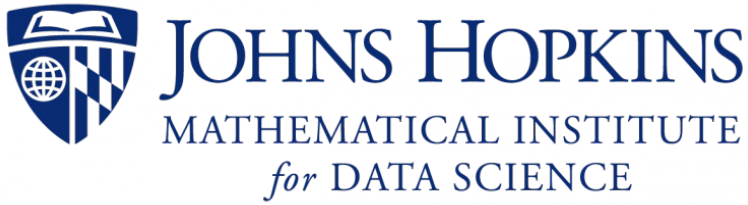
On the Convergence of Gradient Flow on Multi-layer Linear Models

Hancheng Min



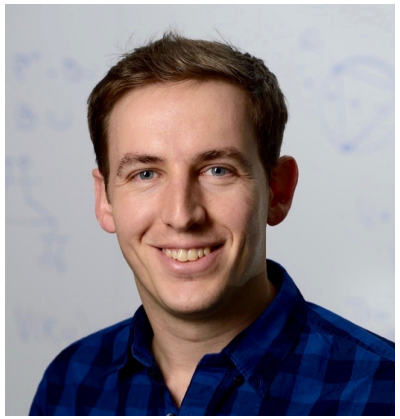
Vision Lab Retreat
September 2nd

Acknowledgements

Salma Tarmoun

Ziqing Xu

René Vidal



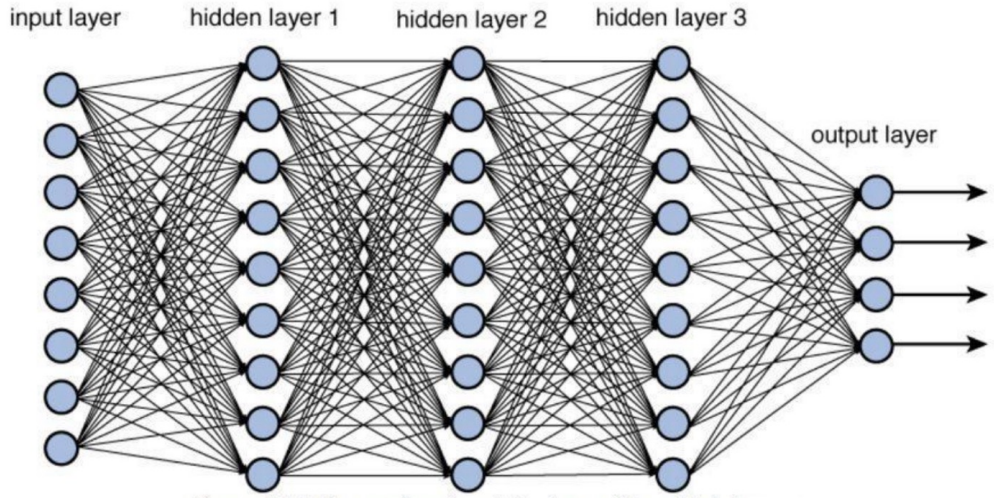
Enrique Mallada

NSF

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

Introduction

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



# of learned weights $\gg$ # of training examples

  - Highly underdetermined problem, many solutions
  - Variants of gradient descent often find those with good generalization

- Question: What is the effect of **overparameterization** on the learning dynamics of optimization algorithms?

Introduction

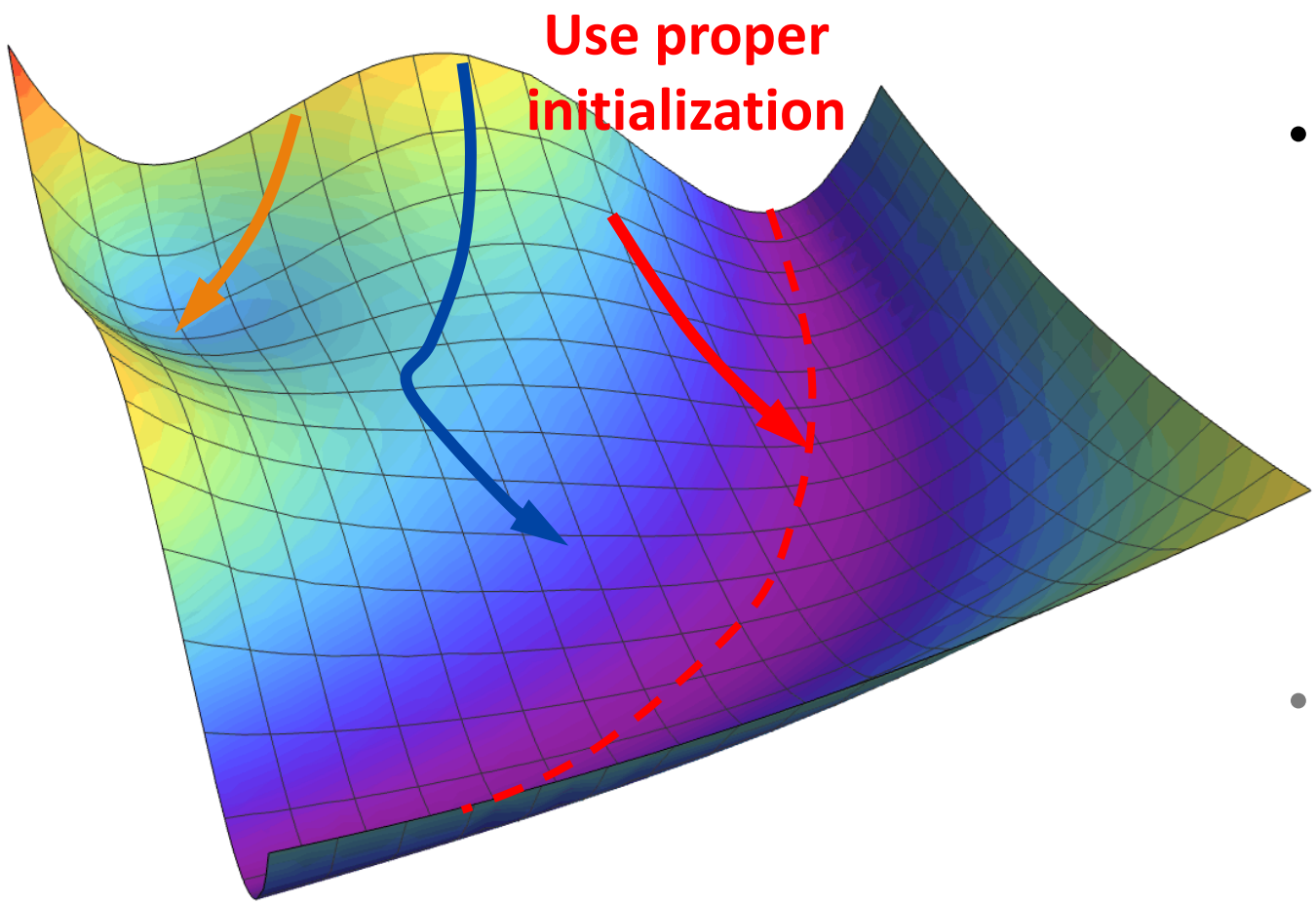
- Prior work suggests that in this overparametrized regime, **specific initialization** may:
  - Accelerate convergence (*implicit acceleration*)
  - Promote generalization (*implicit bias*)

- Question: Are there general properties of **initialization** that benefit **convergence** (This talk) and implicit bias?

- For overparametrized linear models, $\mathcal{L}(W_1, \cdots, W_L) := f(W_1 W_2 \cdots W_L)$
  gradient flow, $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$
  or gradient descent, $W_l^{k+1} = W_l^k - \eta \, \partial \mathcal{L} / \partial W_l$

  the answer is YES!

Non-convex Optimization Landscape



- Loss function for neural network is generally non-convex
- The gradient flow/descent
  - may get stuck at local minimum (non-optimality)
  - may take long time to escape some saddle point (slow convergence)
- Infinitely many global optimal solutions, how can GF/GD reach one that generalizes well? (implicit bias) [Min'21]

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.

Existing Analyses for Specific Initialization

- NTK Initialization [Jacot'18]: Large hidden layer width, random initialization
  - Exponential convergence for GF
  - "lazy regime": rarely seen in practical networks [Chizat'19]

- Small initialization [Stöger'21]: All weight parameters are initialized close to zero
  - Interesting studies on implicit bias: low-rank, sparse models
  - Slow convergence (initialized close to origin, a stationary point)

[Li'21]: $init.scale$: α, $\#\ of\ iter.required$: $\mathcal{O}\left(\frac{1}{\alpha^{(L-2)}}\right)$

- **Restrictive assumptions on initialization**

A Jacot, F Gabriel, and C Hongler. Neural tangent kernel: Convergence and generalization in neural networks. NeurIPS 2018
L Chizat, E Oyallon, and F Bach. On lazy training in differentiable programming. NeurIPS 2019.
D Stöger and M Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. NeurIPS 2021.
J Li, T V Nguyen, C Hegde, and R K. W. Wong. Implicit sparse regularization: The impact of depth and early stopping. NeurIPS 2021.

Contribution

- Non-NTK, non-small initialization is mostly studied for linear networks
- Existing analyses for convergence under gradient flow $\dot{\theta} = -\nabla\mathcal{L}(\theta)$ require **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] [Yun'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
C Yun, S Krishnan, and H Mobahi. A unifying view on implicit bias itraining linear neural networks. ICLR2020

Contribution

- Non-NTK, non-small initialization is mostly studied for linear networks
- Existing analyses for convergence under gradient flow $\dot{\theta} = -\nabla\mathcal{L}(\theta)$ require **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] [Yun'21]	**Our work**

- We show

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

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

Outline

- Problem Setting
- Warm-up Example
- Convergence Analysis for Multi-layer Linear Model
- Convergence Rate Bound
- Conclusion

Problem Setting

- **Problem**: Find solution that obtains

$$f^* = \min_{W \in \mathbb{R}^{n \times m}} f(W)$$

- **Assumptions**: Objective f is μ-strongly convex, and K-smooth

- **Overparametrization**: Multi-layer linear model:

$$\min_{W_1, \cdots, W_L} \mathcal{L}(W_1, \cdots, W_L) := f(W_1 W_2 \cdots W_L)$$

- Examples:
  - Asymmetric matrix factorization: $f(W) = \|Y - W\|_F^2/2, \;\; W = W_1 W_2$
  - Multi-layer linear networks: $f(W) = \|Y - XW\|_F^2/2, \quad W = W_1 W_2 \cdots W_L$

Problem Setting

- **Problem**: Find solution that obtains

$$f^* = \min_{W \in \mathbb{R}^{n \times m}} f(W)$$

- **Assumptions**: Objective f is μ-strongly convex, and has K-Lipschiz gradient

- **Overparametrization**: Multi-layer linear model:

$$\min_{W_1, \cdots, W_L} \mathcal{L}(W_1, \cdots, W_L) := f(W_1 W_2 \cdots W_L)$$

- $\mathcal{L}(W_1, \cdots, W_L)$ is non-convex, and its gradient is not globally Lipschiz. How can gradient flow or gradient descent find the global optimal f^*?

- I will mainly discuss gradient flow $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$ in this talk

Problem Setting: Overparametrized Linear Model

- Multi-layer linear model(network):

  $\mathcal{L}(W_1, \cdots, W_L) := f(W_1 W_2 \cdots W_L)$

- Overparametrized:

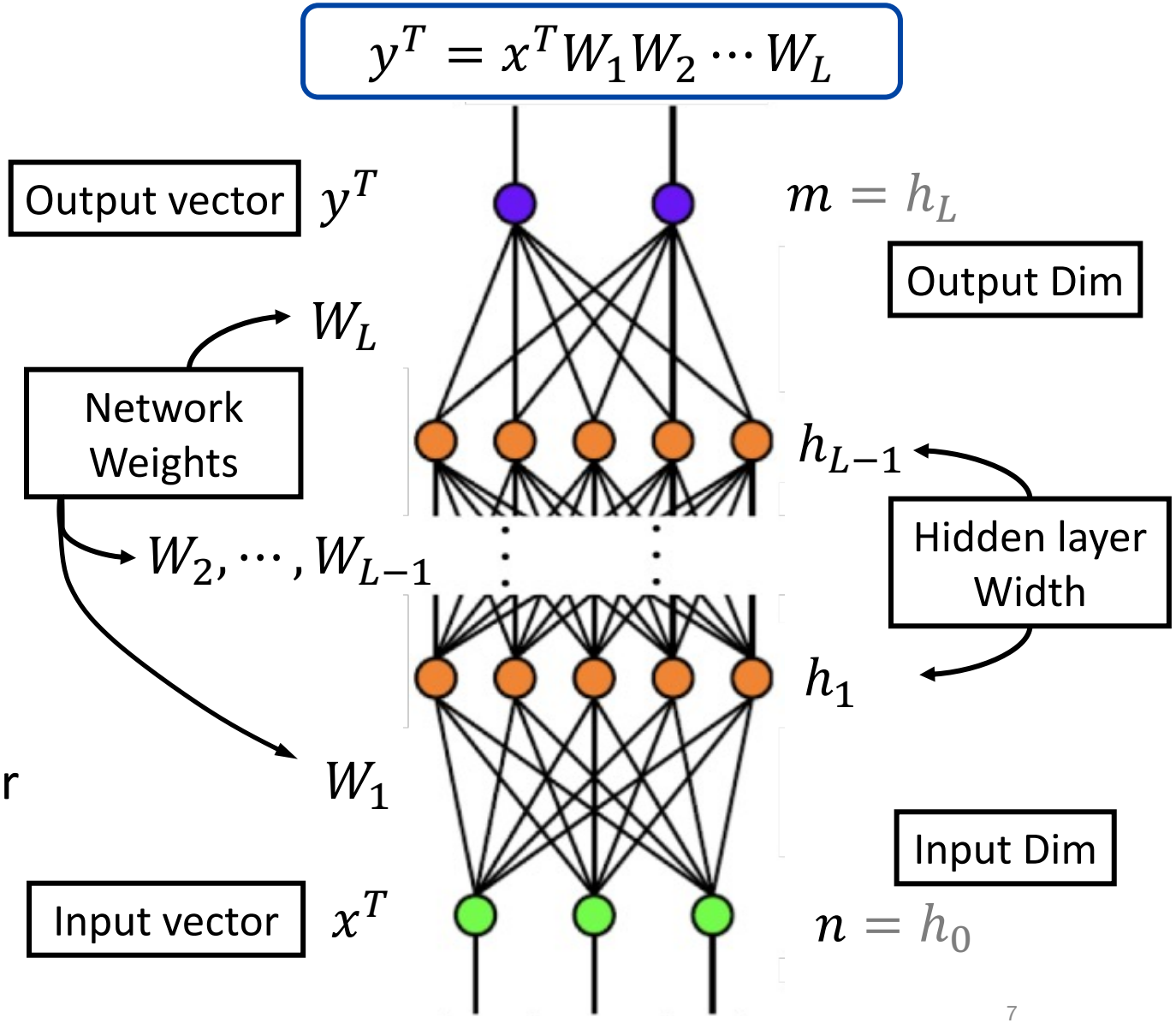
  $W_l \in \mathbb{R}^{h_{l-1} \times h_l}, \quad l = 1, \cdots, L$

  $h_0 = n, h_L = m$

  $\min\{h_1, \cdots, h_{L-1}\} \geq \min\{n, m\}$

  $\Longrightarrow \quad (\mathcal{L}^* = f^*)$

- A deep linear network is FAR simpler than practical neural networks, yet not fully understood.



Problem Setting: Assumptions

- Find solution that obtains

$$f^* = \min_{W \in \mathbb{R}^{n \times m}} f(W)$$

- **Assumptions**: Objective f is μ-strongly convex, and K-smooth

  - satisfies Polyak-Łojasiewicz(PL)-inequality:

$$\|\nabla f(W)\|_F^2 \geq \gamma(f(W) - f^*), \qquad \forall W$$

  - is μ-strongly convex, and K-smooth (Non-essential for convergence of GF)

Convergence with PL-inequality

Non-overparametrized

- Gradient Flow: $\dot{W} = -\nabla f(W)$
- Global PL-Inequality
  $\|\nabla f(W)\|_F^2 \geq \gamma(f(W) - f^*)$

- $\dot{f}(W) = \langle \nabla f(W), \quad \dot{W} \rangle_F = -\|\nabla f(W)\|_F^2 \leq -\gamma(f(W) - f^*)$

Convergence with PL-inequality

Non-overparametrized

- Gradient Flow: $\dot{W} = -\nabla f(W)$

Grönwall's inequality

$$\dot{x}(t) \leq -\gamma x(t)$$
$$\Rightarrow x(t) \leq \exp(-\gamma t)\, x(0)$$

- Global PL-Inequality
  $\|\nabla f(W)\|_F^2 \geq \gamma(f(W) - f^*)$

- $\dot{f}(W) = \langle \nabla f(W), \dot{W} \rangle_F = -\|\nabla f(W)\|_F^2 \leq -\gamma(f(W) - f^*)$
  (by Grönwall's inequality)
  $\rightarrow (f(W(t)) - f^*) \leq \exp(-\gamma t)\left(f\left(W(0)\right) - f^*\right)$
  $f\left(W(t)\right)$ converges to f^* exponentially
- Rate: PL-Constant γ

Convergence under overparametrization

Non-overparametrized

$f(W)$

- Gradient Flow: $\dot{W} = -\nabla f(W)$
- Global PL-Inequality
  $\|\nabla f(W)\|_F^2 \geq \gamma(f(W) - f^*)$
- f converges **exponentially** to f^* regardless of initialization
- Rate = PL-Constant γ

Overparametrized

$\mathcal{L}(W_1, \cdots, W_L) = f(W_1 W_2 \cdots W_L)$

- Gradient Flow: $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$
- **Local (Weight-dependent) PL-inequality**
  ??
- $\mathcal{L}$ converges **exponentially** to $\mathcal{L}^*$ under **proper initialization**
- $Rate \geq \gamma\sqrt{(Imbalance)^2 + 4(Margin)^2}$

Outline

- Problem Settings
- Warm-up Example
- Convergence Analysis for Multi-layer Linear Model
- Convergence Rate Bound
- Conclusion

Warm-up Example: Scalar dynamics

- $f(w)$ is a function of scalar $w \in \mathbb{R}$

- PL-inequality

$$|f'(w)|^2 \geq \gamma(f(w) - f^*), \qquad \forall w$$

- Simplest overparametrization $w \to uv$

$$\mathcal{L}(u, v) = f(uv)$$

Scalar Dynamics: Imbalance

- Gradient flow induces **conservation law**

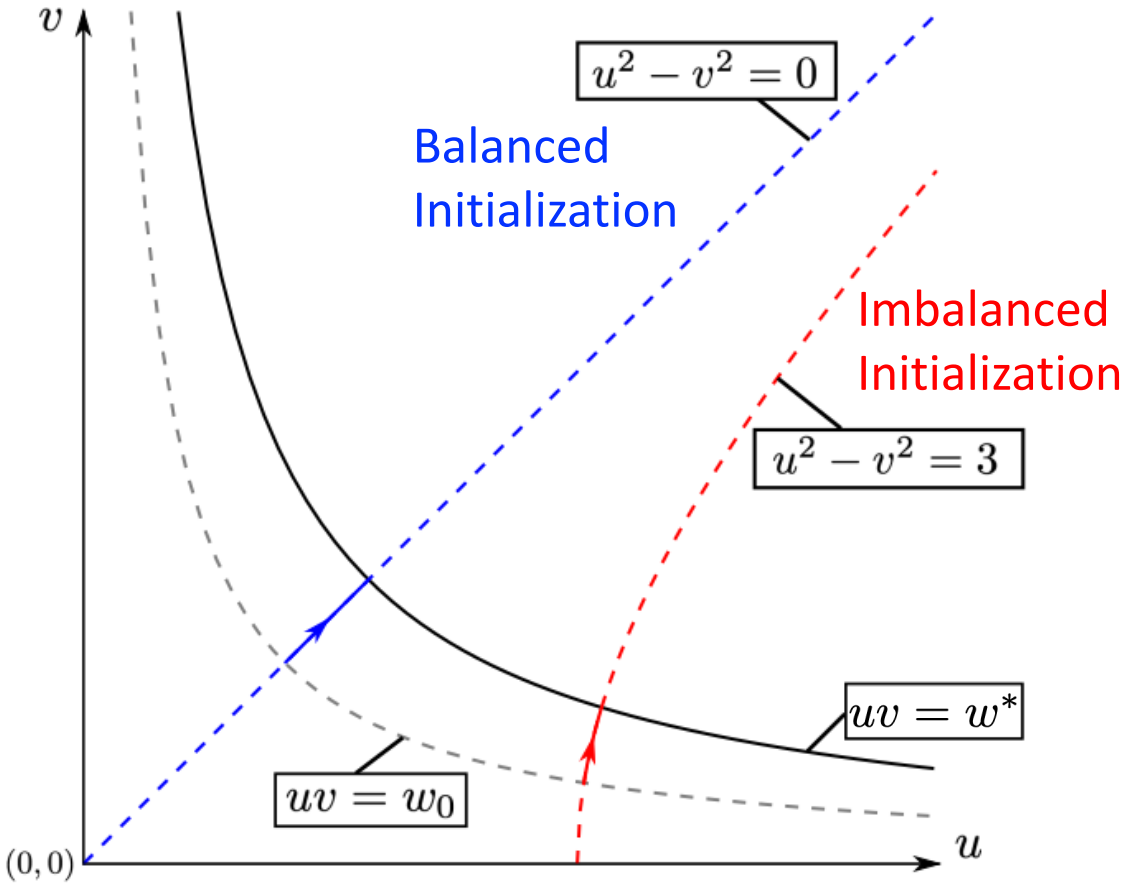
$$\dot{u} = -f'(uv)v$$
$$\dot{v} = -f'(uv)u$$
$$\Longrightarrow \quad d \coloneqq u^2 - v^2 \ , \dot{d} \equiv 0$$

- **imbalance** $d \coloneqq u^2 - v^2$ is time-invariant

- Conservation law arises due to scaling **symmetry**

$$u \to su, \qquad v \to \frac{v}{s}$$

(Noether's Theorem connects symmetry to conservation law)



Scalar Dynamics: Weight-dependent PL inequality

- Gradient flow on $\mathcal{L}(u,v) = f(uv)$

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

- $\|\nabla\mathcal{L}\|_F^2 = |f'(uv)|^2(u^2+v^2)$

PL-inequality $|f'|^2 \geq \gamma(f - f^*)$

- $\|\nabla\mathcal{L}(u,v)\|_F^2 \geq \gamma(u^2+v^2)(\mathcal{L}(u,v) - \mathcal{L}^*)$
  (weight-dependent PL-inequality)

- Given initialization $u(0), v(0)$, find a lower bound for $u^2(t) + v^2(t)$

Scalar Dynamics: Rate Bound

- $\|\nabla\mathcal{L}\|_F^2 \geq \gamma(u^2 + v^2)(\mathcal{L} - \mathcal{L}^*)$
- $\|\nabla\mathcal{L}\|_F^2 \geq \gamma\sqrt{d^2 + 4(uv)^2}(\mathcal{L} - \mathcal{L}^*)$

Express u^2, v^2 by

imbalance $d := u^2 - v^2$

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: Rate Bound

- $\|\nabla\mathcal{L}\|_F^2 \geq \gamma(u^2+v^2)(\mathcal{L}-\mathcal{L}^*)$

- $\|\nabla\mathcal{L}\|_F^2 \geq \gamma\sqrt{d^2+4(uv)^2}(\mathcal{L}-\mathcal{L}^*)$

- f is $\boldsymbol{\mu}$-strongly convex, and $\boldsymbol{K}$-smooth
- Loss $\mathcal{L}$ is non-increasing

imbalance d is time invariant

$$|d(t)| = |d(0)| := Imbalance$$

A lower bound on product uv

$$|u(t)v(t)| \geq \left[|w^*| - \sqrt{\boldsymbol{K}/\boldsymbol{\mu}}\,|w^* - u(0)v(0)|\right]_+ := Margin$$

$$(u^2+v^2) = \gamma\sqrt{(Imbalance)^2+4(Product)^2} \geq \gamma\sqrt{(Imbalance)^2+4(Margin)^2}$$

Scalar Dynamics: Summary

Local (Weight-dependent) PL-inequality

$$\|\nabla\mathcal{L}\|_F^2 \geq \gamma(u^2+v^2)(\mathcal{L}-\mathcal{L}^*)$$

"Weight" to imbalance and product

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

Control imbalance and product by initialization

- **Imbalance** is time invariant
- $Product \geq Margin$

Initialization-dependent PL-inequality → Exponential Convergence

$$\|\nabla\mathcal{L}\|_F^2 \geq \gamma\sqrt{(Imbalance)^2+4(Margin)^2}(\mathcal{L}-\mathcal{L}^*)$$

(Grönwall)

$$\Longrightarrow (\mathcal{L}(t)-\mathcal{L}^*) \leq \exp\left(-\gamma\sqrt{(Imbalance)^2+4(Margin)^2}t\right)(\mathcal{L}(0)-\mathcal{L}^*)$$

Outline

- Problem Settings
- Warm-up Example
- Convergence Analysis for Multi-layer Linear Model
- Convergence Rate Bound
- Conclusion

To General Case

	Warm-up Example: $f(uv)$	**General Case**: $f(W_1 W_2 \cdots W_L)$
Imbalance	$d := u^2 - v^2$	$\{D_l := W_l^T W_l - W_{l+1} W_{l+1}^T\}_{l=1}^{L-1}$
Margin	$\left[\lvert w^* \rvert - \sqrt{K/\mu}\,\lvert w^* - u(0)v(0)\rvert\right]_+$	$\left[\sigma_{\min}(W^*) - \sqrt{K/\mu}\,\lVert W^* - W(0)\rVert_F\right]_+$
Local PL-ineq	$\lVert\nabla\mathcal{L}\rVert_F^2 \geq \gamma(u^2+v^2)(\mathcal{L}-\mathcal{L}^*)$	$\lVert\nabla\mathcal{L}\rVert_F^2 \geq \gamma \cdot \lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right)(\mathcal{L}-\mathcal{L}^*)$
Weight to imbalance and product	$(u^2+v^2) = \sqrt{d^2 + 4(uv)^2}$	$\lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right) \geq \alpha(Imbalance, \sigma_{\min}(W))$

Control imbalance and product by initialization

Exponential Convergence:

$$Rate \geq \gamma\, \alpha(Imbalance, Margin)$$

	Warm-up Example: $f(uv)$	**General Case**: $f(W_1 W_2 \cdots W_L)$
Imbalance	$d := u^2 - v^2$	$\{D_l := W_l^T W_l - W_{l+1} W_{l+1}^T\}_{l=1}^{L-1}$
Margin	$\left[\lvert w^* \rvert - \sqrt{K/\mu}\,\lvert w^* - u(0)v(0)\rvert\right]_+$	$\left[\sigma_{min}(W^*) - \sqrt{K/\mu}\,\lVert W^* - W(0)\rVert_F\right]_+$
Local PL-ineq	$\lVert\nabla\mathcal{L}\rVert_F^2 \geq \gamma(u^2+v^2)(\mathcal{L}-\mathcal{L}^*)$	$\lVert\nabla\mathcal{L}\rVert_F^2 \geq \gamma \cdot \lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right)(\mathcal{L}-\mathcal{L}^*)$
Weight to imbalance and product	$(u^2+v^2) = \sqrt{d^2 + 4(uv)^2}$	$\lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right) \geq \alpha(Imbalance, \sigma_{\min}(W))$

Control imbalance and product by initialization

Exponential Convergence:

$$Rate \geq \gamma\, \alpha(Imbalance, Margin)$$

General Convergence Analysis: Imbalance

- $\mathcal{L}(W_1, \cdots, W_L) = f(W_1 W_2 \cdots W_L)$
- Gradient Flow: $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$

- **Imbalance matrices** $D_1, \ D_2, \cdots, D_{L-1}$

$$W = \underbrace{W_1}_{(n \times h_1)} \cdot \underbrace{W_2}_{(h_1 \times h_2)} \cdot W_3 \cdots W_{L-1} \cdot W_L$$

Symmetry: $W_1 \to W_1 S, \ W_2 \to S^{-1} W_2$

Conservation law

$$D_1 = W_1^T W_1 - W_2 W_2^T, \qquad \dot{D}_1 \equiv 0$$

General Convergence Analysis: Imbalance

- $\mathcal{L}(W_1, \cdots, W_L) = f(W_1 W_2 \cdots W_L)$
- Gradient Flow: $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$

- **Imbalance matrices** $D_1, \boxed{D_2}, \cdots, D_{L-1}$

$$W = W_1 \cdot \underbrace{W_2}_{(h_1 \times h_2)} \cdot \underbrace{W_3}_{(h_2 \times h_3)} \cdots W_{L-1} \cdot W_L$$

$$D_2 = W_2^T W_2 - W_3 W_3^T$$

General Convergence Analysis: Imbalance

- $\mathcal{L}(W_1, \cdots, W_L) = f(W_1 W_2 \cdots W_L)$
- Gradient Flow: $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$

- **Imbalance matrices** $D_1,\ D_2, \cdots, \boxed{D_{L-1}}$

$$W = W_1 \cdot W_2 \cdot W_3 \cdots \underset{(h_{L-2} \times h_{L-1})}{W_{L-1}} \cdot \underset{(h_{L-1} \times m)}{W_L}$$

$$D_{L-1} = W_{L-1}^T W_{L-1} - W_L W_L^T$$

General Convergence Analysis: Imbalance

- $\mathcal{L}(W_1, \cdots, W_L) = f(W_1 W_2 \cdots W_L)$
- Gradient Flow: $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$

- Imbalance matrices $\left\{ D_l \coloneqq W_l^T W_l - W_{l+1} W_{l+1}^T \right\}_{l=1}^{L-1}$

- **Imbalance matrices are time-invariant under GF**

$$\dot{D}_l \equiv 0, \qquad l = 1, \cdots, L-1$$

	Warm-up Example: $f(uv)$	**General Case**: $f(W_1 W_2 \cdots W_L)$
Imbalance	$d := u^2 - v^2$	$\{D_l := W_l^T W_l - W_{l+1} W_{l+1}^T\}_{l=1}^{L-1}$
Margin	$\left[\lvert w^* \rvert - \sqrt{K/\mu}\,\lvert w^* - u(0)v(0)\rvert\right]_+$	$\left[\sigma_{min}(W^*) - \sqrt{K/\mu}\,\lVert W^* - W(0)\rVert_F\right]_+$
Local PL-ineq	$\lVert\nabla\mathcal{L}\rVert_F^2 \geq \gamma(u^2+v^2)(\mathcal{L}-\mathcal{L}^*)$	$\lVert\nabla\mathcal{L}\rVert_F^2 \geq \gamma \cdot \lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right)(\mathcal{L}-\mathcal{L}^*)$
Weight to imbalance and product	$(u^2+v^2) = \sqrt{d^2 + 4(uv)^2}$	$\lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right) \geq \alpha(Imbalance, \sigma_{\min}(W))$

Control imbalance and product by initialization

Exponential Convergence:

$$Rate \geq \gamma\, \alpha(Imbalance, Margin)$$

General Convergence Analysis: PL-inequality

- $\mathcal{L}(W_1, \cdots, W_L) = f(W_1 W_2 \cdots W_L)$
- Gradient Flow: $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$

Recall warm-up example:

$$\|\nabla \mathcal{L}\|_F^2 = |f'(uv)|^2 (u^2 + v^2) = \langle (u^2 + v^2) f', f' \rangle,$$

- $\left\| \nabla \mathcal{L}(\{W_l\}_{l=1}^L) \right\|_F^2 = \left\langle \mathcal{T}_{\{W_l\}_{l=1}^L} \nabla f(W), \quad \nabla f(W) \right\rangle_F$

$\mathcal{T}_{\{W_l\}_{l=1}^L}$ is a positive semi-definite operator on $\mathbb{R}^{n \times m}$

- $\mathcal{L} = f(W_1 W_2)$,

$$\mathcal{T}_{\{W_1, W_2\}} E = W_1 W_1^T E + E W_2^T W_2$$

- $\mathcal{L} = f(W_1 W_2 W_3)$,

$$\mathcal{T}_{\{W_1, W_2, W_3\}} E = W_1 W_2 W_2^T W_1^T E + W_1 W_1^T E W_3^T W_3 + E W_3^T W_2^T W_2 W_3$$

General Convergence Analysis: PL-inequality

- $\mathcal{L}(W_1, \cdots, W_L) = f(W_1 W_2 \cdots W_L)$, Gradient Flow: $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$

- $$\left\| \nabla \mathcal{L}(\{W_l\}_{l=1}^L) \right\|_F^2 = \left\langle \mathcal{T}_{\{W_l\}_{l=1}^L} \nabla f(W), \quad \nabla f(W) \right\rangle_F$$

$$\geq \lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right) \|\nabla f(W)\|_F^2$$

$$\geq \gamma \lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right) (\mathcal{L} - \mathcal{L}^*)$$

Min-max theorem

PL: $\|\nabla f\|_F^2 \geq \gamma (f - f^*)$

- **Local (weight-dependent) PL-inequality**

$$\left\| \nabla \mathcal{L}(\{W_l\}_{l=1}^L) \right\|_F^2 \geq \gamma \lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right) (\mathcal{L} - \mathcal{L}^*)$$

	Warm-up Example: $f(uv)$	**General Case**: $f(W_1 W_2 \cdots W_L)$
Imbalance	$d := u^2 - v^2$	$\{D_l := W_l^T W_l - W_{l+1} W_{l+1}^T\}_{l=1}^{L-1}$
Margin	$\left[\lvert w^* \rvert - \sqrt{K/\mu}\, \lvert w^* - u(0)v(0) \rvert\right]_+$	$\left[\sigma_{min}(W^*) - \sqrt{K/\mu}\, \lVert W^* - W(0) \rVert_F\right]_+$
Local PL-ineq	$\lVert \nabla \mathcal{L} \rVert_F^2 \geq \gamma (u^2 + v^2)(\mathcal{L} - \mathcal{L}^*)$	$\lVert \nabla \mathcal{L} \rVert_F^2 \geq \gamma \cdot \lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right)(\mathcal{L} - \mathcal{L}^*)$
Weight to imbalance and product	$(u^2 + v^2) = \sqrt{d^2 + 4(uv)^2}$	$\lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right) \geq \alpha(Imbalance, \sigma_{\min}(W))$

Control imbalance and product by initialization

Exponential Convergence:

$$Rate \geq \gamma\, \alpha(Imbalance, Margin)$$

Lower Bound on Convergence Rate: Summary

Linear model	Rate Bound	Expression
Multi-layer Scalar weights $f(w_1 w_2 \cdots w_L)$	$\alpha(\{d_l\}_{l=1}^{L-1}, w)$	$\sqrt{\left(\prod Cumulative\ imbalance\right)^2 + (Lw^{2-2/L})^2}$
Two-layer Matrix weights $f(W_1 W_2)$	$\alpha(D_1, \sigma_{\min}(W))$	$-Spread + \sqrt{(Spread + Gap)^2 + 4\sigma_{\min}^2(W)}$
Three-layer Matrix weights $f(W_1 W_2 W_3)$	$\alpha(D_1, D_2)$	A complicated expression $\approx \sum \prod Cumulative\ imbalance$

Lower Bound on Convergence Rate: Summary

Linear model	Rate Bound	Expression
Multi-layer Scalar weights $f(w_1 w_2 \cdots w_L)$	$\alpha(\{d_l\}_{l=1}^{L-1}, w)$	$\sqrt{\left(\prod Cumulative\ imbalance\right)^2 + (Lw^{2-2/L})^2}$
Two-layer Matrix weights $f(W_1 W_2)$	$\alpha(D_1, \sigma_{\min}(W))$	$-Spread + \sqrt{(Spread + Gap)^2 + 4\sigma_{\min}^2(W)}$
Three-layer Matrix weights $f(W_1 W_2 W_3)$	$\alpha(D_1, D_2)$	A complicated expression $\approx \sum \prod Cumulative\ imbalance$

Details to come!

Convergence under overparametrization: Summary

Non-overparametrized

$f(W)$

- Gradient Flow: $\dot{W} = -\nabla f(W)$
- Global PL-Inequality
  $\|\nabla f(W)\|_F^2 \geq \gamma(f(W) - f^*)$
- f converges **exponentially** to f^* regardless of initialization
- Rate = PL-Constant γ

Overparametrized

$L(W_1, \cdots, W_L) = f(W_1 W_2 \cdots W_L)$

- Gradient Flow: $\dot{W}_l = -\partial L / \partial W_l$
- **Local (Weight-dependent) PL-inequality**
  $\left\|\nabla \mathcal{L}(\{W_l\}_{l=1}^L)\right\|_F^2 \leq \gamma \cdot \lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right)(\mathcal{L} - \mathcal{L}^*)$
- $\mathcal{L}$ converges **exponentially** to $\mathcal{L}^*$ under **proper initialization**
- $Rate \geq \gamma\, \alpha(Imbalance, Margin)$

Outline

- Problem Settings
- Warm-up Example
- Meta-proof for Convergence
- Convergence Rate Bound

$$\lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^{L}}\right) \geq \alpha(\, Imbalance, \sigma_{\min}(W)\,)$$

- Conclusion

Lower Bound on Convergence Rate: Overview

- We want a lower bound that depends on both imbalance and product

$$\lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^{L}}\right) \geq \alpha(\ \{D_l\}_{l=1}^{L-1}, \sigma_{\min}(W)\ )$$

- The (arguably) optimal bound is given by

$$(*) \quad \min_{\{W_l\}_{l=1}^{L}} \ \lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^{L}}\right)$$
$$s.t. \ \ W_l^T W_l - W_{l+1} W_{l+1}^T = D_l, \ \ l = 1, \cdots, L-1$$
$$W_1 W_2 \cdots W_L = W$$

- We will compare our bound to the optimal value of (*)

Lower Bound on Convergence Rate

Linear model	Rate Bound	Expression
Multi-layer Scalar weights $f(w_1 w_2 \cdots w_L)$	$\alpha(\{d_l\}_{l=1}^{L-1}, w)$	$\sqrt{\left(\prod Cumulative\ imbalance\right)^2 + (Lw^{2-2/L})^2}$
Two-layer Matrix weights $f(W_1 W_2)$	$\alpha(D_1, \sigma_{\min}(W))$	$-Spread + \sqrt{(Spread + Gap)^2 + 4\sigma_{\min}^2(W)}$
Three-layer Matrix weights $f(W_1 W_2 W_3)$	$\alpha(D_1, D_2)$	A complicated expression $\approx \sum \prod Cumulative\ imbalance$

Multi-layer Scalar Networks

- $f(w)$ is a function of scalar $w \in \mathbb{R}$

- Multi-layer scalar networks

$$f(w_1 w_2 \cdots w_L), \qquad w_l \in \mathbb{R}, l = 1, \cdots L$$

- Imbalance

$$d_l \coloneqq w_l^2 - w_{l+1}^2, \qquad l = 1, \cdots L-1$$

- Analysis for scalar networks arises when studying general matrix model
  - under specific initialization (spectral initialization)
  - with additional network structure (diagonal linear networks)

Multi-layer Scalar Networks: Formulation

- We want a lower bound that depends on both imbalance and product

$$\lambda_{\min}\left(\mathcal{T}_{\{w_l\}_{l=1}^{L}}\right) \geq \alpha(\{d_l\}_{l=1}^{L-1}, w)$$

- Ideally, we want to solve

$$\min_{\{w_l\}_{l=1}^{L}} \lambda_{\min}\left(\mathcal{T}_{\{w_l\}_{l=1}^{L}}\right) = \sum_{l=1}^{L}\prod_{i\neq l} w_i^2 = \sum_{l=1}^{L}\frac{w^2}{w_l^2}$$

$$s.t. \quad w_l^2 - w_{l+1}^2 = d_l, \ l = 1, \cdots, L-1$$

$$w_1 w_2 \cdots w_L = w$$

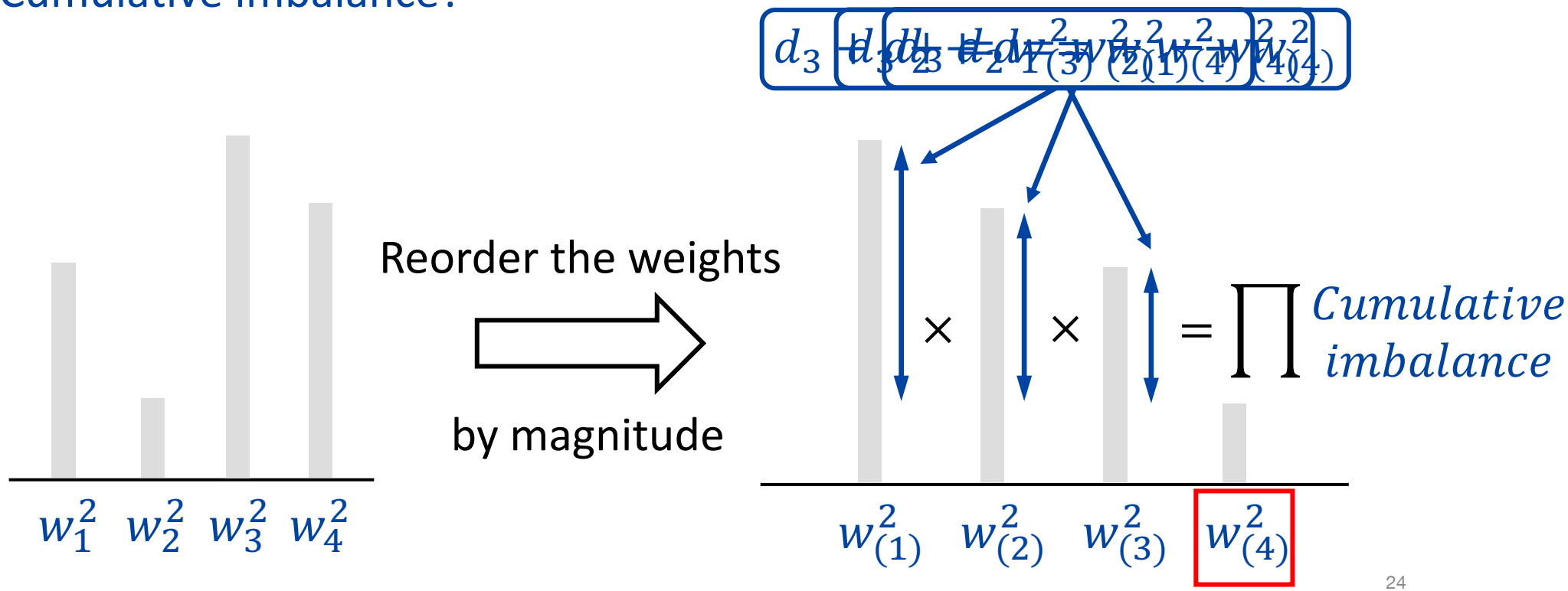
  - Only isolated points in the feasible set
  - **All feasible points** have the **same** objective value $\alpha^*(\{d_l\}_{l=1}^{L-1}, w)$
  - $\alpha^*(\{d_l\}_{l=1}^{L-1}, w)$ has **no closed-form expression** in general
    (need to solve an L-th order polynomial)

Multi-layer Scalar Networks: Rate Bound

- $\alpha^*(\{d_l\}_{l=1}^{L-1}, w)$ has **no closed-form expression** in general

***Proposition** 1.* $\alpha^*(\{d_l\}_{l=1}^{L-1}, w) \geq \sqrt{\left(\prod \mathit{Cumulative\ imbalance}\right)^2 + (Lw^{2-2/L})^2}$

- Cumulative imbalance?

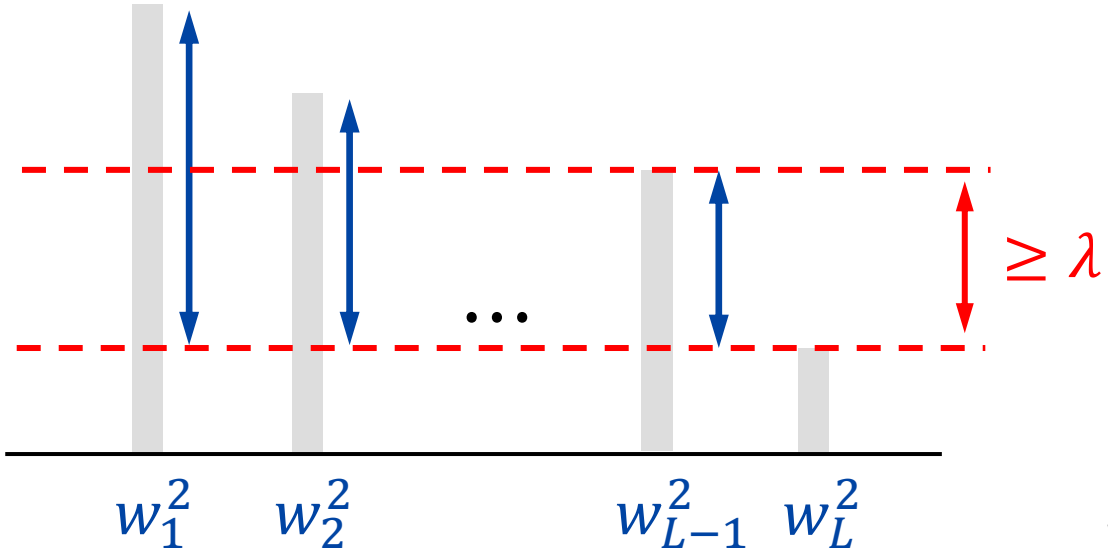


Multi-layer Scalar Networks: Effect of Imbalance

- $\alpha^*(\{d_l\}_{l=1}^{L-1}, w)$ has **no closed-form expression** in general

***Proposition** 1.* $\alpha^*(\{d_l\}_{l=1}^{L-1}, w) \geq \sqrt{\left(\prod \textit{Cumulative imbalance}\right)^2 + (Lw^{2-2/L})^2}$

- Effect of imbalance in **deep** networks



$$\sqrt{(\lambda^{L-1})^2 + (Lw^{2-2/L})^2}$$

[Yun et al. 20] derived rate bound λ^{L-1} (w.o. product)

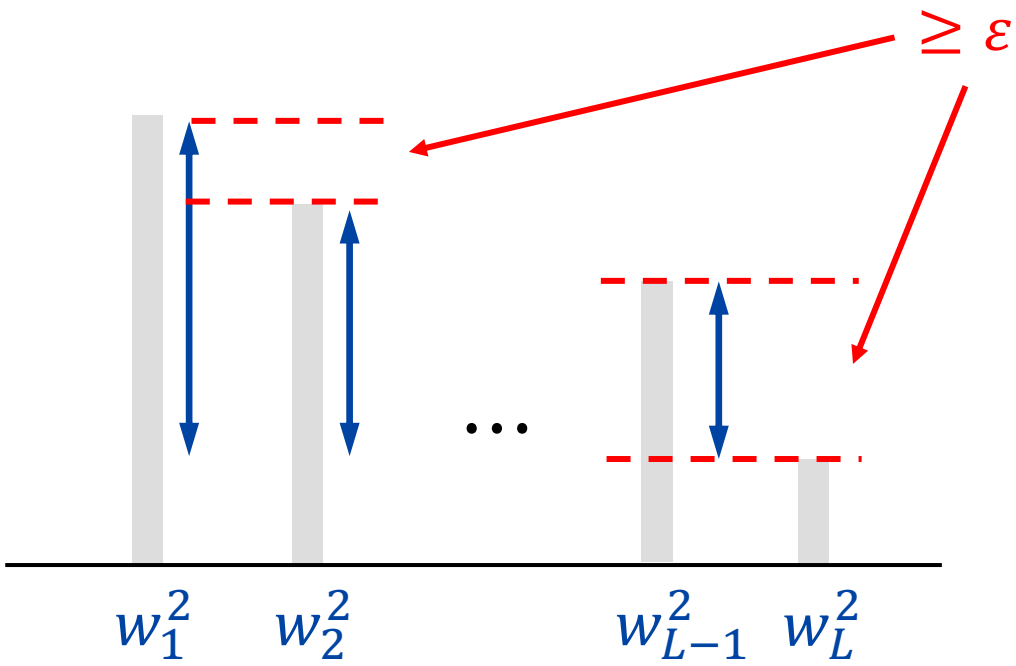
Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. ICLR 2020

Multi-layer Scalar Networks: Effect of Imbalance

- $\alpha^*(\{d_l\}_{l=1}^{L-1}, w)$ has **no closed-form expression** in general

***Proposition** 1.* $\alpha^*(\{d_l\}_{l=1}^{L-1}, w) \geq \sqrt{\left(\prod \begin{matrix} Cumulative \\ imbalance \end{matrix}\right)^2 + (Lw^{2-2/L})^2}$

- Effect of imbalance in **deep** networks



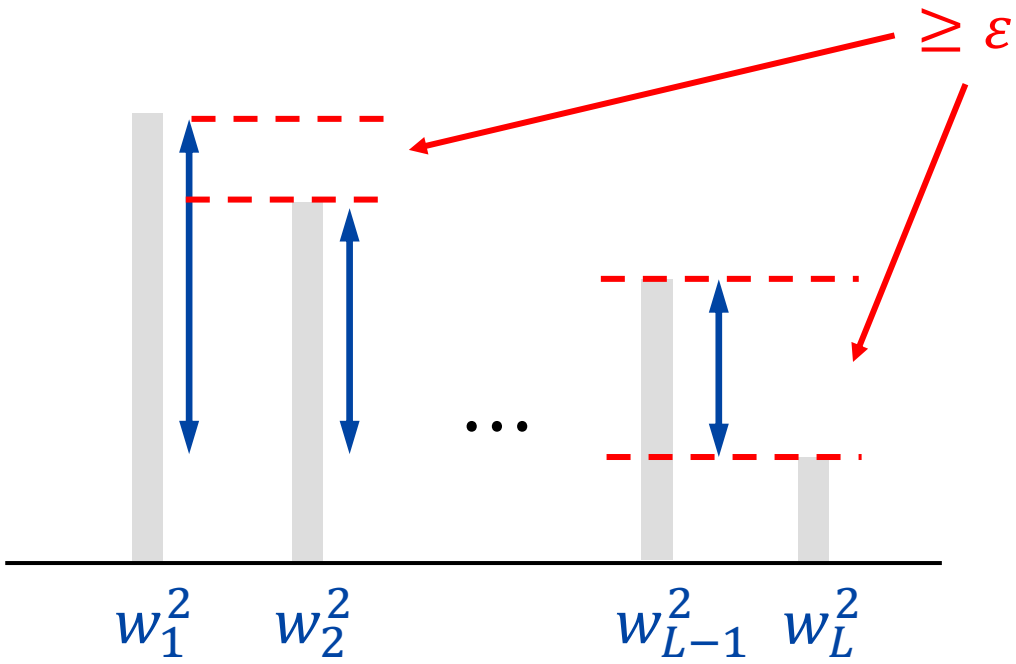
$$\sqrt{(\varepsilon^{L-1}(L-1)!)^2 + (Lw^{2-2/L})^2}$$

Multi-layer Scalar Networks: Effect of Imbalance

- $\alpha^*(\{d_l\}_{l=1}^{L-1}, w)$ has **no closed-form expression** in general

***Proposition** 1.* $\alpha^*(\{d_l\}_{l=1}^{L-1}, w) \geq \sqrt{\left(\prod \begin{matrix} Cumulative \\ imbalance \end{matrix}\right)^2 + (Lw^{2-2/L})^2}$

- Effect of imbalance in **deep** networks



- Imbalanced initialization could **accelerate convergence** significantly for deep networks
  (For gradient flow!)
- In practice, this is more related to **exploding gradient**

Lower Bound on Convergence Rate

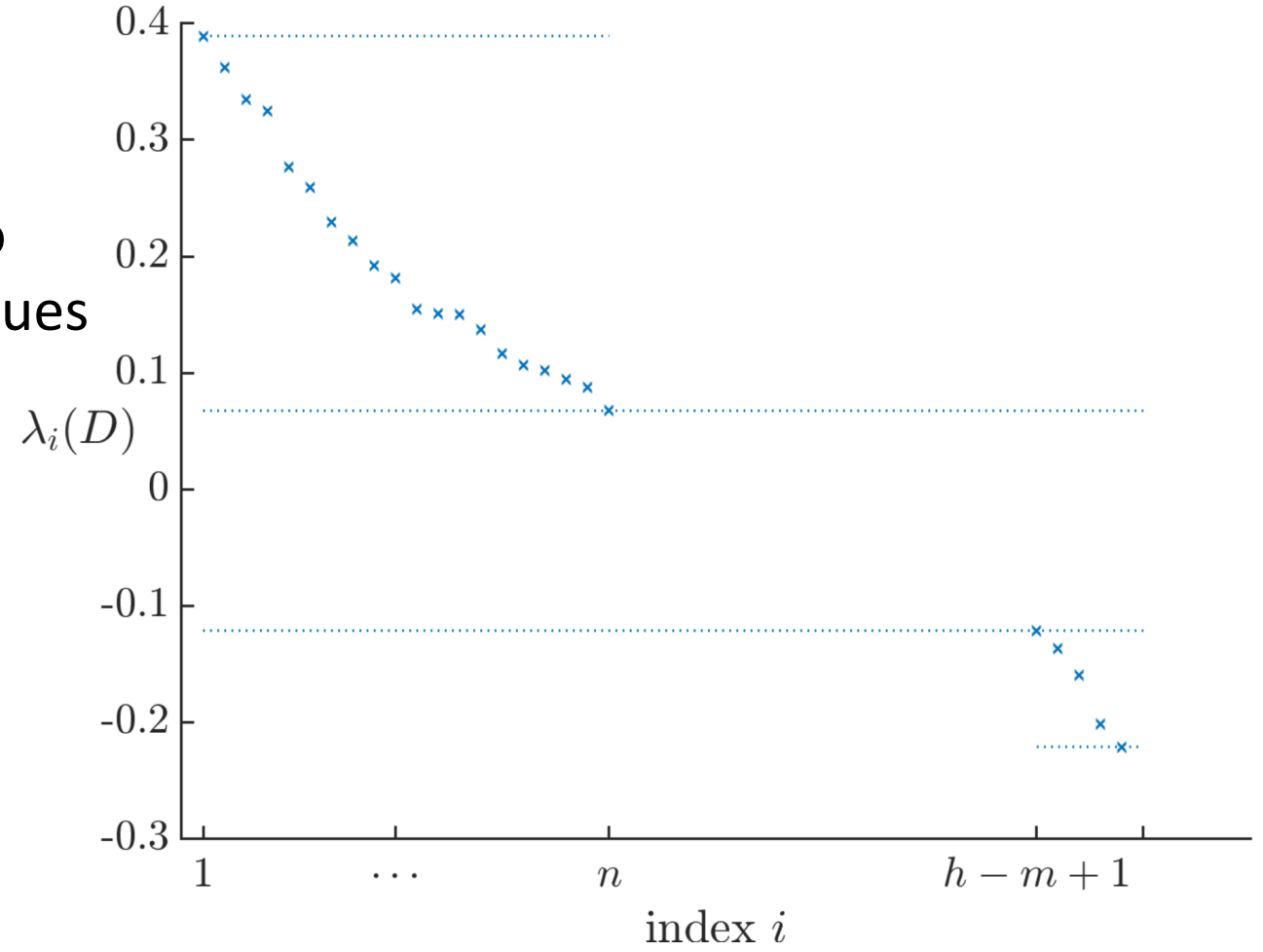
Linear model	Rate Bound	Expression
Multi-layer Scalar weights $f(w_1 w_2 \cdots w_L)$	$\alpha(\{d_l\}_{l=1}^{L-1}, w)$	$\sqrt{\left(\prod Cumulative\ imbalance\right)^2 + (Lw^{2-2/L})^2}$
Two-layer Matrix weights $f(W_1 W_2)$	$\alpha(D_1, \sigma_{\min}(W))$	$-Spread + \sqrt{(Spread + Gap)^2 + 4\sigma_{\min}^2(W)}$
Three-layer Matrix weights $f(W_1 W_2 W_3)$	$\alpha(D_1, D_2)$	A complicated expression $\approx \sum \prod Cumulative\ imbalance$

https://arxiv.org/abs/2105.06351

Imbalance quantities

- $\mathcal{L} = f(W_1 W_2), \quad (W_1 \in \mathbb{R}^{n\times h}, W_2 \in \mathbb{R}^{h\times m})$
- Imbalance $D_1 = W_1^T W_1 - W_2 W_2^T := D$

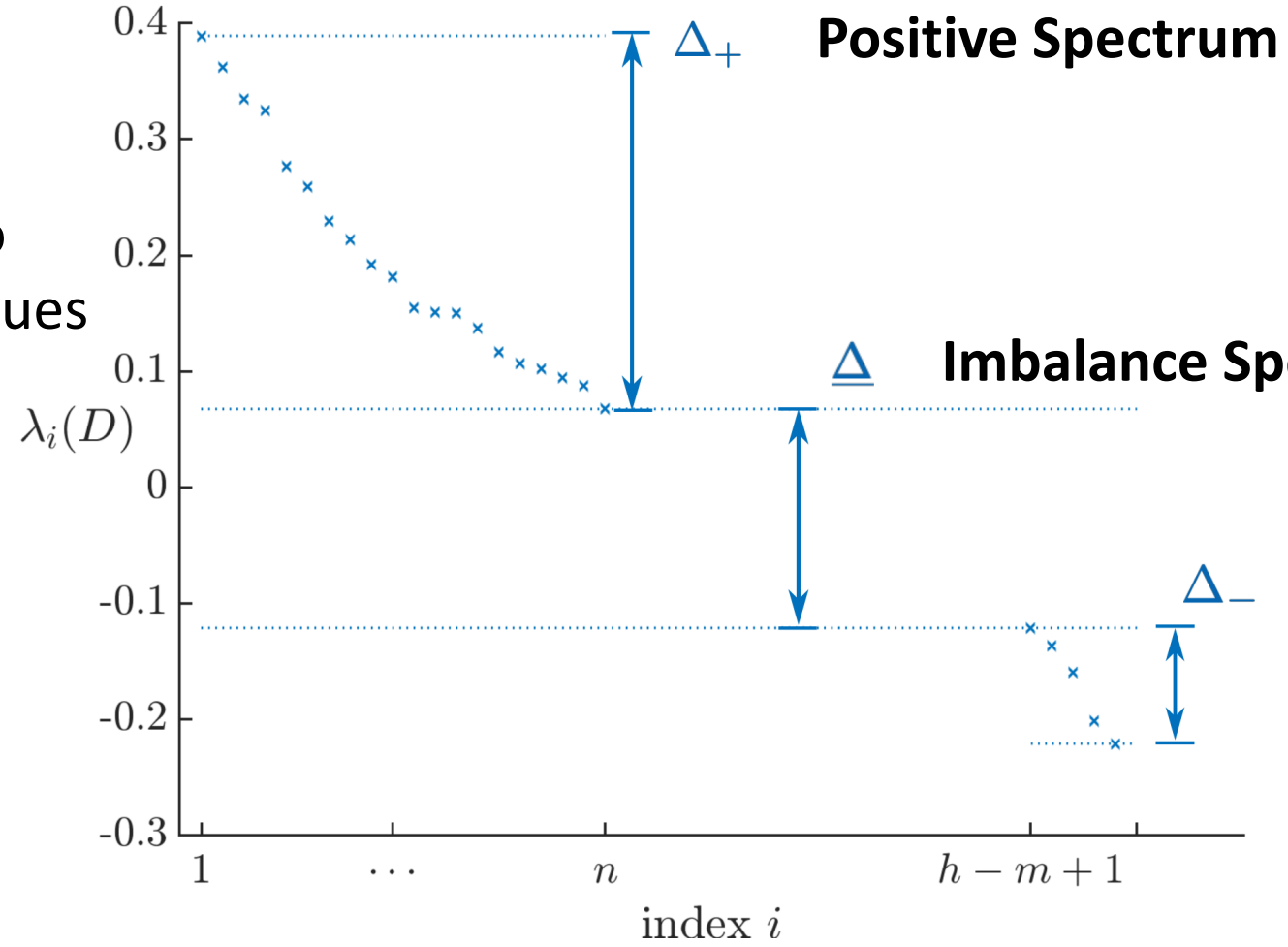
Nonzero Eigenvalues of D



Imbalance quantities

- $\mathcal{L} = f(W_1 W_2), \quad (W_1 \in \mathbb{R}^{n\times h}, W_2 \in \mathbb{R}^{h\times m})$
- Imbalance $D_1 = W_1^T W_1 - W_2 W_2^T := D$

Nonzero Eigenvalues of D



Positive Spectrum Spread

Imbalance Spectrum Gap

Negative Spectrum Spread

Two-layer Linear Networks

- We want a lower bound that depends on both imbalance and product

$$\lambda_{\min}\left(\mathcal{T}_{\{W_1,W_2\}}\right) \geq \alpha(D_1, \sigma_{\min}(W))$$

and

$$\mathcal{T}_{\{W_1,W_2\}}E = W_1W_1^TE + EW_2^TW_2$$

- Ideally, we want to find

$$\alpha^*(D_1, W) = \min_{\{W_1,W_2\}} \lambda_{\min}\left(\mathcal{T}_{\{W_1,W_2\}}\right) = \lambda_{\min}(W_1W_1^T) + \lambda_{\min}(W_2^TW_2)$$

$$s.t.\ \ W_1^TW_1 - W_2W_2^T = D_1$$

$$W_1W_2 = W$$

Two-layer Linear Networks: Rate Bound

***Proposition** 2.*

$$\alpha^*(D_1, W) \geq$$

$$-\Delta_+ + \sqrt{\left(\Delta_+ + \underline{\Delta}\right)^2 + 4\sigma_m^2(W)} \; -\Delta_- + \sqrt{\left(\Delta_- + \underline{\Delta}\right)^2 + 4\sigma_n^2(W)}$$

Equality holds when $n \neq m$

- $-Spread + \sqrt{(Spread + Gap)^2 + 4\sigma_{\min}^2(W)}$

Rate Bound Approximate form

- For the warm-up example $f(uv)$

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

- For the matrix case $f(W_1 W_2)$

$$Rate \geq -Spread + \sqrt{(Spread + Gap)^2 + 4\sigma^2(Product)}$$

  - When $Spread$ is <u>small</u>

$$Bound \approx \sqrt{(Gap)^2 + 4\sigma^2(Product)}$$

  - When $Spread$ is <u>large</u>

$$Bound \approx Gap$$

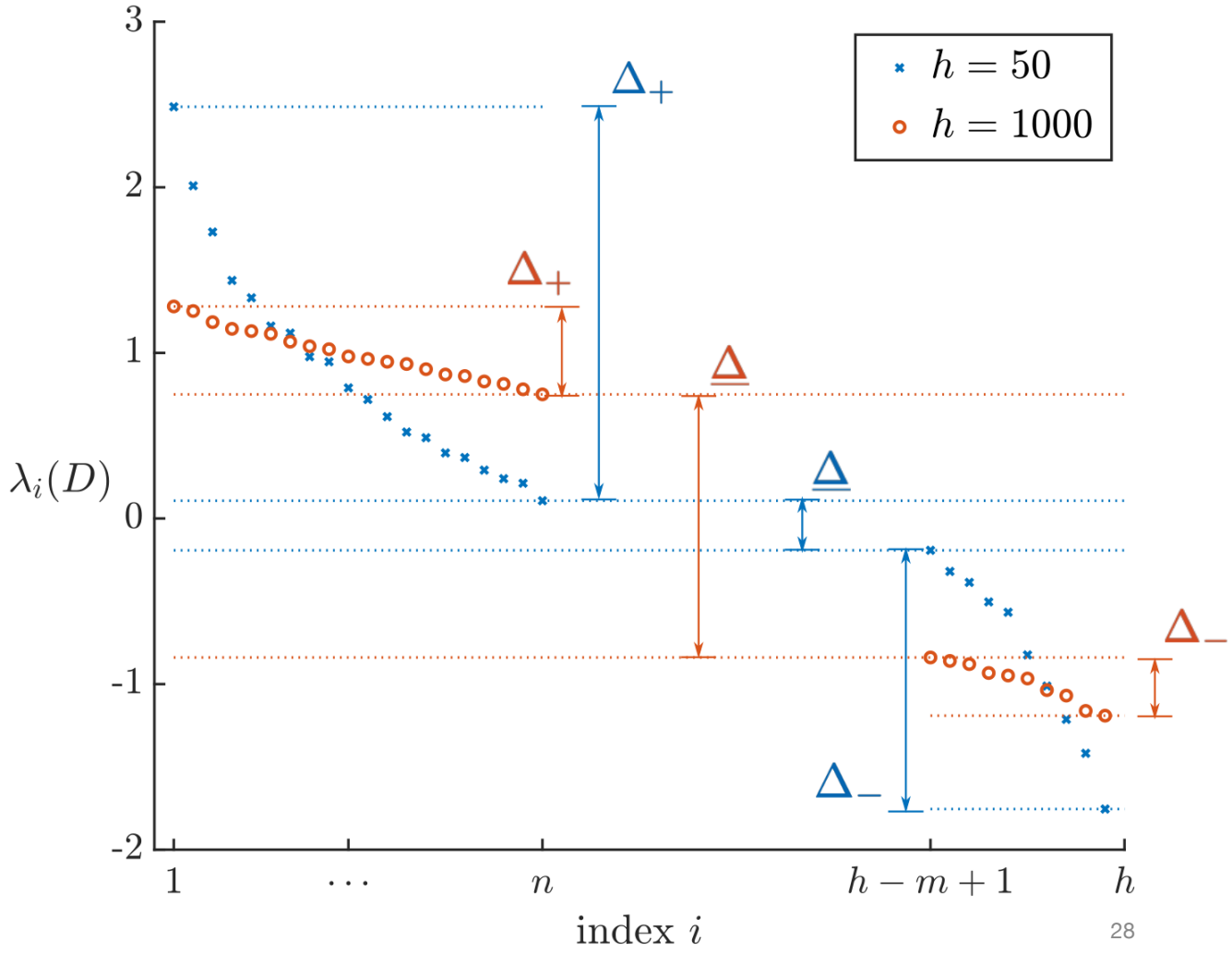
Width affects imbalance quantities

- $n = 20, m = 10$
- Random initialization

$$[W_1]_{ij}, [W_2]_{ij} \sim \mathcal{N}\left(0, \frac{1}{h}\right)$$

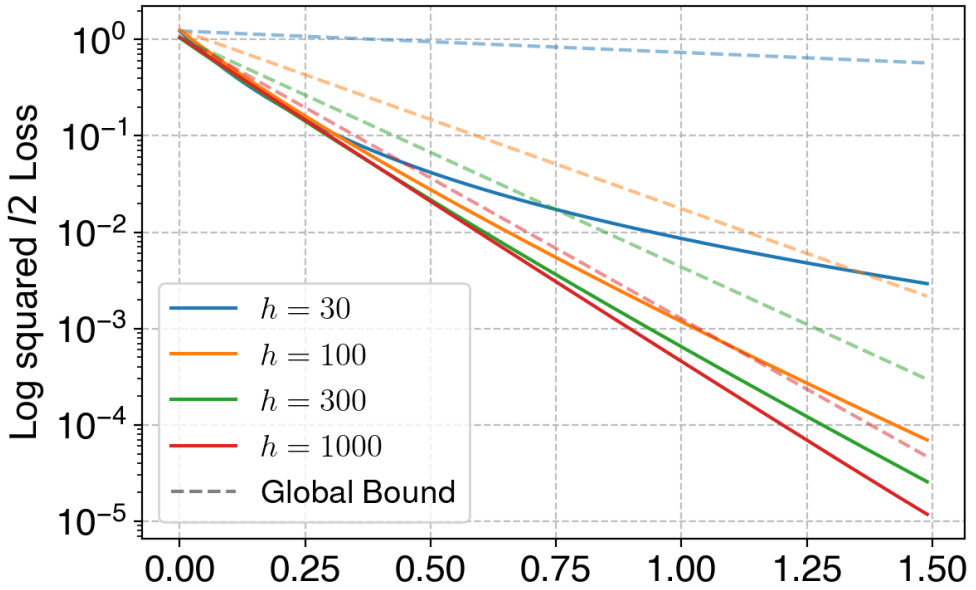
- Compare Gap with $Spread$
  - Small Width (h=50)
    Large $Spread$, small Gap
  - Large Width (h=1000)
    Small $Spread$, large Gap

All non-zero imbalance eigenvalues ($h \geq n + m$)

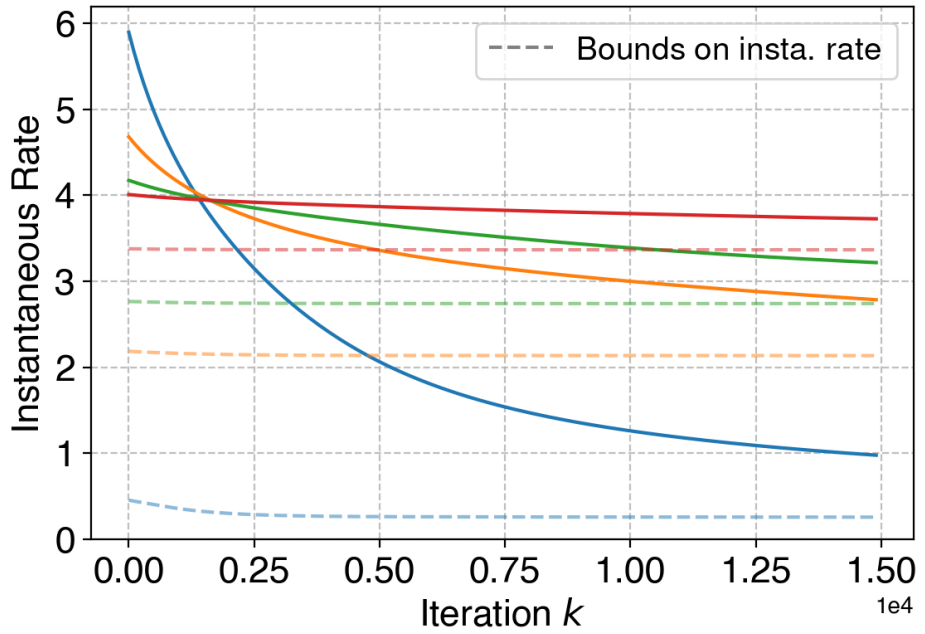


Imbalance quantities affects convergence "non-trivially"

- $n = 20, m = 5, L = \frac{1}{2}\left\|Y - \frac{1}{\sqrt{h}} W_1 W_2\right\|_F^2$
- Random init. $[W_1]_{ij}, [W_2]_{ij} \sim \mathcal{N}(0,1)$
- Target Y has small norm
  (Rate mainly depends on imbalance)
- Different loss curve
  - h=30, large spectrum spread,
    large initial rate then slows down,
    bound is good at late stage
  - h=1000, small spectrum spread,
    rate does not change too much,
    bound is good

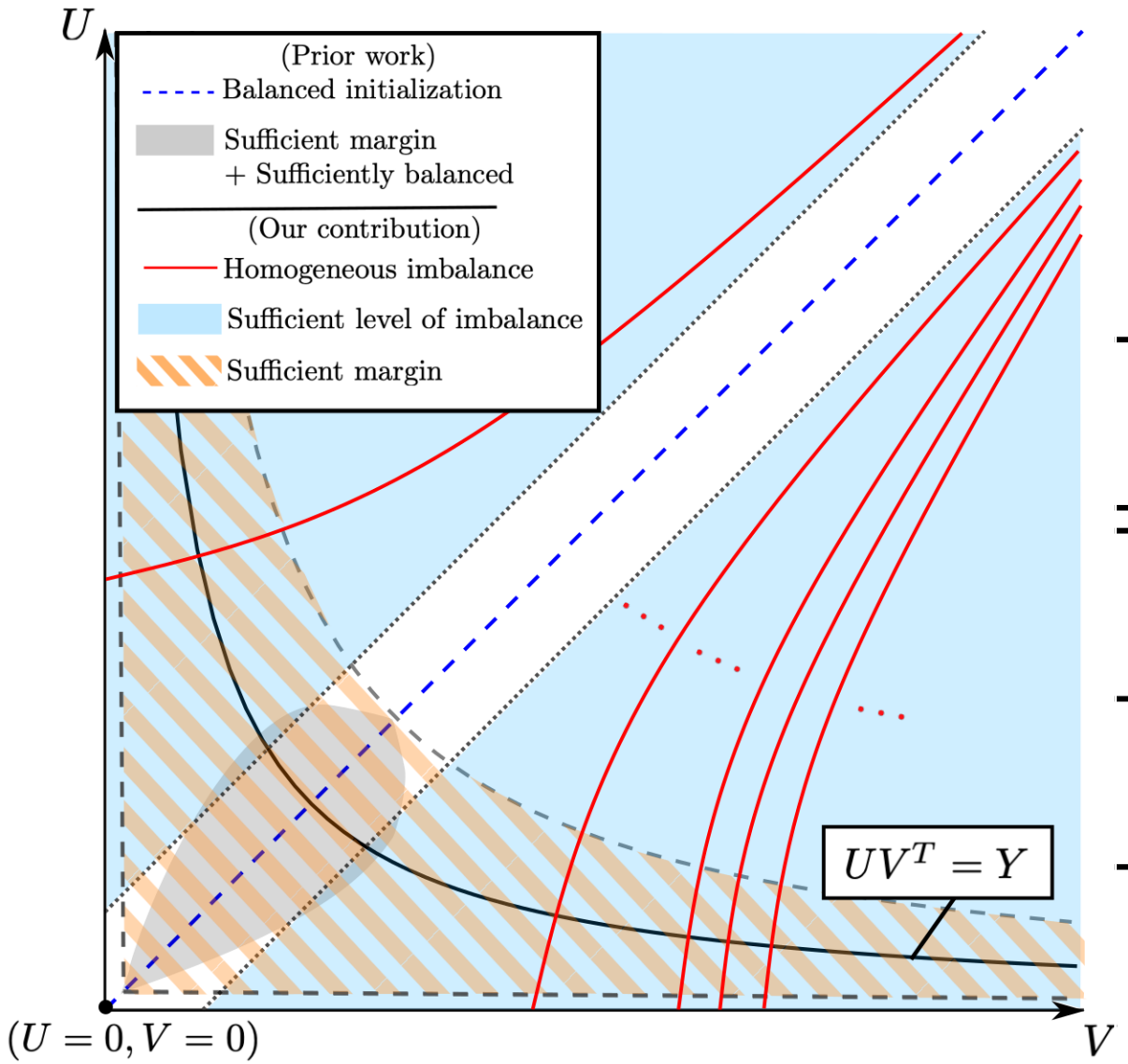


$\log L(k)$



$-\frac{\dot{L}(k)}{L(k)}$

Two-layer Linear Networks: Summary



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

Balanced initialization	$D := U^T U - V^T V = 0$

Outline

- Problem Settings
- Warm-up Example
- Meta-proof for Convergence
- Convergence Rate Bound
- Conclusion

Conclusion

We study the gradient flow on $\mathcal{L}(W_1, \cdots, W_L) = f(W_1 W_2 \cdots W_L)$:

$$Rate \geq \gamma\, \alpha(Imbalance, Margin)$$

Our analysis also works for classification task with exponential loss

$$\|\nabla f(W)\|_F \geq \gamma(f(W) - f^*) \Longrightarrow \mathcal{L}(t) = \mathcal{O}\left(\frac{1}{t}\right)$$

Future work:

- Gradient Descent (Ongoing work)
- Extension to 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