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

Hancheng Min, René Vidal†, Enrique Mallada**

Electrical and Computer Engineering, Johns Hopkins University, †Center for Innovation in Data Engineering and Science, University of Pennsylvania

INTRODUCTION

Goal: Understand the effect of overparametrization on the convergence of gradient flow for training linear models:

Prior work studied specific initialization:

- NTK [1]: requires extremely large width
- Spectral [2,3], balanced [4]: satisfied by a zero-measure set

Our work studies **general initialization**:

- Propose a **unified** convergence analysis for **deep linear networks** under any initialization shape
- Show that the rate of convergence is determined by **certain properties** of the initialization

EFFECT OF OVERPARAMETERIZATION

Overparametrized model: $W \coloneqq W_1 W_2 \cdots W_L$

Loss: $\mathcal{L}(\{W_l\}_{l=1}^L) = f(W_1 W_2 \cdots W_L)$

Overparametrization ≈ Preconditioning:

Gradient flow: $\dot{W}_l = -\nabla_{W_l}\mathcal{L}, \quad l = 1,2,\cdots,L$

Full gradient **(a)**: $\nabla_{\{W_l\}_{l=1}^L}\mathcal{L} = \tau_{\{W_l\}_{l=1}^L} \cdot \nabla f(W)$ "$\frac{\mathrm{d}W}{\mathrm{d}\{W_l\}_{l=1}^L}$"

Induced flow: $\dot{W} = \mathcal{T}_{\{W_l\}_{l=1}^L} \cdot \nabla f(W)$

$$\dot{\mathcal{L}} = -\left\|\nabla_{\{W_l\}_{l=1}^L}\mathcal{L}\right\|_F^2 = -\left\langle \nabla f, \mathcal{T}_{\{W_l\}_{l=1}^L}\nabla f\right\rangle$$

- $\mathcal{T}_{\{W_l\}_{l=1}^L} \coloneqq \tau^*_{\{W_l\}_{l=1}^L} \circ \tau_{\{W_l\}_{l=1}^L}$ is a p.s.d. linear operator
- The overparametrized flow proceeds as if we are running **gradient flow on f** w.r.t. the product W, with a weight-dependent **preconditioner** $\mathcal{T}_{\{W_l\}_{l=1}^L}$
- NTK analysis: $\mathcal{T}_{\{W_l(t)\}_{l=1}^L} \approx \mathcal{T}_{\{W_l(0)\}_{l=1}^L}$

Outside NTK: $\mathcal{T}_{\{W_l\}_{l=1}^L}$ is time-varying (**Main Challenge**)

UNIFIED CONVERGENCE ANALYSIS

- **(1)** Weight-dependent PL (from **(a)+(b)**):

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

- **(2) Initialization-dependent** lower bound:

$$\alpha^*(\{W_l(0)\}_{l=1}^L) = \min_{\{W_l\}_{l=1}^L} \lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right)$$
$$s.t. \{W_l\}_{l=1}^L \in \boldsymbol{ConstraintSet}(\{W_l(0)\}_{l=1}^L)$$

- **(1) + (2) = Exponential convergence**: $\mathcal{L}(t) - \mathcal{L}^* \leq \exp(-\alpha^*(\{W_l(0)\}_{l=1}^L)\,\gamma\, t)(\mathcal{L}(0)-\mathcal{L}^*)$

Assumptions on f:

- **(b)** PL inequality: $\|\nabla f(W)\|_F^2 \geq \gamma(f(W) - f^*)$
- **(c)** strongly convex and has Lipschitz gradient

CONVERGENCE RATE FOR DEEP SCALAR NETWORKS

Deep scalar networks: $\mathcal{L}(\{w_l\}_{l=1}^L) = \left|y - \Pi_{l=1}^L w_l\right|^2, w_l \in \mathbb{R}$

- **(1):** $\|\nabla\mathcal{L}\|_F^2 \geq \left(\Sigma_{l=1}^L \frac{w^2}{w_l^2}\right)(\mathcal{L}-\mathcal{L}^*)$ $(\gamma = 1)$
- **(2):** $\alpha^* = \min_{\{w_l\}_{l=1}^L} \Sigma_{l=1}^L \frac{w^2}{w_l^2}$

$s.t.$ ***Imbalance constraints***

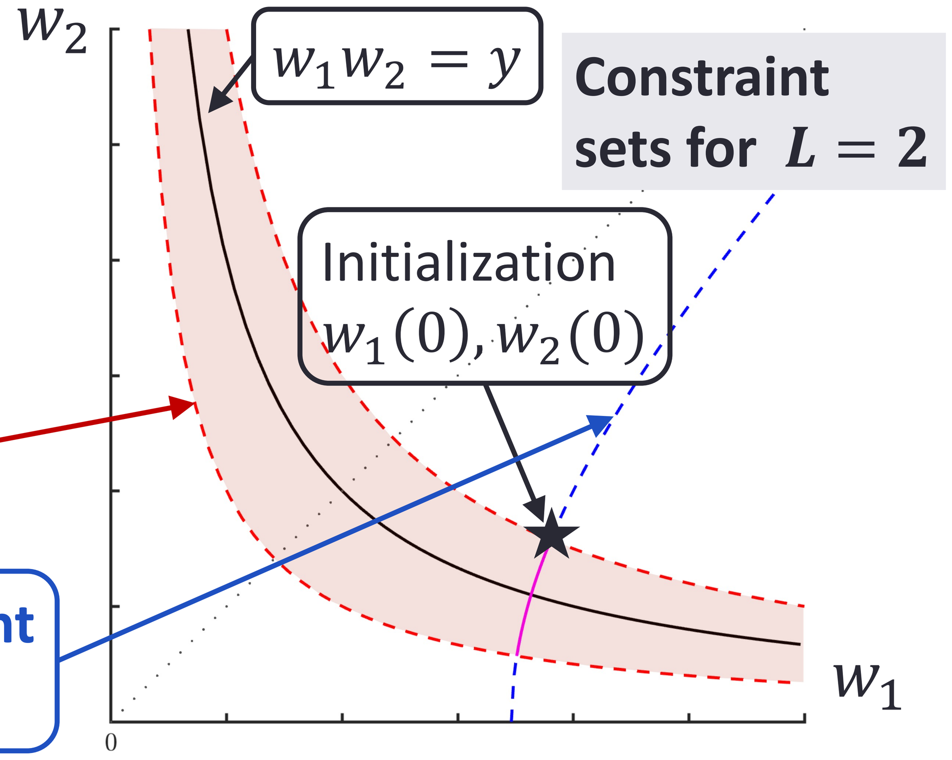
$w_l^2 - w_{l+1}^2 = w_l^2(0) - w_{l-1}^2(0), \quad l = 1,\cdots,L-1$

Margin constraint

$|w| \geq |y| - |y - w(0)| \coloneqq margin$

Loss is non-increasing + **(c)**: $|y - w| \leq |y - w(0)|$

Imbalance is **time-invariant** $d_l = w_l^2 - w_{l+1}^2, \boldsymbol{\dot{d}_l = 0}$



Constraint sets for $L = 2$

Special case $L = 2$:

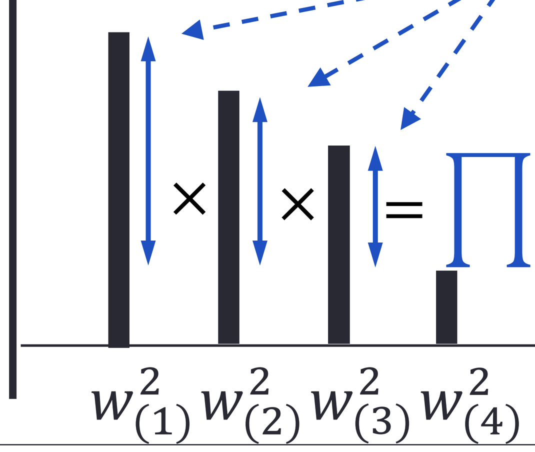
$$\alpha^* = \min_{w_1,w_2} w_1^2 + w_2^2$$
$$s.t.\ w_1^2 - w_2^2 = d$$
$$|w_1 w_2| \geq margin$$

$$\alpha^* = \sqrt{d^2 + 4(margin)^2}$$

$w_1^2 + w_2^2 = \sqrt{(w_1^2 - w_2^2)^2 + 4(w_1w_2)^2}$

General case $L > 2$:

$$\alpha^* \geq \sqrt{\left(\Pi_{l=1}^{L-1} d_{(l)}\right)^2 + (L(margin)^{2-2/L})^2}$$



α^* has **no closed-form** (solution of L-th order poly.)

weights reordered by magnitude

CONVERGENCE RATE FOR GENERAL NETWORKS

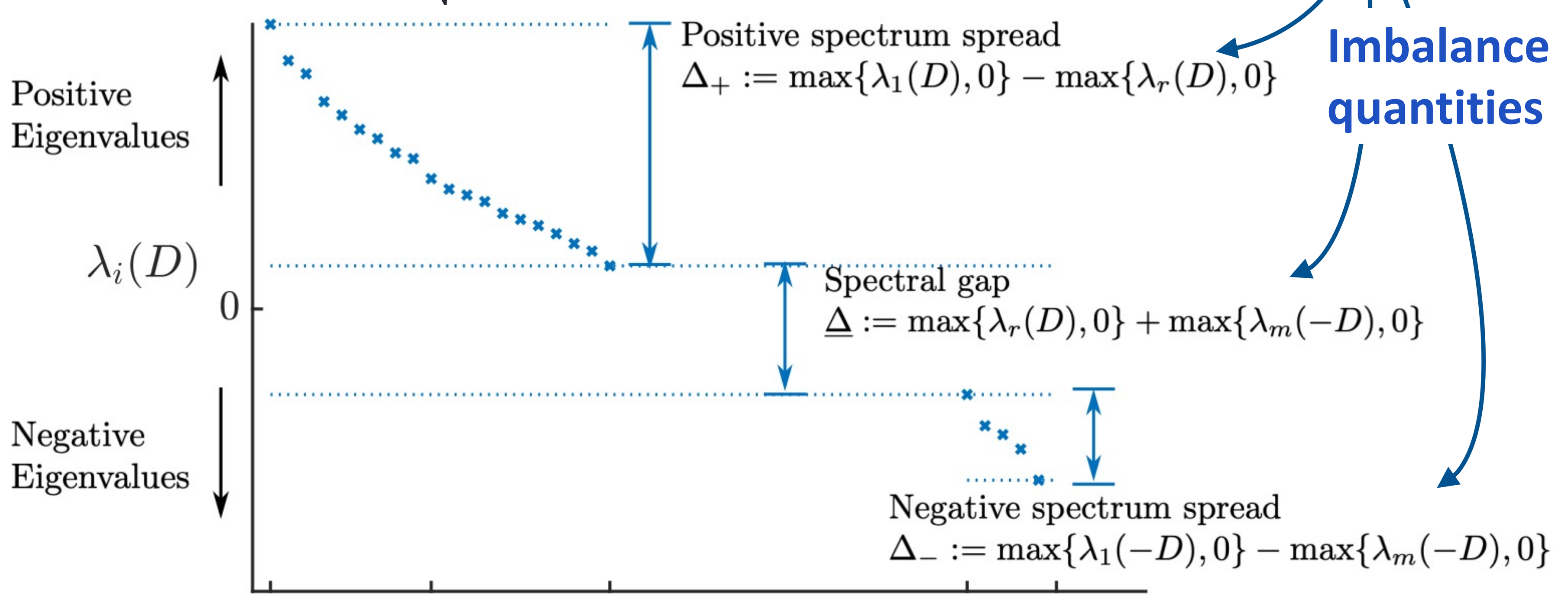
Two-layer networks: $\mathcal{L}(W_1, W_2) = f(W_1 W_2)$

(1): $\|\nabla\mathcal{L}\|_F^2 \geq \left(\lambda_{\min}(W_1W_1^\top) + \lambda_{\min}(W_2^\top W_2)\right)(\mathcal{L}-\mathcal{L}^*)$

(2): $\alpha^* = -\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4(margin)^2}$

$$-\Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4(margin)^2}$$

Imbalance matrix: $D = W_1^\top W_1 - W_2 W_2^\top$



Three-layer networks:

for **general imbalanced initialization**

$$\alpha^* \geq \prod \textit{cumulative imbalance}^{\dagger}$$

†: a complicated expression

Deep networks:

under **homogeneous imbalance** assumption

$$\alpha^* \geq \sqrt{\left(\prod \textit{cumulative imbalance}\right)^2 + (L(margin)^{2-2/L})^2}$$

For certain imbalanced initialization,

$$\prod \textit{cumulative imbalance} = \Theta(L!)$$

- Super-exponential in depth
- Related to exploding gradient

REFERENCES

[1] Jacot, A., Gabriel,F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. NeurIPS, 2018

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

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

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