

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

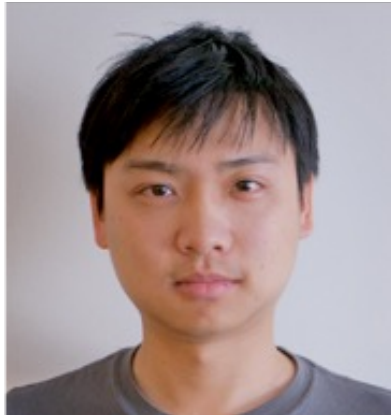
Enrique Mallada



Massachusetts Institute of Technology

August 25, 2022

Acknowledgements



Hancheng Min



Salma Tarmoun



Ziqing Xu



René Vidal

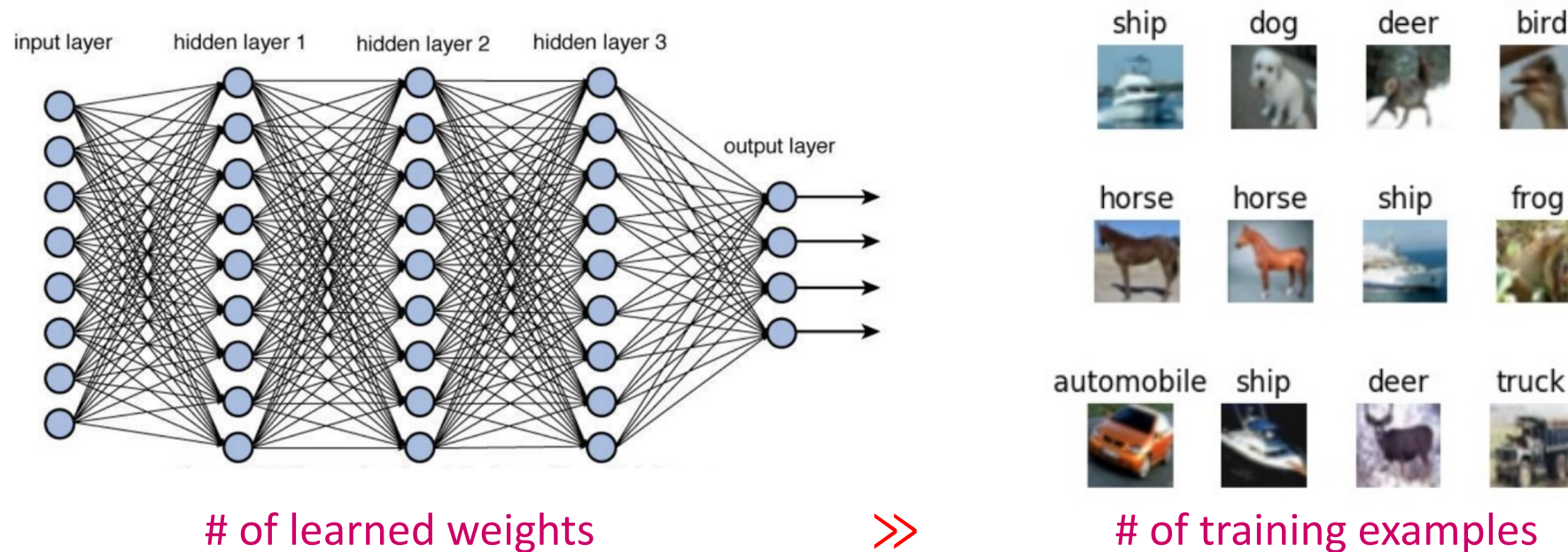


JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE



Introduction

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



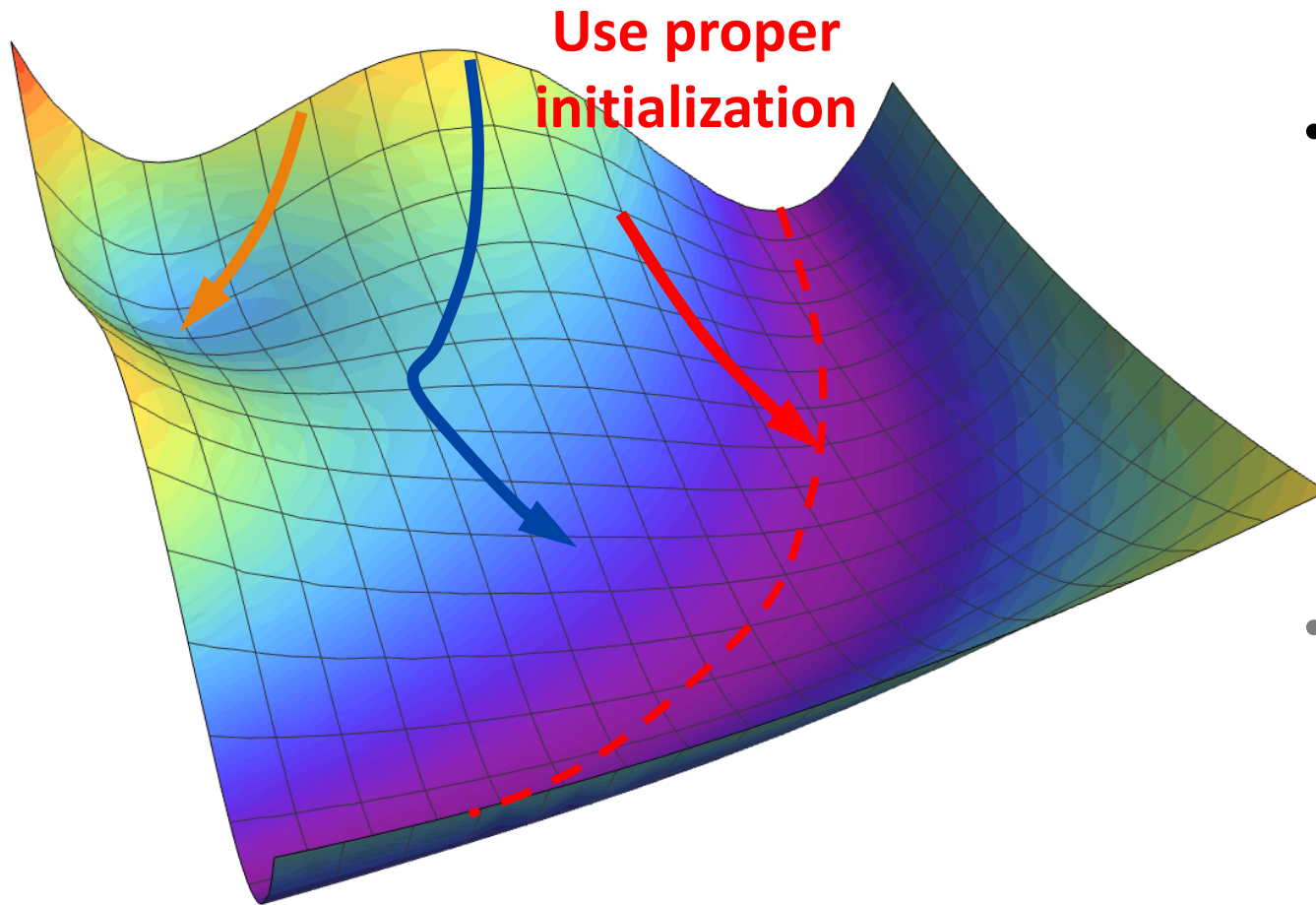
- Highly underdetermined problem, many solutions
- Variants of gradient descent often find those with good generalization
- Question: What is the effect of **overparameterization** on the learning dynamics of optimization algorithms?

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

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]

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)

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

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.

• **Restrictive assumptions on initialization**
D Stöger and M Sussangkarn. 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 in training 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

- Find solution that obtains

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

 **overparametrization**

- Gradient flow on multi-layer linear model:

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

$$\dot{W}_l = -\partial \mathcal{L} / \partial W_l$$

- Examples:

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

Problem Setting: Overparametrized Linear Model

- Multi-layer linear model(network):

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

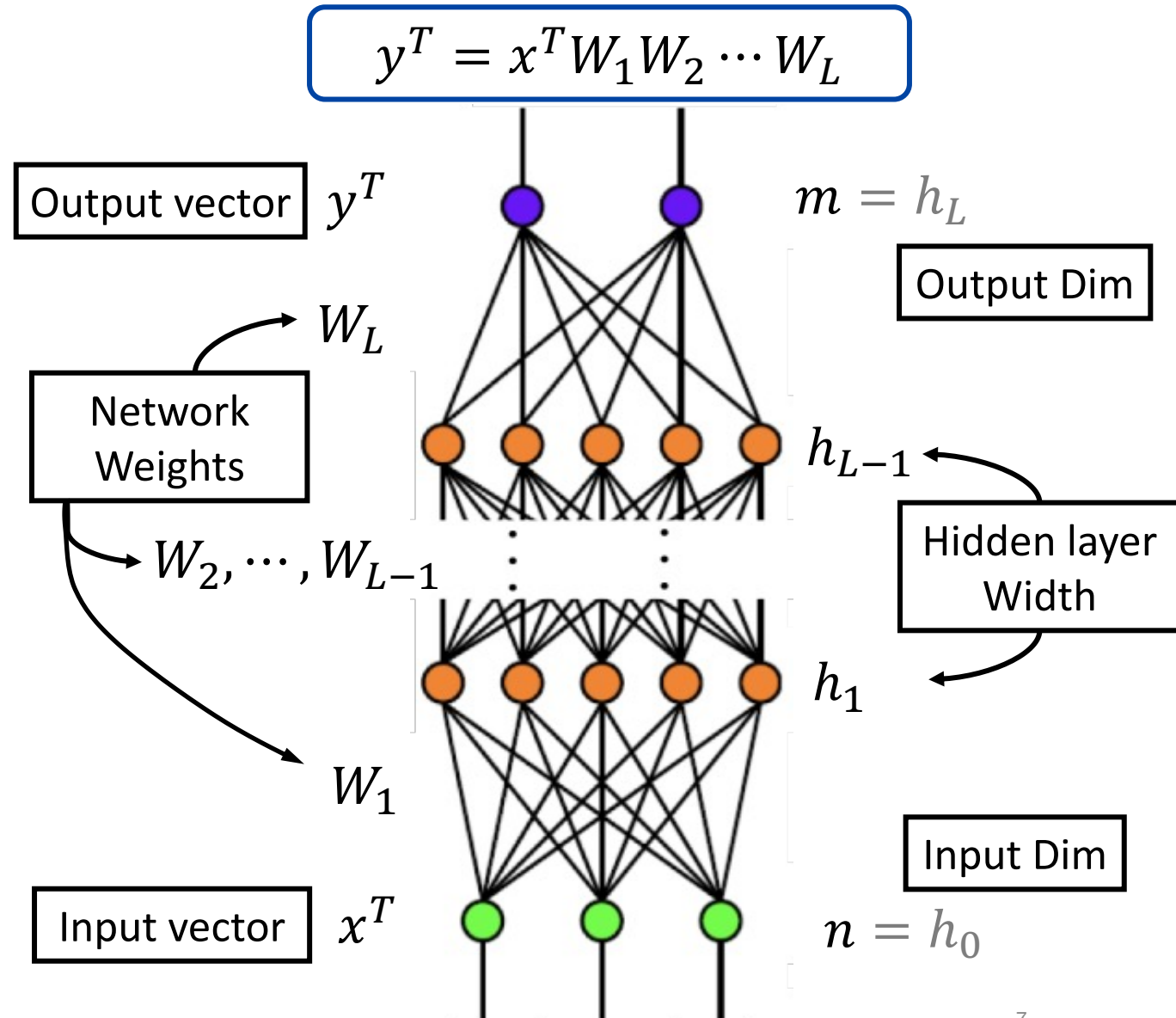
- Overparametrized:

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

$$h_0 = n, h_L = m$$

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

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



Problem Setting: Assumptions

- Find solution that obtains

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

- **Assumptions:** Objective f

- satisfies Polyak-Łojasiewicz(PL)-inequality:

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

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

Problem Setting: Assumptions

- Find solution that obtains

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

- **Assumptions:** Objective f

- satisfies Polyak-Łojasiewicz(PL)-inequality:

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

Examples

Matrix factorization: $f(W) = \|Y - W\|_F^2/2$

Linear regression: $f(W) = \|Y - XW\|_F^2/2$

Matrix sensing: $f(W) = \sum_{i=1}^N (y_i - \langle A_i, W \rangle)^2$

Extends to classification losses
assuming general
Łojasiewicz inequality

Convergence with PL-inequality

Non-overparametrized

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

- Global PL-Inequality

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



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

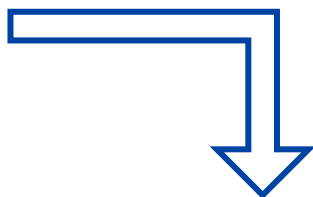
Convergence with PL-inequality

Non-overparametrized

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

- 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)

$$\rightarrow (f(W(t)) - f^*) \leq \exp(-\gamma t) (f(W(0)) - f^*)$$

$f(W(t))$ converges to f^* exponentially

- Rate: PL-Constant γ

Grönwall's inequality

$$\dot{x}(t) \leq -\gamma x(t)$$

$$\Rightarrow x(t) \leq \exp(-\gamma t) x(0)$$

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

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

- 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

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

- Simplest overparameterization $w \rightarrow uv$

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

Scalar Dynamics: Imbalance

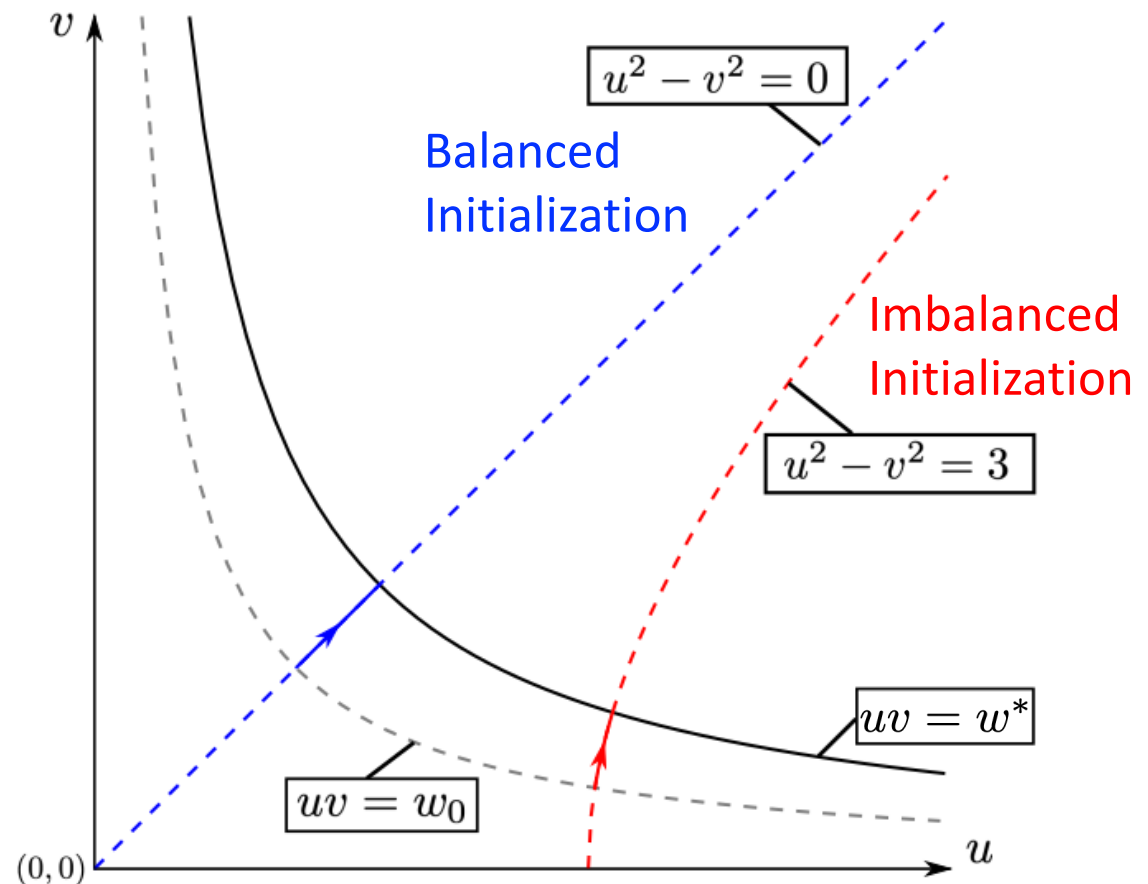
- Gradient flow induces **conservation law**

$$\begin{aligned} \dot{u} &= -f'(uv)v \\ \dot{v} &= -f'(uv)u \end{aligned} \quad \Rightarrow \quad d := u^2 - v^2, \quad \dot{d} \equiv 0$$

- **imbalance** $d := u^2 - v^2$ is time-invariant
- $\mathcal{L}(u, v) = |y - uv|^2$
- Conservation law arises due to scaling **symmetry**

$$u \rightarrow su, \quad v \rightarrow \frac{v}{s}$$

(Noether's Theorem connects symmetry to conservation law)



Scalar Dynamics: Weight-dependent PL inequality

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

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

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



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 μ -strongly convex, and K -smooth
- Loss \mathcal{L} is non-increasing

imbalance d is time invariant
 $|d(t)| = |d(0)|$
 $:= \textit{Imbalance}$

A lower bound on product uv

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

$:= \textit{Margin}$

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

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**

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



Initialization-dependent PL-inequality → Exponential Convergence

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

(Grönwall)

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

Control **imbalance** and **product** by initialization

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



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

$$\left[|w^*| - \sqrt{K/\mu} |w^* - u(0)v(0)| \right]_+$$

$$\left[\sigma_{\min}(W^*) - \sqrt{K/\mu} \|W^* - W(0)\|_F \right]_+$$

Local PL-ineq

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

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

Weight to imbalance and product

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

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

Control imbalance and product by initialization

Exponential Convergence:

$$\text{Rate} \geq \gamma \alpha(\text{Imbalance}, \text{Margin})$$

Imbalance

Warm-up Example: $f(uv)$

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

General Case: $f(W_1 W_2 \cdots W_L)$

$$\{D_l := W_l^T W_l - W_{l+1} W_{l+1}^T\}_{l=1}^{L-1}$$

Margin

$$\left[|w^*| - \sqrt{K/\mu} |w^* - u(0)v(0)| \right]_+$$

$$\left[\sigma_{\min}(W^*) - \sqrt{K/\mu} \|W^* - W(0)\|_F \right]_+$$

Local PL-ineq

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

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

Weight to imbalance and product

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

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

Control imbalance and product by initialization

Exponential Convergence:

$$\text{Rate} \geq \gamma \alpha(\text{Imbalance}, \text{Margin})$$

General Convergence Analysis: Imbalance

- $\mathcal{L}(W_1, \dots, W_L) = f(W_1 W_2 \dots W_L)$
- Gradient Flow: $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$
- Imbalance matrices D_1, D_2, \dots, D_{L-1}

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

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

Conservation law

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

General Convergence Analysis: Imbalance

- $\mathcal{L}(W_1, \dots, W_L) = f(W_1 W_2 \cdots W_L)$
- Gradient Flow: $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$
- Imbalance matrices D_1, D_2, \dots, D_{L-1}

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

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

General Convergence Analysis: Imbalance

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

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

$$(h_{L-2} \times h_{L-1}) \quad (h_{L-1} \times m)$$

$$W = W_1 \cdot W_2 \cdot W_3 \cdots W_{L-1} \cdot W_L$$

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

General Convergence Analysis: Imbalance

- $\mathcal{L}(W_1, \dots, W_L) = f(W_1 W_2 \cdots W_L)$
- Gradient Flow: $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$
- Imbalance matrices $\{D_l := W_l^T W_l - W_{l+1} W_{l+1}^T\}_{l=1}^{L-1}$
- **Imbalance matrices are time-invariant under GF**
$$\dot{D}_l \equiv 0, \quad l = 1, \dots, L - 1$$

Warm-up Example: $f(uv)$

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

General Case: $f(W_1 W_2 \cdots W_L)$

$$\{D_l := W_l^T W_l - W_{l+1} W_{l+1}^T\}_{l=1}^{L-1}$$

Imbalance

Margin

$$\left[|w^*| - \sqrt{K/\mu} |w^* - u(0)v(0)| \right]_+$$

$$\left[\sigma_{\min}(W^*) - \sqrt{K/\mu} \|W^* - W(0)\|_F \right]_+$$

Local PL-ineq

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

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

Weight to imbalance and product

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

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

Control imbalance and product by initialization

Exponential Convergence:

$$\text{Rate} \geq \gamma \alpha(\text{Imbalance}, \text{Margin})$$

General Convergence Analysis: PL-inequality

- $\mathcal{L}(W_1, \dots, W_L) = f(W_1 W_2 \cdots W_L)$
- Gradient Flow: $\dot{W}_l = -\partial \mathcal{L} / \partial W_l$
- $\|\nabla \mathcal{L}(\{W_l\}_{l=1}^L)\|_F^2 = \langle \mathcal{J}_{\{W_l\}_{l=1}^L} \nabla f(W), \nabla f(W) \rangle_F$

Recall warm-up example:

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

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

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

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

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

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

General Convergence Analysis: PL-inequality

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

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

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

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

Min-max theorem

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

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

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

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

$$\left[|w^*| - \sqrt{K/\mu} |w^* - u(0)v(0)| \right]_+$$

$$\left[\sigma_{\min}(W^*) - \sqrt{K/\mu} \|W^* - W(0)\|_F \right]_+$$

Local PL-ineq

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

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

Weight to imbalance
and product

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

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

Control imbalance and product by initialization

Exponential Convergence:

$$\text{Rate} \geq \gamma \alpha(\text{Imbalance}, \text{Margin})$$

Lower Bound on Convergence Rate: Summary

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

Lower Bound on Convergence Rate: Summary

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

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

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

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

Outline

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

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

- Conclusion

Lower Bound on Convergence Rate: Overview

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

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

- The (arguably) optimal bound is given by

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

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

Lower Bound on Convergence Rate

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

Multi-layer Scalar Networks

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

- Multi-layer scalar networks

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

- Imbalance

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

- 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**

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

- Ideally, we want to solve

$$\begin{aligned} \min_{\{w_l\}_{l=1}^L} \lambda_{\min} \left(\mathcal{J}_{\{w_l\}_{l=1}^L} \right) &= \sum_{l=1}^L \prod_{i \neq l} w_i^2 = \sum_{l=1}^L \frac{w^2}{w_l^2} \\ \text{s. t. } w_l^2 - w_{l+1}^2 &= d_l, \quad l = 1, \dots, L-1 \\ w_1 w_2 \cdots w_L &= w \end{aligned}$$

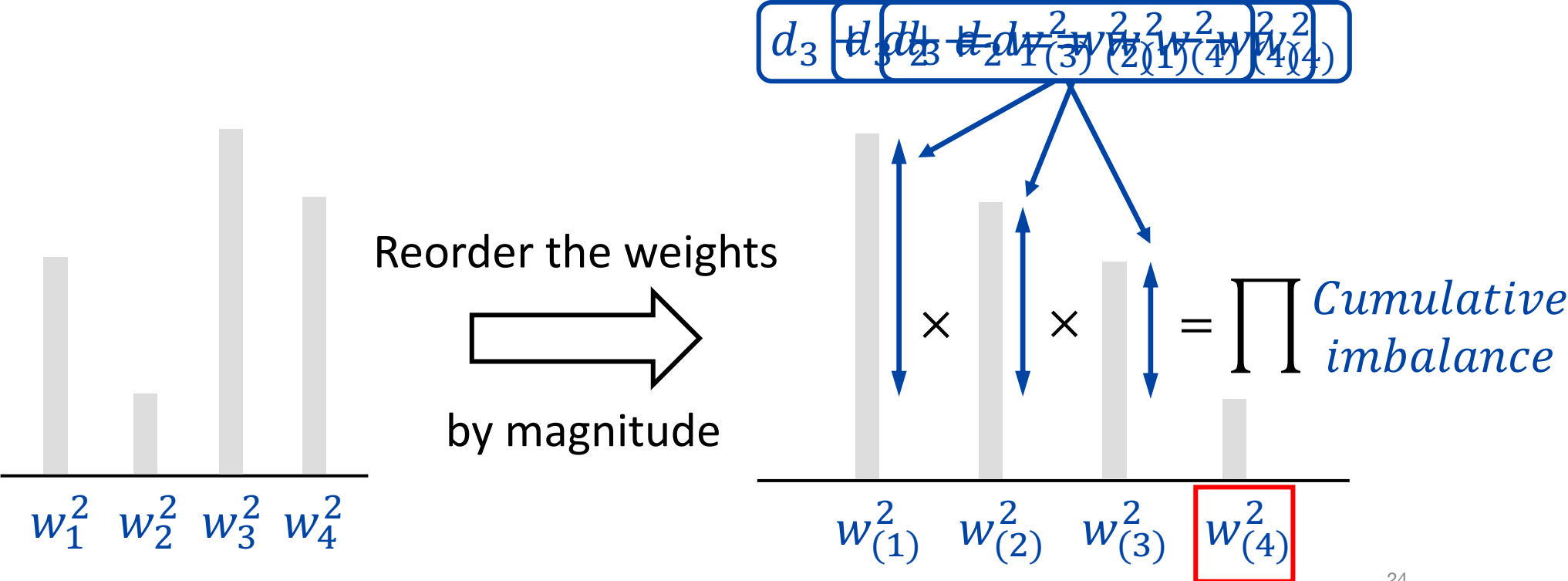
- Only isolated points in the feasible set
- **All feasible points** have the **same** objective value $\alpha^* \left(\{d_l\}_{l=1}^{L-1}, w \right)$
- $\alpha^* \left(\{d_l\}_{l=1}^{L-1}, w \right)$ has **no closed-form expression** in general
(need to solve an L -th order polynomial)

Multi-layer Scalar Networks: Rate Bound

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

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

- Cumulative imbalance?

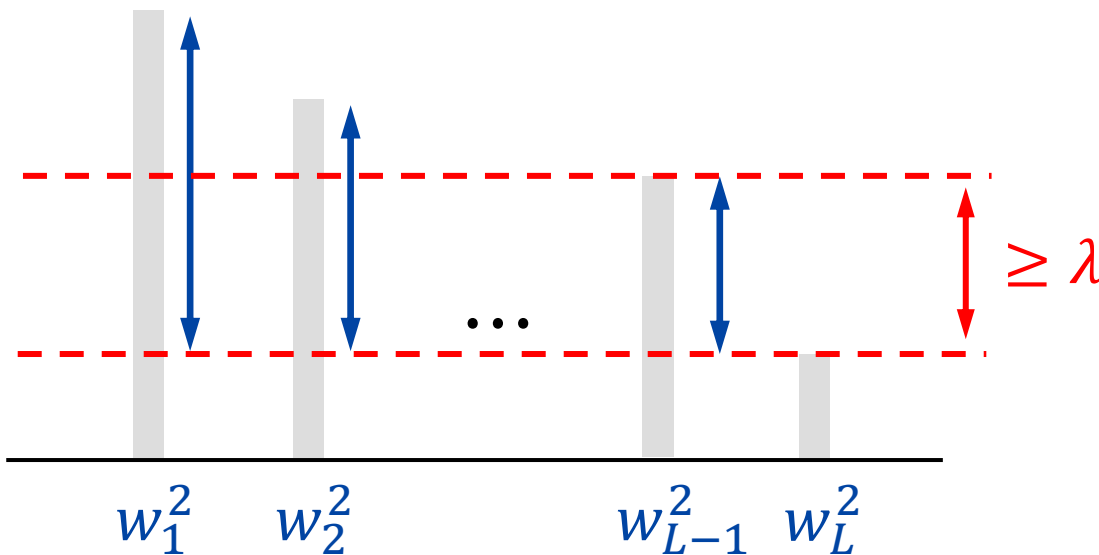


Multi-layer Scalar Networks: Effect of Imbalance

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

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

- Effect of **imbalance** in **deep** networks



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

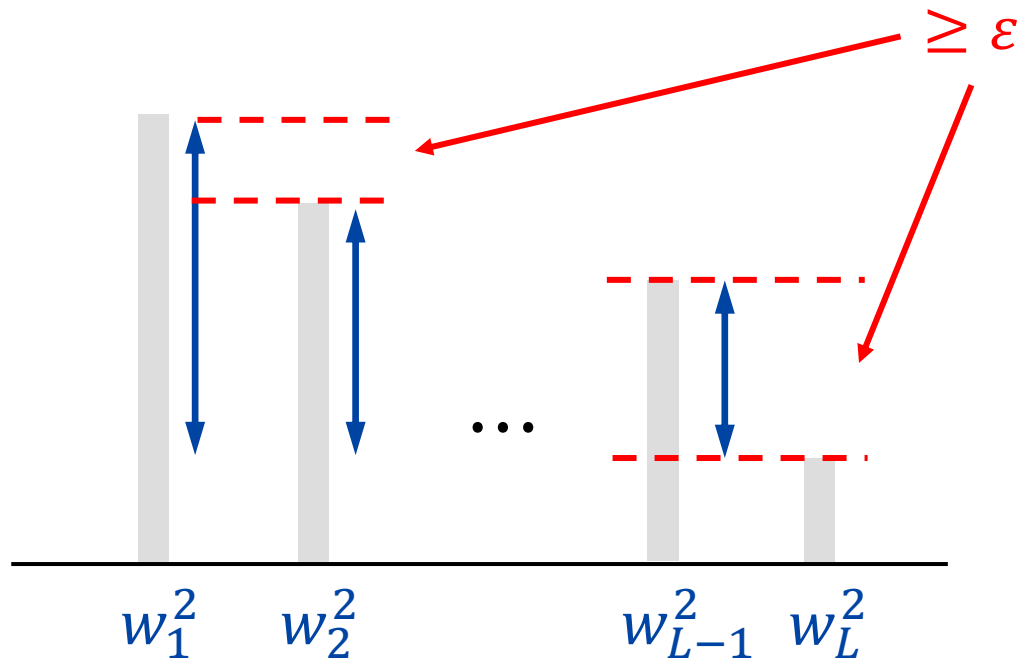
[Yun et al. 20] derived rate bound λ^{L-1} (w.o. product)

Multi-layer Scalar Networks: Effect of Imbalance

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

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

- Effect of imbalance in deep networks



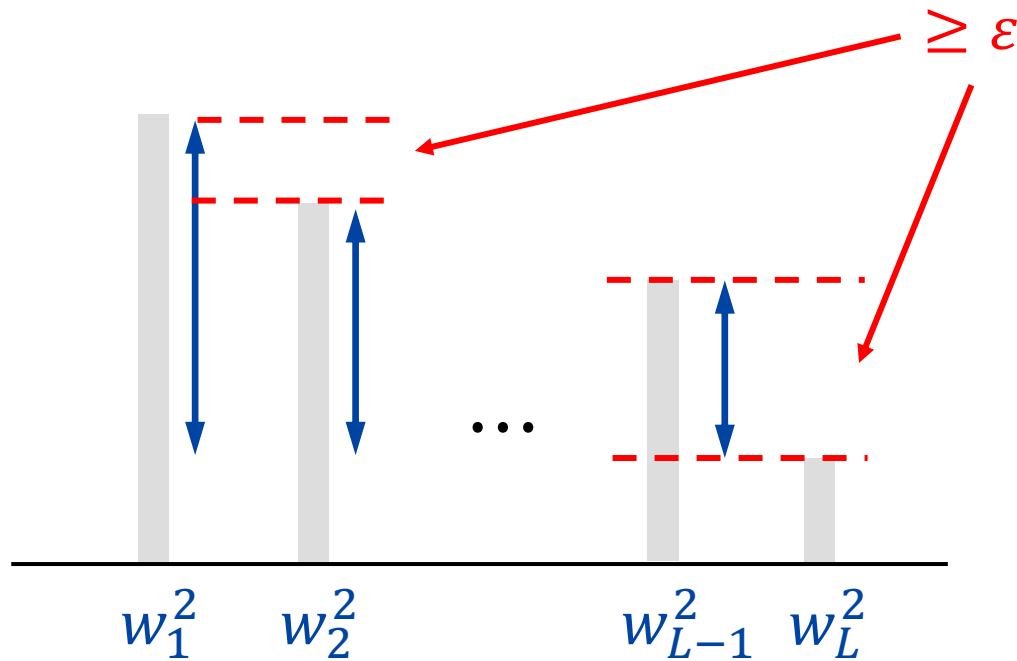
$$\sqrt{(\epsilon^{L-1}(L-1)!)^2 + (Lw^{2-2/L})^2}$$

Multi-layer Scalar Networks: Effect of Imbalance

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

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

- Effect of **imbalance** in **deep** networks



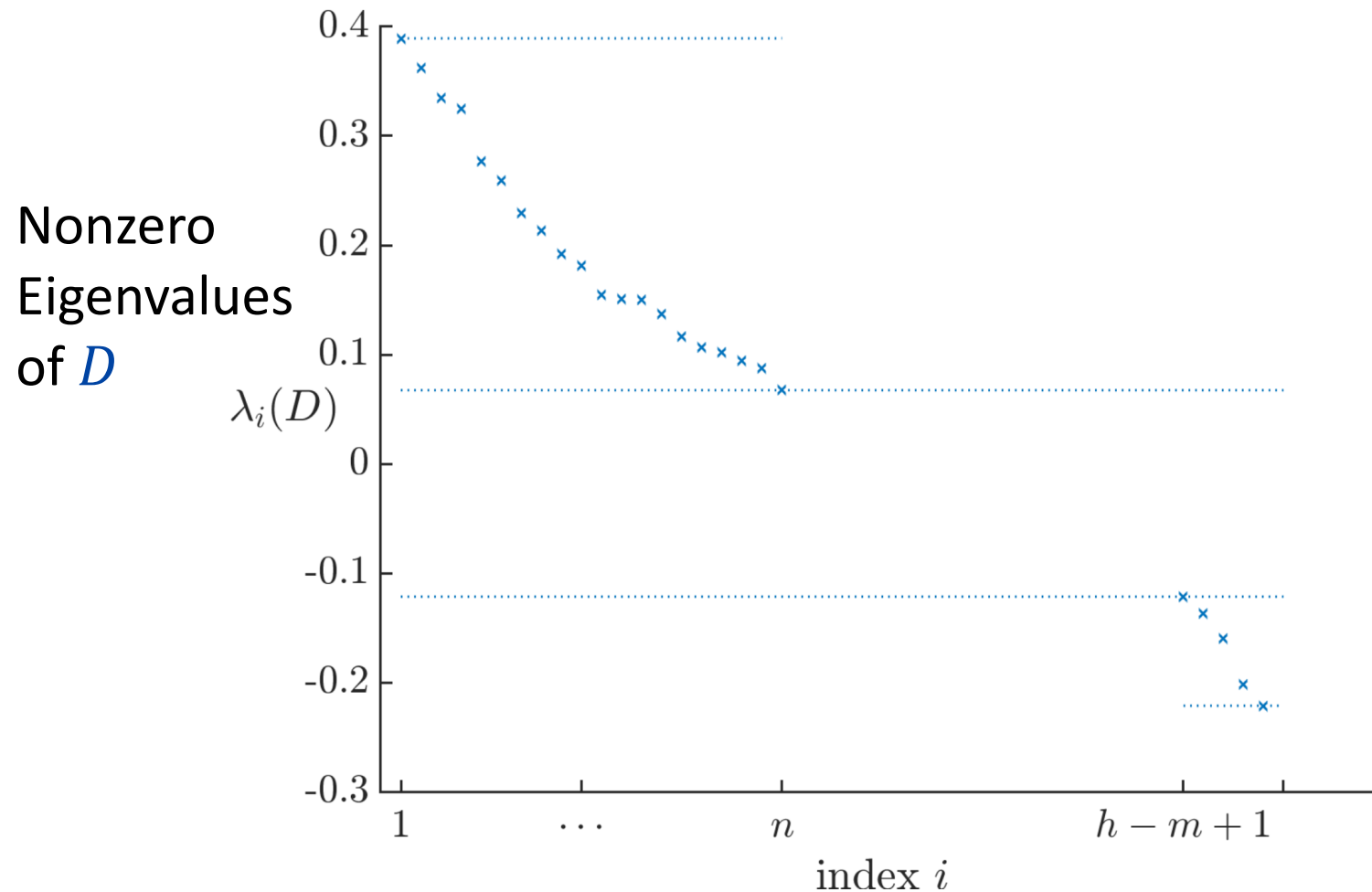
- Imbalanced initialization could **accelerate convergence** significantly for deep networks (For gradient flow!)
- In practice, this is more related to **exploding gradient**

Lower Bound on Convergence Rate

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

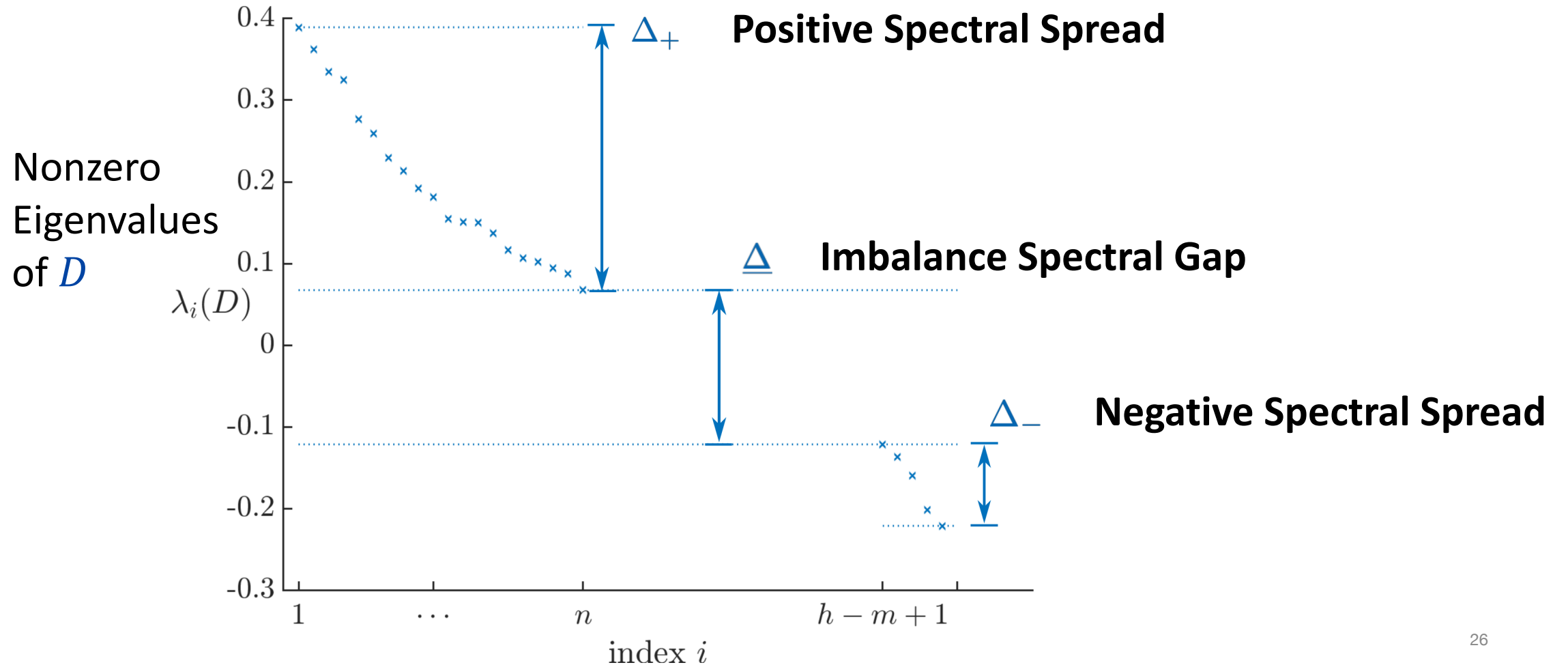
Imbalance quantities

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



Imbalance quantities

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



Two-layer Linear Networks

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

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

and

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

- Ideally, we want to find

$$\alpha^*(D_1, W) = \min_{\{W_1, W_2\}} \lambda_{\min}(\mathcal{J}_{\{W_1, W_2\}}) = \lambda_{\min}(W_1 W_1^T) + \lambda_{\min}(W_2^T W_2)$$
$$s. t. \quad W_1^T W_1 - W_2 W_2^T = D_1$$
$$W_1 W_2 = W$$

Two-layer Linear Networks: Rate Bound

Proposition 2.

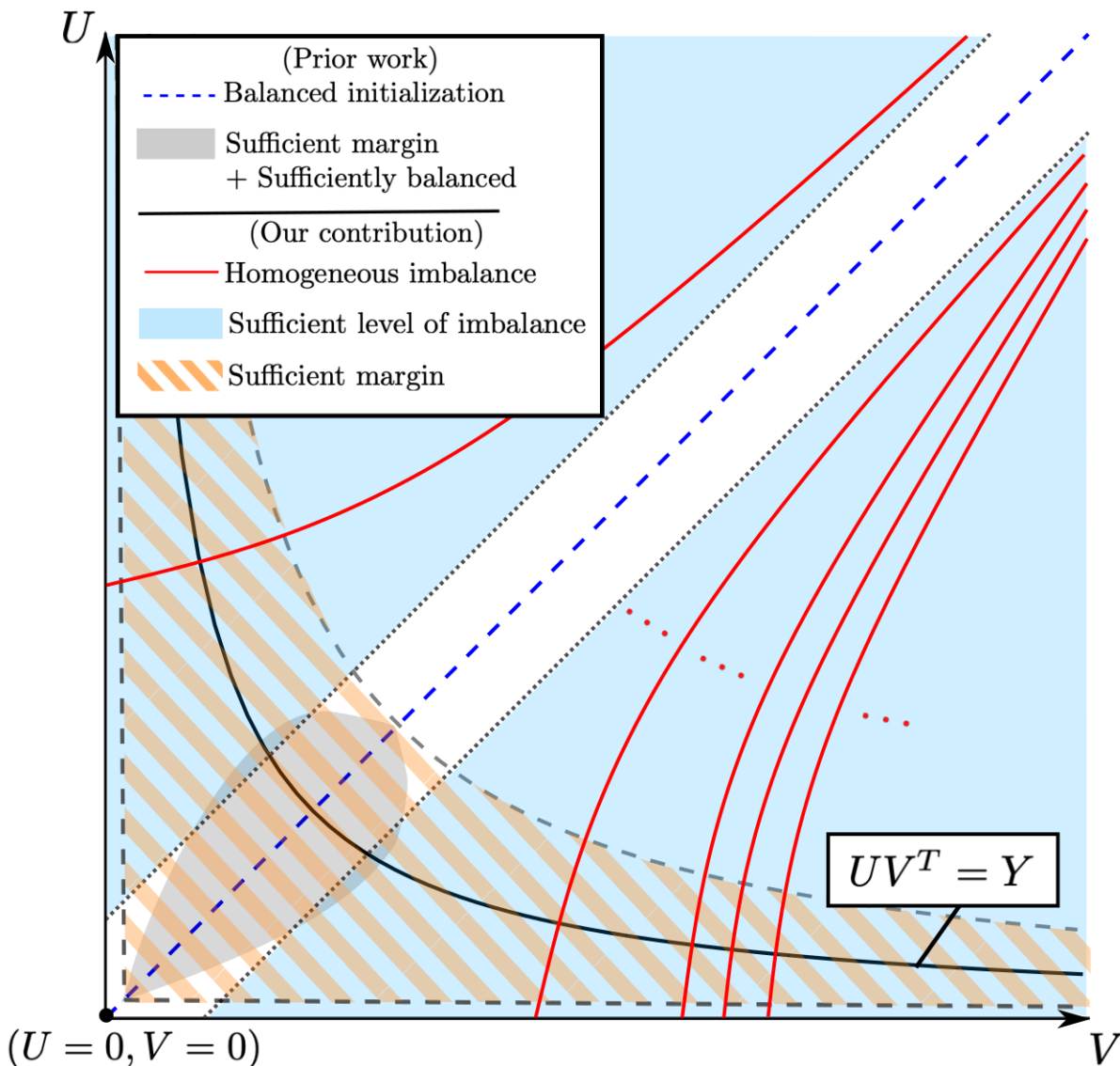
$$\alpha^*(D_1, W) \geq$$

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

Equality holds when $n \neq m$

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

Two-layer Linear Networks: Summary



Initialization for the gradient flow on

$$\frac{1}{2} \|Y - UV^T\|_F^2$$

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

Lower Bound on Convergence Rate

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

Three-layer Linear Model

- $\mathcal{L} = f(W_1 W_2 W_3)$. Find a lower bound on $\lambda_{\min}(\mathcal{J}_{\{W_1, W_2, W_3\}})$

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

- When $w = W_1 W_2 W_3$ is a **scalar**, so is $\mathcal{J}_{\{W_1, W_2, W_3\}}$

$$\lambda_{\min}(\mathcal{J}_{\{W_1, W_2, W_3\}}) = (W_1 W_2 W_2^T W_1^T) + (W_1 W_1^T) (W_3^T W_3) + (W_3^T W_2^T W_2 W_3)$$

- When $W = W_1 W_2 W_3$ is a **matrix**,

- **no closed-form expression** for $\lambda_{\min}(\mathcal{J}_{\{W_1, W_2, W_3\}})$

- $\lambda_{\min}(\mathcal{J}_{\{W_1, W_2, W_3\}})$
 $\geq \lambda_{\min}(W_1 W_2 W_2^T W_1^T) + \lambda_{\min}(W_1 W_1^T) \lambda_{\min}(W_3^T W_3) + \lambda_{\min}(W_3^T W_2^T W_2 W_3)$

Three-layer Linear Model: Formulation

- Our result: Find a lower bound on $\lambda_{\min}(\mathcal{J}_{\{W_1, W_2, W_3\}})$ **considering only the effect of imbalance**
- Ideally, we may want to solve

$$\begin{aligned} & \min_{W_1, W_2, W_3} \lambda_{\min}(\mathcal{J}_{\{W_1, W_2, W_3\}}) \\ & s. t. \quad W_1^T W_1 - W_2 W_2^T = D_1 \\ & \quad \quad W_2^T W_2 - W_3 W_3^T = D_2 \end{aligned}$$

Technical difficulty: **no closed-form expression** for $\lambda_{\min}(\mathcal{J}_{\{W_1, W_2, W_3\}})$

Three-layer Linear Model: Rate Bound

- Our result: Find a lower bound on $\lambda_{\min}(\mathcal{J}_{\{W_1, W_2, W_3\}})$ **considering only the effect of imbalance**

- Instead, we solve

$$\begin{aligned} \min_{W_1, W_2, W_3} & \lambda_{\min}(W_1 W_2 W_2^T W_1^T) + \lambda_{\min}(W_1 W_1^T) \lambda_{\min}(W_3^T W_3) + \lambda_{\min}(W_3^T W_2^T W_2 W_3) \\ \text{s. t.} & W_1^T W_1 - W_2 W_2^T = D_1 \\ & W_2^T W_2 - W_3 W_3^T = D_2 \end{aligned}$$

- The optimal value is

$$\alpha(D_1, D_2) = \frac{1}{2} \left(\Delta_{21}^{(2)} + \Delta_{21} \right) + \Delta_{21} \cdot \Delta_{23} + \frac{1}{2} \left(\Delta_{23}^{(2)} + \Delta_{23} \right)$$

- $\Delta_{21}, \Delta_{21}^{(2)}, \Delta_{23}, \Delta_{23}^{(2)}$ have complicated expressions

Three-layer Linear Model: Rate Bound

- Our result: Find a lower bound on $\lambda_{\min}(\mathcal{J}_{\{W_1, W_2, W_3\}})$ **considering only the effect of imbalance**

- Instead, we solve

$$\begin{aligned} \min_{W_1, W_2, W_3} & \lambda_{\min}(W_1 W_2 W_2^T W_1^T) + \lambda_{\min}(W_1 W_1^T) \lambda_{\min}(W_3^T W_3) + \lambda_{\min}(W_3^T W_2^T W_2 W_3) \\ \text{s. t. } & W_1^T W_1 - W_2 W_2^T = D_1 \\ & W_2^T W_2 - W_3 W_3^T = D_2 \end{aligned}$$

- The optimal value **is obtained by solving 4 optimization problems**

$$\alpha(D_1, D_2) = \frac{1}{2} \left(\Delta_{21}^{(2)} + \Delta_{21} \right) + \Delta_{21} \cdot \Delta_{23} + \frac{1}{2} \left(\Delta_{23}^{(2)} + \Delta_{23} \right)$$

Three-layer Linear Model: Rate Bound

- Our result: Find a lower bound on $\lambda_{\min}(\mathcal{J}_{\{W_1, W_2, W_3\}})$ **considering only the effect of imbalance**

- Instead, we solve

$$\begin{aligned} \min_{W_1, W_2, W_3} \quad & \lambda_{\min}(W_1 W_2 W_2^T W_1^T) + \lambda_{\min}(W_1 W_1^T) \lambda_{\min}(W_3^T W_3) + \lambda_{\min}(W_3^T W_2^T W_2 W_3) \\ \text{s. t.} \quad & W_1^T W_1 - W_2 W_2^T = D_1 \\ & W_2^T W_2 - W_3 W_3^T = D_2 \end{aligned}$$

- The optimal value **is obtained by solving 4 optimization problems**

$$\alpha(D_1, D_2) = \frac{1}{2} \left(\Delta_{21}^{(2)} + \Delta_{21} \right) + \Delta_{21} \cdot \Delta_{23} + \frac{1}{2} \left(\Delta_{23}^{(2)} + \Delta_{23} \right)$$

Three-layer Linear Model: Rate Bound

- Our result: Find a lower bound on $\lambda_{\min}(\mathcal{J}_{\{W_1, W_2, W_3\}})$ **considering only the effect of imbalance**

- Instead, we solve

$$\begin{aligned} \min_{W_1, W_2, W_3} & \lambda_{\min}(W_1 W_2 W_2^T W_1^T) + \lambda_{\min}(W_1 W_1^T) \lambda_{\min}(W_3^T W_3) + \lambda_{\min}(W_3^T W_2^T W_2 W_3) \\ \text{s. t.} & W_1^T W_1 - W_2 W_2^T = D_1 \\ & W_2^T W_2 - W_3 W_3^T = D_2 \end{aligned}$$

- The optimal value **is obtained by solving 4 optimization problems**

$$\alpha(D_1, D_2) = \frac{1}{2} \left(\Delta_{21}^{(2)} + \Delta_{21} \right) + \Delta_{21} \cdot \Delta_{23} + \frac{1}{2} \left(\Delta_{23}^{(2)} + \Delta_{23} \right)$$

Three-layer Linear Model: Rate Bound

- Our result: Find a lower bound on $\lambda_{\min}(\mathcal{J}_{\{W_1, W_2, W_3\}})$ **considering only the effect of imbalance**

- Instead, we solve

$$\begin{aligned} \min_{W_1, W_2, W_3} & \lambda_{\min}(W_1 W_2 W_2^T W_1^T) + \lambda_{\min}(W_1 W_1^T) \lambda_{\min}(W_3^T W_3) + \lambda_{\min}(W_3^T W_2^T W_2 W_3) \\ \text{s. t.} & W_1^T W_1 - W_2 W_2^T = D_1 \\ & W_2^T W_2 - W_3 W_3^T = D_2 \end{aligned}$$

- The optimal value **is obtained by solving 4 optimization problems**

$$\alpha(D_1, D_2) = \frac{1}{2} \left(\Delta_{21}^{(2)} + \Delta_{21} \right) + \Delta_{21} \cdot \Delta_{23} + \frac{1}{2} \left(\Delta_{23}^{(2)} + \Delta_{23} \right)$$

- Solving these optimization problems relies on **novel eigenvalue bound** based on generalized eigenvalue interlacing relation

Three-layer Linear Model: Cumulative imbalance

- Our bound:

$$\alpha(D_1, D_2) = \frac{1}{2} \left(\Delta_{21}^{(2)} + \Delta_{21} \right) + \Delta_{21} \cdot \Delta_{23} + \frac{1}{2} \left(\Delta_{23}^{(2)} + \Delta_{23} \right)$$

Three-layer Linear Model: Cumulative imbalance

- Consider the case

$$D_1 = W_1^T W_1 - W_2 W_2^T \succcurlyeq 0 \quad \Rightarrow \quad W_1^T W_1 \succcurlyeq W_2 W_2^T$$

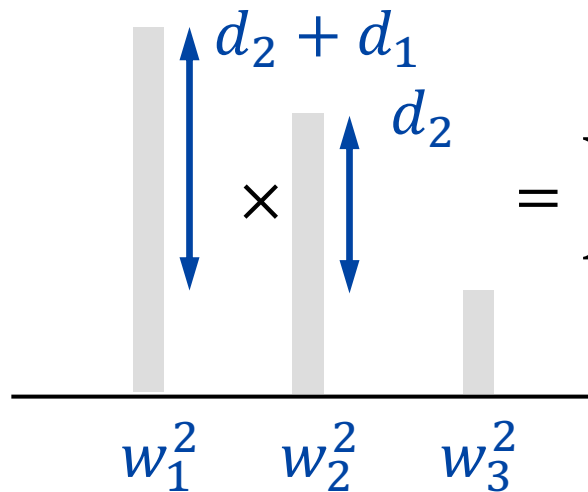
$$D_2 = W_2^T W_2 - W_3 W_3^T \succcurlyeq 0 \quad \Rightarrow \quad W_2^T W_2 \succcurlyeq W_3 W_3^T$$

- Our bound (**Simplified Expression**):

$$\alpha(D_1, D_2) = \boxed{\frac{1}{2} (\Delta_{21}^{(2)} + \Delta_{21})} + \boxed{\Delta_{21} \cdot \Delta_{23}} + \boxed{\frac{1}{2} (\Delta_{23}^{(2)} + \Delta_{23})}$$

$= 0$
 $= 0$

Recall: multi-layer scalar weights



$$\left(\Delta_{21}^{(2)} + \Delta_{21} \right) / 2 = \lambda_n(D_2) [\lambda_n(D_1) + \lambda_n(D_2)]$$

$$= \prod \text{Cumulative imbalance} = d_2(d_1 + d_2)$$

Three-layer Linear Model: Cumulative imbalance

- Our bound:

$$\begin{aligned}\alpha(D_1, D_2) &= \frac{1}{2} \left(\Delta_{21}^{(2)} + \Delta_{21} \right) + \Delta_{21} \cdot \Delta_{23} + \frac{1}{2} \left(\Delta_{23}^{(2)} + \Delta_{23} \right) \\ &\approx \sum \prod \text{Cumulative} \\ &\quad \text{imbalance}\end{aligned}$$

- Each $\left(\prod \text{Cumulative imbalance} \right)$ is associated with some **positive semi-definite ordering on the weights**
- We conjecture that similar bound can be derived for deep networks $L > 3$.

General Convergence Analysis: Rate Bound

- Local PL-inequality: $\|\nabla\mathcal{L}(\{W_l\}_{l=1}^L)\|_F^2 \geq \gamma \lambda_{\min}(\mathcal{J}_{\{W_l\}_{l=1}^L}) (\mathcal{L} - \mathcal{L}^*)$

- Given **initialization** $\{W_l(0)\}_{l=1}^L$, find a lower bound on $\lambda_{\min}(\mathcal{J}_{\{W_l(t)\}_{l=1}^L})$

Our bound: $\lambda_{\min}(\mathcal{J}_{\{W_l\}_{l=1}^L}) \geq \alpha(\{D_l\}_{l=1}^{L-1}, \sigma_{\min}(W))$

α increasing function of $\sigma_{\min}(W)$

Consider the effect of imbalance only

$$\lambda_{\min}(\mathcal{J}_{\{W_l\}_{l=1}^L}) \geq \alpha(\{D_l\}_{l=1}^{L-1}, \mathbf{0})$$

Consider balanced initialization

$$\begin{aligned} \lambda_{\min}(\mathcal{J}_{\{W_l\}_{l=1}^L}) &\geq \alpha(\mathbf{0}, \sigma_{\min}(W)) \\ &\geq \alpha(\mathbf{0}, \text{margin}) \end{aligned}$$

(use directly)

$$\lambda_{\min}(\mathcal{J}_{\{W_l\}_{l=1}^L}) \geq \alpha(\{D_l\}_{l=1}^{L-1}, \text{margin})$$

Exponential Convergence

“trajectory” PL-inequality: $\|\nabla\mathcal{L}(\{W_l\}_{l=1}^L)\|_F^2 \geq \gamma \alpha(\{D_l\}_{l=1}^{L-1}, \text{margin}) (\mathcal{L} - \mathcal{L}^*)$

General Convergence Analysis: Comparison

| # of layers | $\alpha(\{D_l\}_{l=1}^{L-1}, 0)$ | $\alpha(0, \textit{margin})$ | $\alpha(\{D_l\}_{l=1}^{L-1}, \textit{margin})$ |
|--------------------------------|---------------------------------------|--------------------------------------|--|
| $(L \geq 2)$ Scalar weights | [Yun et al. 20]* Our work | [Saxes et al. 14] Our work | [Tarmoun et al. 21]* Our work |
| $(L = 2)$ Matrix weights | Our work | [Arora et al. 18] Our work | [Tarmoun et al. 21]* Our work |
| $(L > 2)$ Matrix weights | [Yun et al. 20]* Our work** | [Arora et al. 18] | None |

*: with additional assumptions

** : three-layer networks

Our work: new bounds derived in our paper

Our work: automatically covered by our bound

Outline

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

Conclusion

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

$$\text{Rate} \geq \gamma \alpha(\text{Imbalance}, \text{Margin})$$

Our analysis also works for classification task with exponential loss

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

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