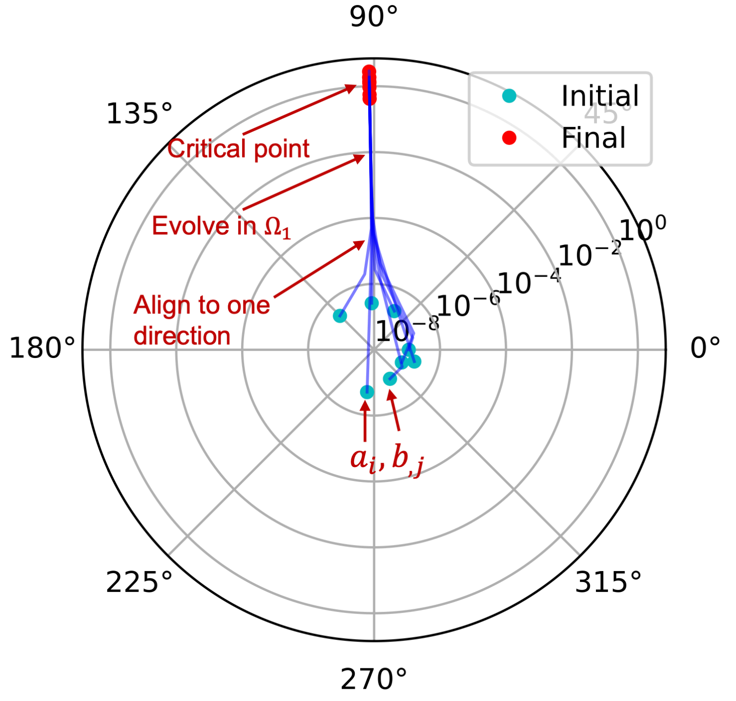
$$\boldsymbol{\theta}(t) = \sum_{i=1}^{s}\sum_{j=1}^{l_i} e^{\lambda_i t}\langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}_c, q_{ij}\rangle q_{ij} + \boldsymbol{\theta}_c$$

- **Dominant eigenvalue trajectory:**

$$\boldsymbol{\theta}(t_0) = \sum_{j=1}^{l_1} e^{\lambda_1 t_0}\langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}_c, q_{1j}\rangle q_{1j} + O(e^{\lambda_2 t_0})$$

- **The first principal component** $\sum_{j=1}^{l_1} e^{\lambda_1 t_0}\langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}_c, q_{1j}\rangle q_{1j}$ corresponds to an Ω_1 invariant manifold under Assump. 1 and 2
- Escaping $\boldsymbol{\theta}_c$ increases rank by 1, **entering** Ω_{k+1}

- **Escape from the top eigen-direction**



Alignment of $\mathrm{row}(\boldsymbol{A})$ and $\mathrm{col}(\boldsymbol{B})$

4. Implicit Regularization Analysis

Minimum Rank Regularization

Minimum Rank Regularization

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

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

- In connected case, experiments provide strong evidence that model achieves an optimal within each invariant manifold $\boldsymbol{\Omega}_k$

Minimum Nuclear Norm Regularization

- In disconnected case, the minimum nuclear norm may still serve as a characterization

Minimum Nuclear Norm Regularization

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

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

5. Discussion and Conclusion

Generalize to Neural Networks: From Linear to Nonlinear

- Matrix Factorization:

$$f_{\boldsymbol{\theta}} = \boldsymbol{A}\boldsymbol{B}$$

- Linear w.r.t input $\boldsymbol{x}$
- Implicit Bias: Low rank

- Neural Networks:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{m} a_i \sigma(\boldsymbol{w}_i^\top \boldsymbol{x})$$

- Non–Linear w.r.t input $\boldsymbol{x}$
- Implicit Bias: ???????

- Model Rank for Non–linear Models:

$$\operatorname{rank}_{f_{\theta}}(\boldsymbol{\theta}^*) := \dim\left(\operatorname{span}\left\{\partial_{\theta_i} f(\cdot; \boldsymbol{\theta}^*)\right\}_{i=1}^{M}\right)$$

- Experiments: Non–linear models has low model rank bias

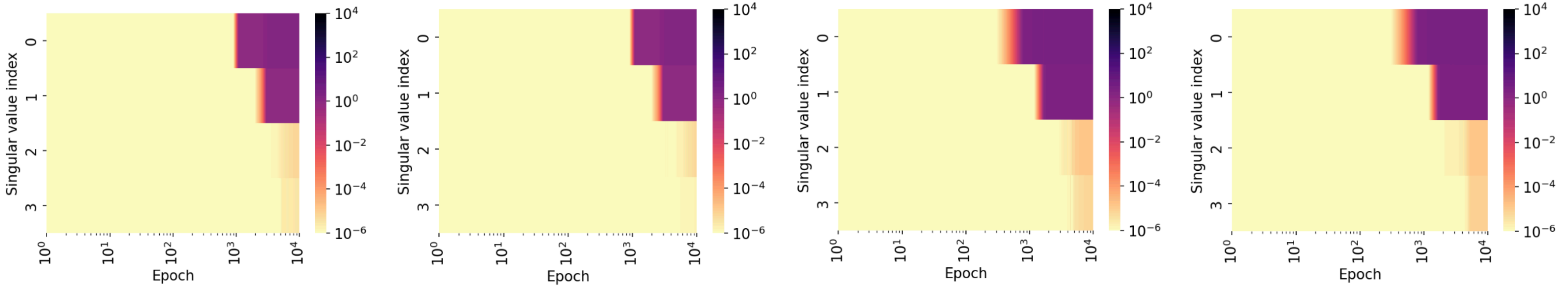
[Zhang et al.] Yaoyu Zhang*, Zhongwang Zhang, Leyang Zhang, *Zhiwei Bai*, Tao Luo, Zhi–Qin John Xu. Optimistic estimate uncovers the potential of nonlinear models. arXiv preprint arXiv: 2307.08921, 2023.

Generalize to Transformer Architecture

- Matrix Factorization Model is a Component of the Transformer Architecture

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

- Low-rank (model rank) behavior in the Attention Module



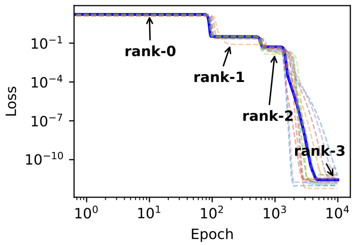
(a) Singular values of $\boldsymbol{W_Q}$ (b) Singular values of $\boldsymbol{W_K}$ (c) Singular values of $\boldsymbol{W_V}$ (d) Singular values of $\boldsymbol{W_O}$

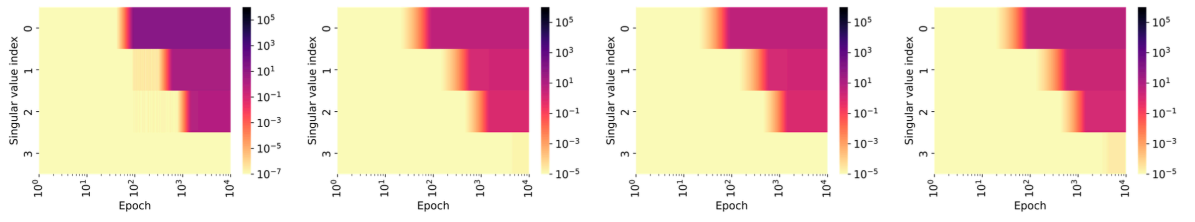
Take Home Messages

- Implicit Regularization of Overparameterized models $\Longrightarrow$ Generalization

- Connected Case: Hierarchical Invariant Manifold Traversal; Model achieves optima within each invariant manifold $\Longrightarrow$ Minimum Rank Regularization

$$\begin{bmatrix} 4 & 0.6 & 1.8 & 0.8 \\ 6 & 0.9 & 2.7 & \star \\ 8 & 2.2 & 2.6 & 1.6 \\ 8 & 2.7 & 5.1 & 3.6 \end{bmatrix}$$

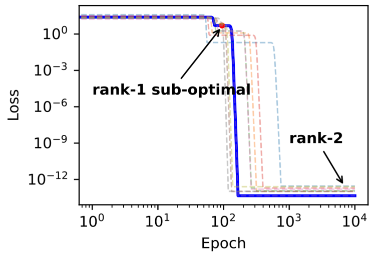


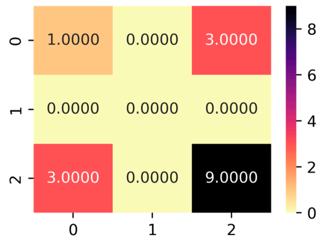


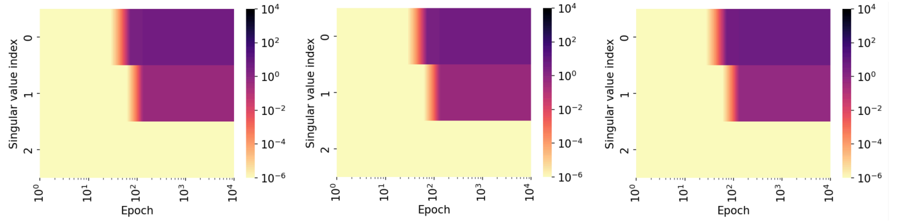
(e) Singular values of $\boldsymbol{W}$ (f) Singular values of $\boldsymbol{A}$ (g) Singular values of $\boldsymbol{B}$ (h) Singular values of $\boldsymbol{W}_{\text{aug}}$

- Disconnected Case: Sub-optima emerges $\Longrightarrow$ preventing low rank; Disconnected with complete bipartite components $\Longrightarrow$ Minimum Nuclear Norm Regularization

$$\begin{bmatrix} 1 & \star & 3 \\ \star & 5 & \star \\ 3 & \star & 9 \end{bmatrix}$$







(b) Singular values of $\boldsymbol{A}$ (c) Singular values of $\boldsymbol{B}$ (d) Singular values of $\boldsymbol{W}_{\text{aug}}$

Thanks!

Mail: bai299@sjtu.edu.cn

WeChat: bai299O_O

Web: https://ZhiweiBai.github.io/