Towards Resolving the Implicit Bias of Gradient Descent for Matrix Factorization: Greedy Low-Rank Learning

Zhiyuan Li, Yuping Luo, **Kaifeng Lyu**

(alphabet order)

PRINCETON UNIVERSITY

清華大學
Tsinghua University

Implicit Bias of GD for Matrix Factorization

Modern neural nets are overparameterized: model complexity ≫ #data
Global minima with bad test errors do exist, but SGD magically avoids them

Need to understand the implicit bias of gradient descent!
(What kind of special properties does the solution found by GD possess?)
However, analyzing deep nets is notoriously difficult…

Work on a simpler model first: Overparameterized Matrix Factorization
[Gunasekar et al., 2017; Li et al., 2018; Arora et al., 2019; Razin and Cohen, 2020; Belabbas, 2020]

(Symmetric) Matrix Factorization

$W_* \in \mathbb{R}^{d \times d}$ is an unknown positive semidefinite (PSD) matrix; W_* is low-rank ($\mathrm{rank}(W_*) \ll d$).
Goal: Recover $W_* \in \mathbb{R}^{d \times d}$ given some observations about W_*.
Method:

1. Formulate a convex empirical risk function $f(W)$ based on the observations;
2. Parameterize W as UU^T, where U is a $d \times r$ matrix ($r \ll d$ is the rank constraint for W);
3. Optimize $f(UU^T)$ over $U \in \mathbb{R}^{d \times r}$.

Example (Matrix Sensing): Each observation $y_i := \langle W_*, X_i \rangle$ is an inner product of the unknown matrix W_* and a measurement matrix $X_i \in \mathbb{R}^{d \times d}$.

$$f(W) = \frac{1}{2} \sum_{i=1}^{m} (\langle W, X_i \rangle - y_i)^2$$

Overparameterized Matrix Factorization

$W_* \in \mathbb{R}^{d \times d}$ is an unknown positive semidefinite (PSD) matrix; W_* is low-rank ($\mathrm{rank}(W_*) \ll d$).
Goal: Recover $W_* \in \mathbb{R}^{d \times d}$ given some observations about W_*.
Method: Optimize $f(UU^T)$ over $U \in \mathbb{R}^{d \times d}$.

No rank constraint!

[Gunasekar et al., 2017]: Empirically, if the initial point is very small (i.e., $U \approx 0$), then GD converges to a solution with low reconstruction error.

How to understand this implicit bias of GD?

Conjecture ([Gunasekar et al., 2017]):
With infinitesimal initialization, gradient flow converges to the minimum nuclear norm solution.

gradient flow: GD with LR $\to 0$.
nuclear norm: a convex relaxation of rank.

Counter-example: Matrix Completion

$$M = \begin{bmatrix} ? & ? & 1 & 10^2 \\ ? & ? & 10^2 & ? \\ 1 & 10^2 & ? & ? \\ 10^2 & ? & ? & ? \end{bmatrix}$$

$$f(W) = \frac{1}{2} \sum_{M_{ij} \neq ?} \left(W_{ij} - M_{ij}\right)^2$$

$$M_{\text{norm}} = \begin{bmatrix} 10^2 & 1 & 1 & 10^2 \\ 1 & 10^2 & 10^2 & 1 \\ 1 & 10^2 & 10^2 & 1 \\ 10^2 & 1 & 1 & 10^2 \end{bmatrix}$$

$\|M_{\text{norm}}\|_* = 4 \times 10^2$
$\text{rank}(M_{\text{norm}}) = 2$

The conjectured solution [Gunasekar et al., 2017]

$$M_{\text{rank}} = \begin{bmatrix} 1 & 10^2 & 1 & 10^2 \\ 10^2 & 10^4 & 10^2 & 10^4 \\ 1 & 10^2 & 1 & 10^2 \\ 10^2 & 10^4 & 10^2 & 10^4 \end{bmatrix}$$

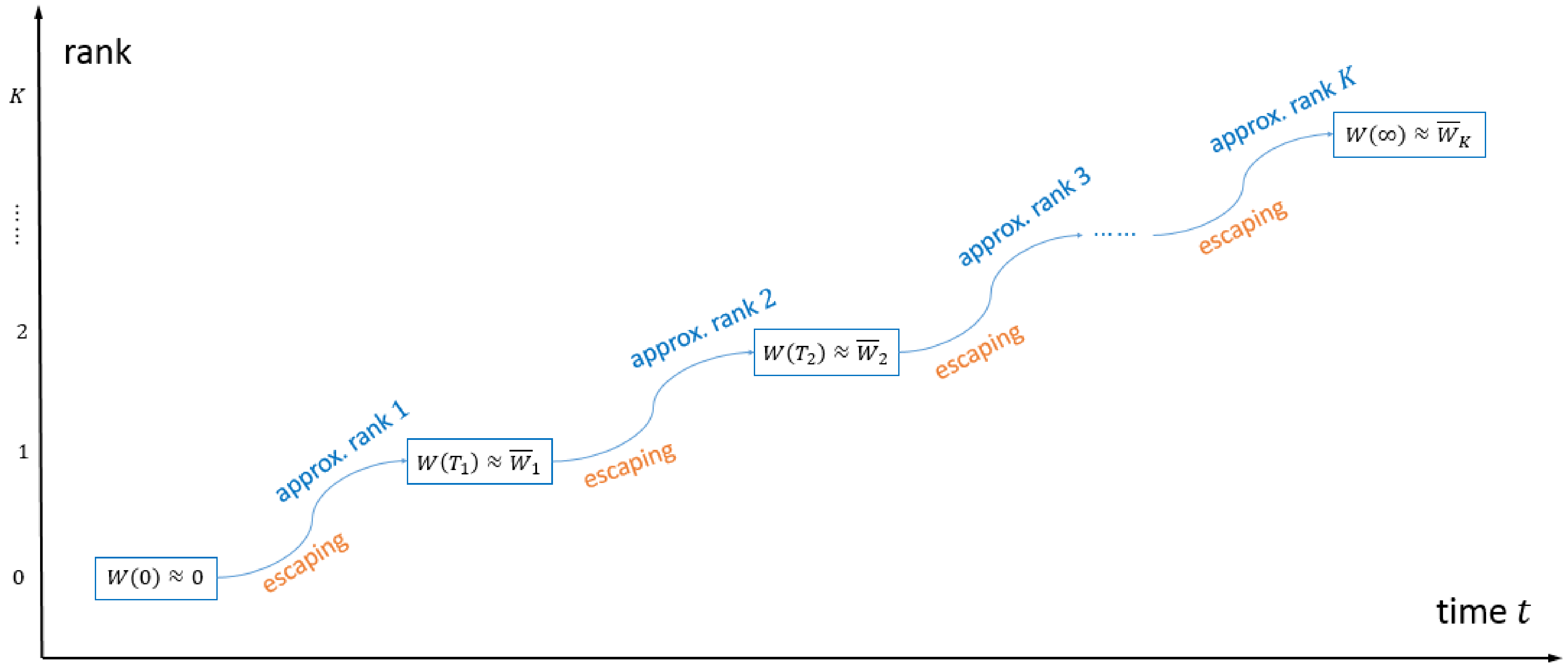
$\|M_{\text{rank}}\|_* = 2 \times 10^4 + 2$
$\text{rank}(M_{\text{rank}}) = 1$

The actual solution found by gradient flow (This work)

Our Result (under certain technical assumptions)
For gradient flow $U(t)$ with very small init, the matrix $W(t) = U(t)U(t)^T$ has rank gradually increasing over time.

For convex $f(W)$, $\exists \overline{W}_1, \overline{W}_2, \ldots, \overline{W}_K$ with rank $1, 2, \ldots, K$ such that $W(t)$ passes through $\overline{W}_1$ to $\overline{W}_K$ in order.

- $\overline{W}_r$ is a local minimizer of $f(W)$ among PSD matrices with rank $\le r$;
- $\overline{W}_K$ is a global minimizer of $f(W)$ among all PSD matrices.



Greedy Low-Rank Learning

Main Results. We provide theoretical evidence that gradient flow with infinitesimal initialization has the same trajectory as that of an algorithm which we call Greedy Low-Rank Learning (GLRL).

GLRL: a greedy algorithm that learns solutions with gradually increasing rank (see our paper)

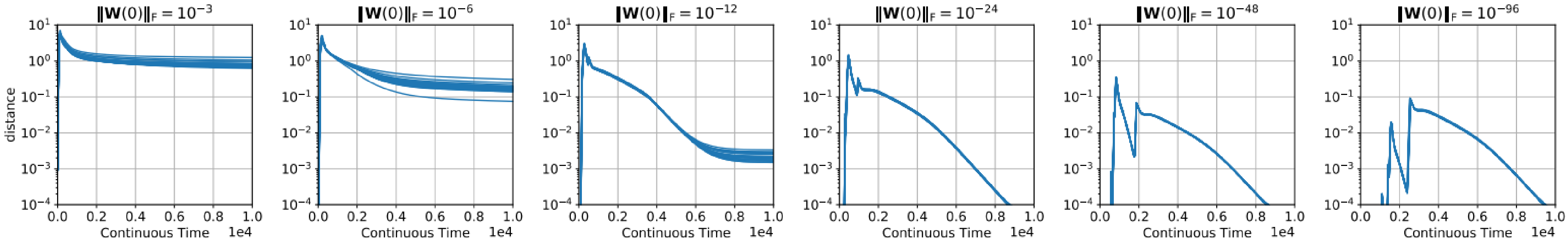


Figure: The pointwise distance between the trajectories of GD and GLRL is getting smaller as init → 0.

Thanks for Listening!

Poster Session 5:
May 4, 9 am - 11 am PDT