On the Convergence and Implicit Bias of Overparametrized Linear Networks

Hancheng Min, Salma Tarmoun, René Vidal and Enrique Mallada

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

2022 MINDS Retreat
Jan. 18th – 21st

Introduction

- In deep learning, neural networks are typically **overparametrized**

$$(\#\ of\ parameters) \quad \gg \quad (\#\ of\ traning\ examples)$$

  - Highly underdetermined problem, many solutions
  - Variants of gradient descent often finds those with good generalization
- Theoretically understand the nonlinear training dynamics of gradient methods
- Prior works suggest that in this overparametrized regime, **specific initialization** may:
  - Accelerate convergence (*implicit acceleration*)
  - Promote generalization (*implicit bias*)
- Question: Are there general properties of **initialization** that benefit **convergence** and **implicit bias**?
- Our setting: <u>two-layer linear networks</u>, <u>gradient flow,</u> the answer is YES!

Outline

- *(Convergence)* **Sufficient imbalance** or **sufficient margin** guarantees exponential convergence

- *(Implicit Bias)* **Orthogonal initialization** leads to min-norm solution

Contributions: Convergence

- Existing analysis for convergence of two-layer linear networks requires **strong assumptions on the initialization (balanced, or spectral)**

	Spectral	Non-spectral (with sufficient **margin**)
Balanced	[Saxes'14] [Gidel'19]	[Arora'18]
Sufficiently **Imbalanced**	[Tarmoun'21]	**Our work**

A Saxe, J Mcclelland, and S Ganguli. "Exact solutions to the nonlinear dynamics of learning in deep linear neural network." ICLR 2014
G Gidel, F Bach, and S Lacoste-Julien. "Implicit regularization of discrete gradient dynamics in linear neural networks." NeurIPS 2019
S Arora, N Cohen, N Golowich, and W Hu. "A convergence analysis of gradient descent for deep linear neural networks." ICLR 2018
S Tarmoun, G França, B D Haeffele, and R Vidal. "Understanding the dynamics of gradient flow in overparameterized linear models." ICML 2021

Contributions: Convergence

- Existing analysis for convergence of two-layer linear networks requires **strong assumptions on the initialization (balanced, or spectral)**

	Spectral	Non-spectral (with sufficient **margin**)
Balanced	[Saxes'14] [Gidel'19]	[Arora'18]
Sufficiently **Imbalanced**	[Tarmoun'21]	**Our work**

- We show

$$\boxed{Rate \geq \sqrt{(Imbalance)^2 + 4(Margin)^2}}$$

- **Exponential convergence** via sufficient imbalance or sufficient margin

Problem Setup

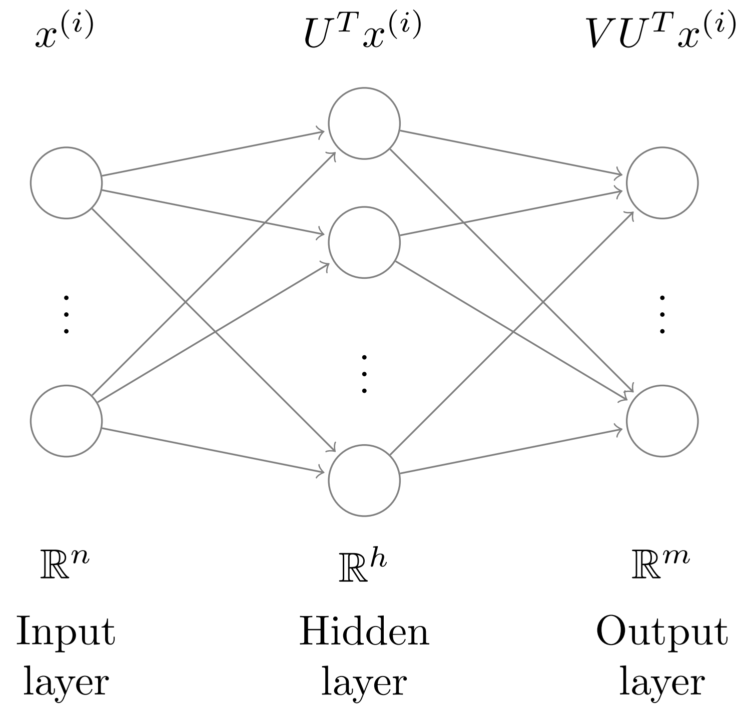
- Training data $X = [x^{(1)} \quad \cdots \quad x^{(P)}]^T \in \mathbb{R}^{P\times n}, Y = [y^{(1)} \quad \cdots \quad y^{(P)}]^T \in \mathbb{R}^{P\times m}$
- Two-layer linear network, squared loss (Regression task)

$$L(U,V) = \frac{1}{2}\|Y - XUV^T\|_F^2, \qquad U \in \mathbb{R}^{n\times h}, V \in \mathbb{R}^{m\times h}$$

- Overparametrized model: $h \geq min\{n, m\}$

- Gradient flow dynamics

$$\dot{U} = -\frac{\partial L}{\partial U}, \quad \dot{V} = -\frac{\partial L}{\partial V}$$



Outline - Convergence

- *(Convergence)* **Sufficient imbalance** or **sufficient margin** guarantees exponential convergence
  - *(Warm-up) Scalar case*: $L_s(u,v) = \frac{1}{2}|y - uv|^2$
  - *Matrix case*: $L(U,V) = \frac{1}{2}\|Y - UV^T\|_F^2$
  - *Convergence results for regression*: $L(U,V) = \frac{1}{2}\|Y - XUV^T\|_F^2$

Scalar Dynamics: Imbalance

- Gradient flow on $L_s(u, v) = \frac{1}{2}|y - uv|^2$

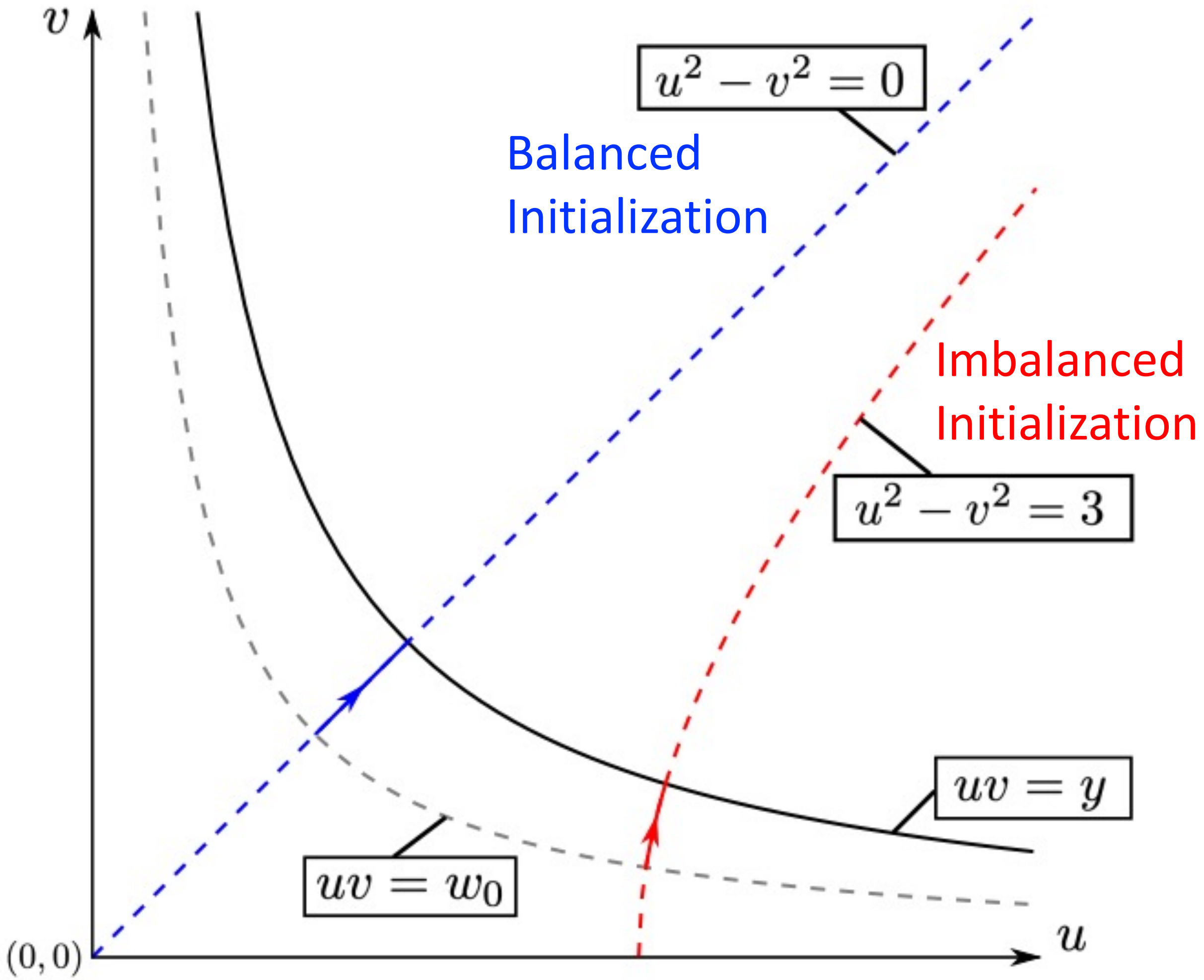
$$\dot{u} = (y - uv)v$$
$$\dot{v} = (y - uv)u$$

- Imbalance of the weights

$$d \coloneqq u^2 - v^2$$

- Imbalance is time-invariant [Saxes'14]

$$\dot{d} = 0$$



A Saxe, J Mcclelland, and S Ganguli. "Exact solutions to the nonlinear dynamics of learning in deep linear neural network." ICLR 2014

Scalar Dynamics: Exponential Convergence

- Gradient flow on $L_s(u,v) = \frac{1}{2}|y - uv|^2$

$$\dot{u} = (y - uv)v, \qquad \dot{v} = (y - uv)u$$

- We need a lower bound on the <u>instantaneous rate</u> $-\dot{L}_s/L_s$

Grönwall's inequality

$$\dot{L}_s(t) \leq -\alpha L_s(t)$$

$$\Rightarrow L_s(t) \leq exp(-\alpha t)\, L_s(0)$$

For **exponential convergence**, show

$-\dot{L}_s/L_s \geq \alpha$ for some $\alpha > 0$

Scalar Dynamics: Exponential Convergence

- Gradient flow on $L_s(u,v) = \frac{1}{2}|y - uv|^2$

  $\dot{u} = (y - uv)v, \qquad \dot{v} = (y - uv)u$

- We need a lower bound on the instantaneous rate $-\dot{L}_s/L_s$
- $-\frac{\dot{L}_s}{L_s} = 2(u^2 + v^2)$
  - d is time-invariant ✓
  - A lower bound on $(uv)^2$??

Express u^2, v^2 by **imbalance** d and **product** uv

$$u^2 = \frac{d + \sqrt{d^2 + 4(uv)^2}}{2}$$

$$v^2 = \frac{-d + \sqrt{d^2 + 4(uv)^2}}{2}$$

Scalar Dynamics: Exponential Convergence

- Gradient flow on $L_s(u, v) = \frac{1}{2}|y - uv|^2$

  $$\dot{u} = (y - uv)v, \qquad \dot{v} = (y - uv)u$$

- We need a lower bound on the instantaneous rate $-\dot{L}_s/L_s$

- $-\frac{\dot{L}_s}{L_s} = 2(u^2 + v^2) = 2\sqrt{d^2 + 4(uv)^2}$
  - d is time-invariant ✓
  - $(uv)^2 \geq (Margin)^2$ ✓

$|uv|$ stays above the **margin**

$$\begin{aligned} |u(t)v(t)| &\geq |y| - |y - u(t)v(t)| \\ &\geq |y| - |y - u(0)v(0)| \\ &:= Margin \end{aligned}$$

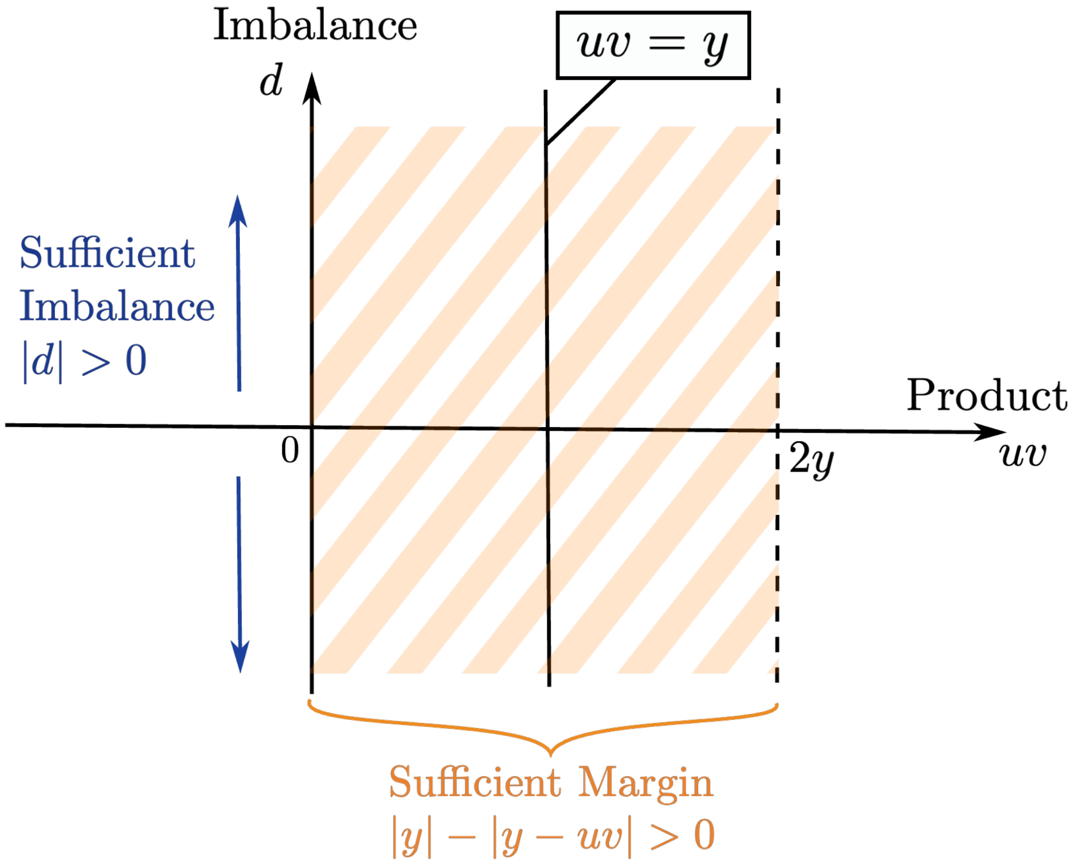
Scalar Dynamics: Exponential Convergence

- Gradient flow on $L_s(u,v) = \frac{1}{2}|y - uv|^2$

$$\dot{u} = (y - uv)v, \qquad \dot{v} = (y - uv)u$$

- We need a lower bound on the <u>instantaneous rate</u> $-\dot{L}_s/L_s$

- $-\frac{\dot{L}_s}{L_s} = 2(u^2 + v^2) = 2\sqrt{d^2 + 4(uv)^2}$
  - d is time-invariant ✓
  - $(uv)^2 \geq (Margin)^2$ ✓



- $-\frac{\dot{L}_s(t)}{L_s(t)} = 2\sqrt{d^2 + 4\big(u(t)v(t)\big)^2} \geq 2\sqrt{d^2 + 4(max\{|y| - |y - u(0)v(0)|, 0\})^2}$

$$\boxed{Rate \geq 2\sqrt{(Imbalance)^2 + 4(Margin)^2}}$$

From Scalar Case to Matrix Case

Scalar Case

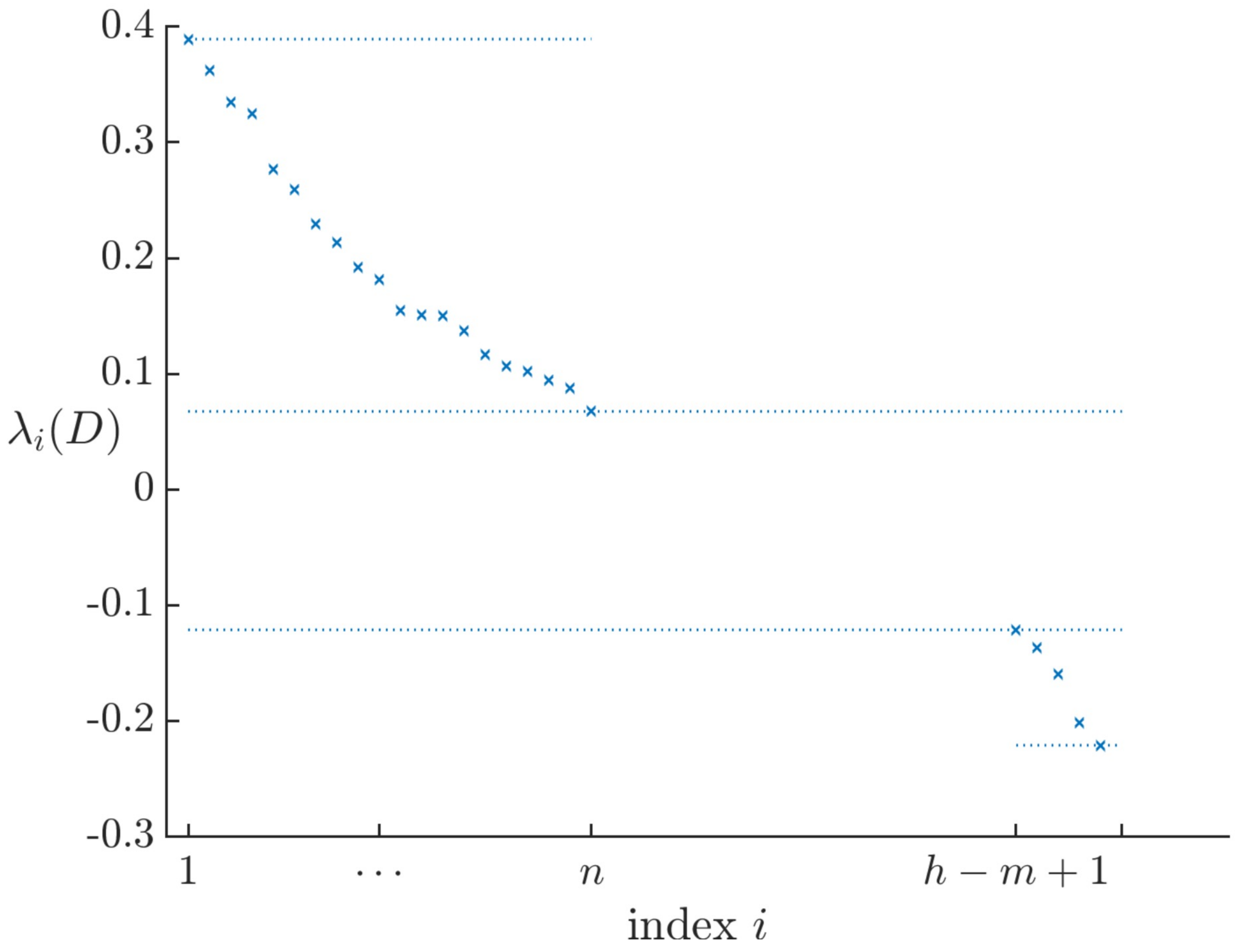
- $L_s(u, v) = \frac{1}{2}|y - uv|^2$
- Imbalance $d = u^2 - v^2$
- Rate depends on imbalance d and product uv
- L_s converges exponentially via
  - **Sufficient imbalance**
  - **Sufficient margin**

Matrix Case

- $L(U, V) = \frac{1}{2}\|Y - UV^T\|_F^2$
  $(U \in \mathbb{R}^{n \times h}, V \in \mathbb{R}^{h \times m}, h \geq min\{n, m\})$
- Imbalance $D = U^T U - V^T V$
- Rate depends on imbalance quantities $\underline{\Delta}, \Delta_+, \Delta_-$ and product UV^T
- L converges exponentially via
  - **Sufficient level of imbalance $\underline{\Delta}$**
  - **Sufficient margin**

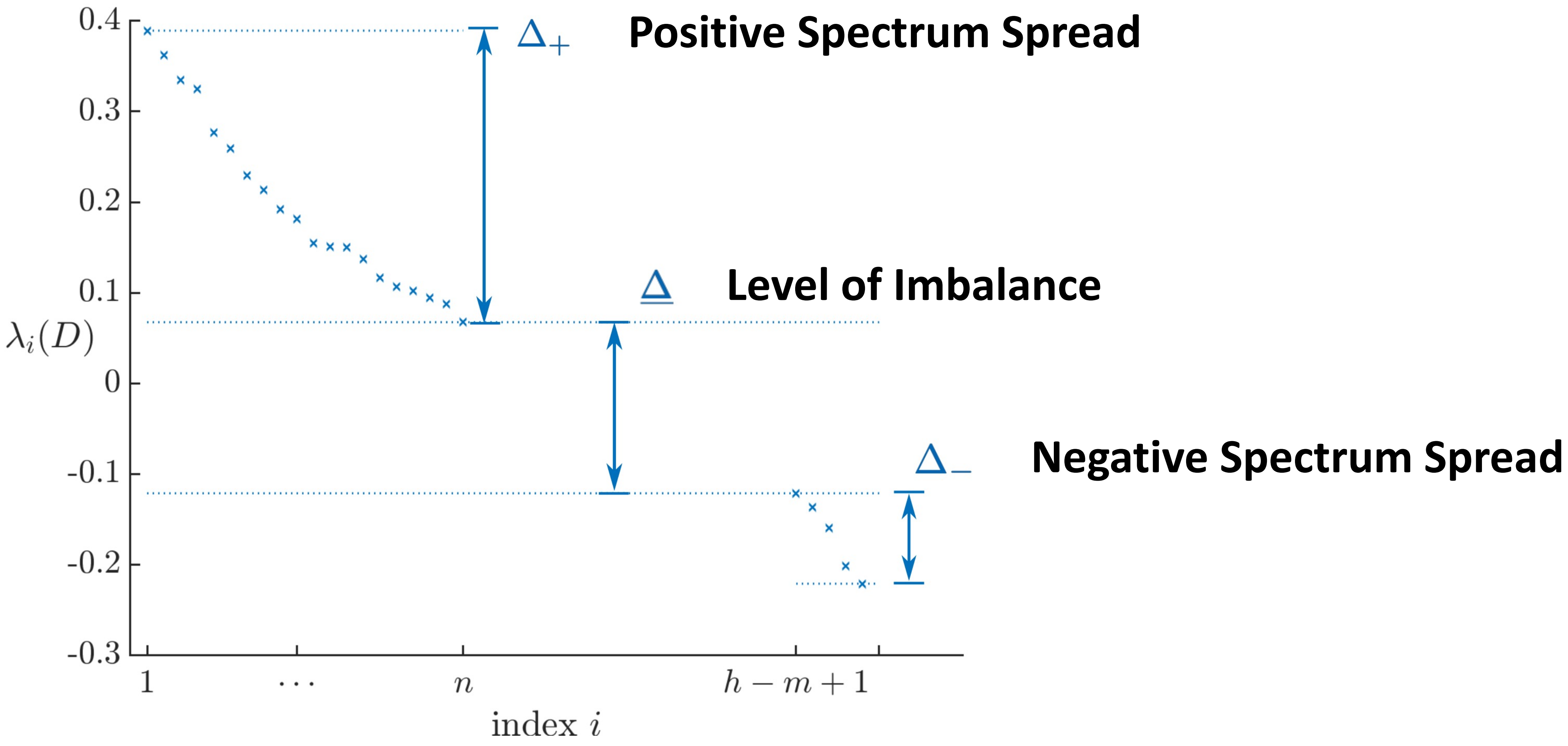
Imbalance quantities

- $L(U,V) = \frac{1}{2}\|Y - UV^T\|_F^2 \ (U \in \mathbb{R}^{n\times h}, V \in \mathbb{R}^{h\times m})$
- Imbalance $D = U^T U - V^T V$



Imbalance quantities

- $L(U,V) = \frac{1}{2}\|Y - UV^T\|_F^2 \; (U \in \mathbb{R}^{n\times h}, V \in \mathbb{R}^{h\times m})$
- Imbalance $D = U^T U - V^T V$



Main Results: Instantaneous Rate

- For the scalar case

$$Rate = 2\sqrt{(Imbalance)^2 + 4(Product)^2}$$

- For the matrix case

$$Rate \geq -Spread + \sqrt{(lvl.of\ imbalance + Spread)^2 + 4\sigma^2(Product)}$$

***Proposition** 1. (Lower bound on instantaneous rate) Define* $D = U^T U - V^T V$. *Consider the gradient flow on* $L(U,V) = \frac{1}{2}\|Y - UV^T\|_F^2$. *Then we have*

$$-\frac{\dot{L}}{L} \geq -\Delta_+ + \sqrt{\left(\Delta_+ + \underline{\Delta}\right)^2 + 4\sigma_m^2(UV^T)} - \Delta_- + \sqrt{\left(\Delta_- + \underline{\Delta}\right)^2 + 4\sigma_n^2(UV^T)},$$

Main Results: Exponential Convergence

- We have a lower bound on the instantaneous rate

$$Rate \geq -Spread + \sqrt{(lvl.\,of\ imbalance + Spread)^2 + 4\sigma^2(Product)}$$

  - Imbalance is time-invariant ✓
  - $\sigma^2(Product) \geq (Margin)^2$ ✓

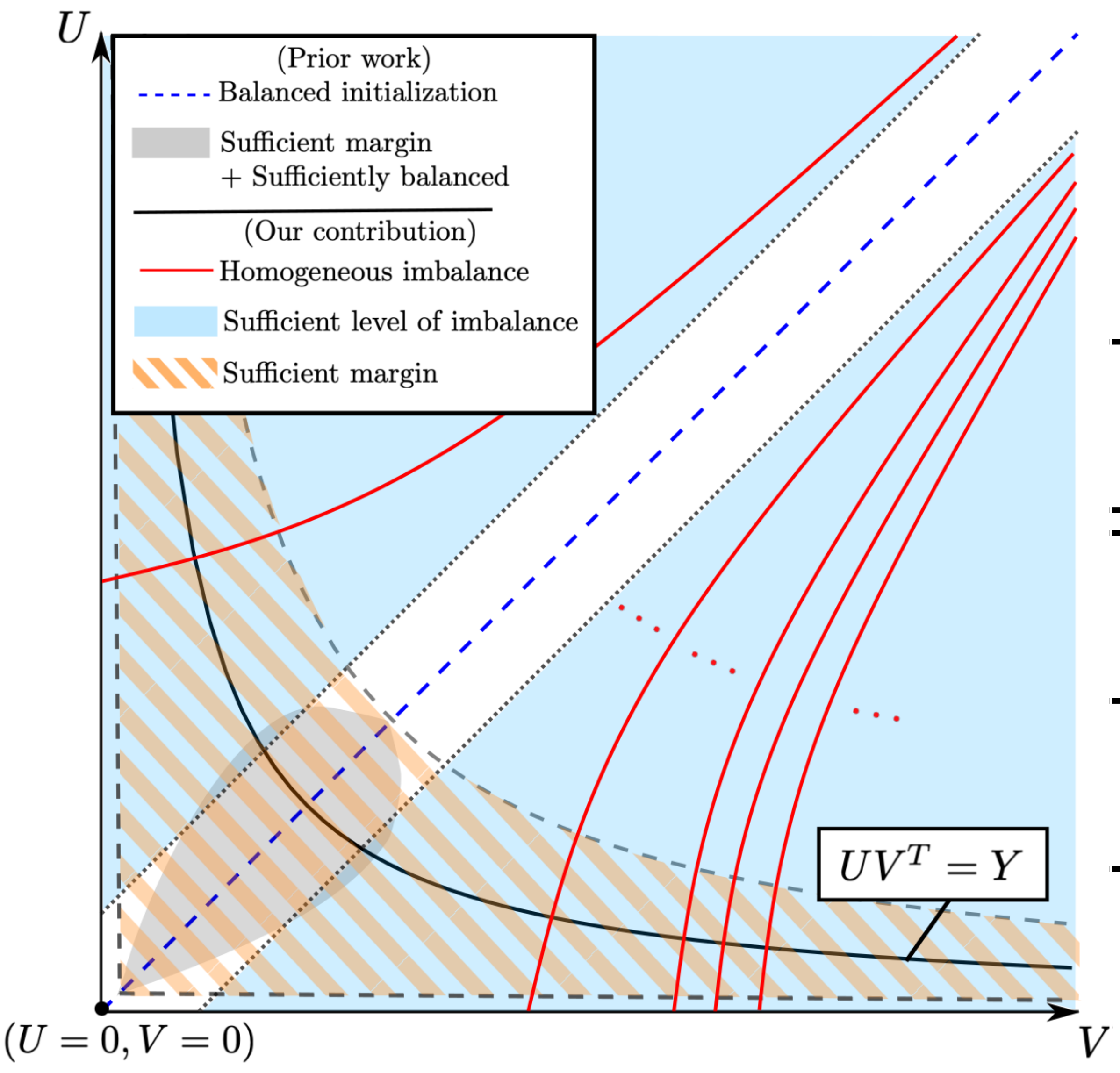
- L converges **exponentially** via
  - **Sufficient level of imbalance**

$$\underline{\Delta} > 0$$

  - **Sufficient margin**

$$\sigma_{min}(Y) - \|Y - UV^T\|_F > 0$$

Exponential Convergence Guarantees: Summary



Non-spectral initializations for the gradient flow on $\frac{1}{2}\|Y - UV^T\|_F^2$

Balanced initialization	$D := U^TU - V^TV = 0$
Margin + approx. balanced [Arora'18]	$\sigma_{min}(Y) - \|Y - UV^T\|_F > \delta$ $\|D\|_F \leq C\delta^2$
Homogeneous imbalance [Tarmoun'21]	$D = \lambda_0 I_h, \|\lambda_0\| > 0$
Sufficient level of imbalance [Min'21]	$\underline{\Delta} > 0$
Sufficient margin	$\sigma_{min}(Y) - \|Y - UV^T\|_F > 0$

Convergence Result for Linear Regression

- For matrix factorization $L(U,V) = \frac{1}{2}\|Y - UV^T\|_F^2$, we have

$$Rate \geq \sqrt{(Imbalance)^2 + 4(Margin)^2}$$

- For linear regression $\tilde{L}(U,V) = \frac{1}{2}\|Y - XUV^T\|_F^2$, we have ($\Sigma_x = X^T X$)

$$Rate \geq \lambda_{min}(\Sigma_x)\sqrt{(Imbalance)^2 + 4(Margin)^2/\lambda_{max}(\Sigma_x)}$$

Outline

- *(Convergence)* **Sufficient imbalance** or **sufficient margin** guarantees exponential convergence

- *(Implicit Bias)* **Orthogonal initialization** leads to min-norm solution

Implicit Bias to Min-norm Solution

- <u>Suppose $X \in \mathbb{R}^{P\times n}$ **DOES NOT** have full row rank, $r = rank(X) < n$</u>

- (Underdetermined) Infinitely many solutions to $min_\Theta \|Y - X\Theta\|_F$

- The **minimum-norm solution**

$$\widehat{\Theta} = argmin_\Theta\{\|\Theta\|_F \colon \|Y - X\Theta\|_F = min_\Theta \|Y - X\Theta\|_F\}$$

- We decompose the weight U using the SVD of X

$$U = \Phi_1 \overbrace{\Phi_1^T U}^{:=U_1} + \Phi_2 \overbrace{\Phi_2^T U}^{:=U_2}, \qquad X = W\begin{bmatrix}\Sigma_x^{1/2} & 0\end{bmatrix}\begin{bmatrix}\Phi_1^T \\ \Phi_2^T\end{bmatrix}$$

- "orthogonality" among $U_1, U_2, V \implies$ exact minimum-norm solution

Main Results: Implicit Bias to Min-norm Solution

***Proposition** 2. (**Orthogonal Initialization**) If one have*

$$V(0)U_2^T(0) = 0, \qquad U_1(0)U_2^T(0) = 0,$$

and that the loss converges to a global minimum, then $U(t)V^T(t)$ converges to exactly the minimum-norm solution $\widehat{\Theta}$

- Orthogonal initialization may not converge (e.g., zero initialization).
- Sufficient imbalance or margin can provide convergence guarantee.

Random Initialization + Large Width

Random initialization

$$[U(0)]_{ij}, [V(0)]_{ij} \sim \mathcal{N}(0, h^{-1})$$

Large hidden layer width h

<u>exact</u> minimum-norm solution

- *(Sufficient level of imbalance)*

$$\underline{\Delta}(0) > 0$$

- *(Orthogonality)*

$$\left\| \begin{bmatrix} V(0)U_2^T(0) \\ U_1(0)U_2^T(0) \end{bmatrix} \right\|_F = 0$$

Non-kernel-regime conditions

<u>approximate</u> minimum-norm solution

- *(Sufficient level of imbalance) w.h.p.*

$$\underline{\Delta}(0) > 1$$

- *(Approximate Orthogonality) w.h.p.*

$$\left\| \begin{bmatrix} V(0)U_2^T(0) \\ U_1(0)U_2^T(0) \end{bmatrix} \right\|_F = \mathcal{O}\left(h^{-1/2}\right)$$

Initialization in the kernel regime

Conclusion

We study the gradient flow on two-layer linear networks:

- **Sufficient imbalance** or **sufficient margin** guarantees exponential convergence
- **Orthogonal Initialization** leads to min-norm solution

Future work:

- Convergence analysis extends to other losses
- Deep linear networks
- Imbalance in nonlinear networks (ReLU net, etc.)

Thank you!

Reference

H Min, S Tarmoun, R Vidal, and E Mallada. “On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks.” ICML 2021.

A Saxe, J Mcclelland, and S Ganguli. “Exact solutions to the nonlinear dynamics of learning in deep linear neural network.” ICLR 2014

G Gidel, F Bach, and S Lacoste-Julien. “Implicit regularization of discrete gradient dynamics in linear neural networks.” NeurIPS 2019

S Arora, N Cohen, N Golowich, and W Hu. “A convergence analysis of gradient descent for deep linear neural networks.” ICLR 2018

S Arora, N Cohen, and E Hazan. “On the optimization of deep networks: Implicit acceleration by overparameterization.” ICML 2018

S Tarmoun, G França, B D Haeffele, and R Vidal. “Understanding the dynamics of gradient flow in overparameterized linear models.” ICML 2021

S Du and W Hu. “Width provably matters in optimization for deep linear neural networks”. ICML 2019

Z Li, Y Luo, and K Lyu. “Towards Resolving the Implicit Bias of Gradient Descent for Matrix Factorization: Greedy Low-Rank Learning.” ICLR 2021

S Arora, S Du, W Hu, Z Li, R Salakhutdinov, and R Wang. “On exact computation with an infinitely wide neural net.” NeurIPS 2019