Early Neuron Alignment in Two-layer ReLU Networks with Small Initialization

Hancheng Min*, Enrique Mallada†, René Vidal*

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

INTRODUCTION

Key result: Two-layer ReLU nets solve binary classification problems by learning features that align with class centers.

Prior work: existing theories are either

- restrictive (# of data, width of network),
- asymptotic (assume infinitely small initialization), or
- heuristics/qualitative (no formal convergence result).

This work: A complete, quantitative, and non-asymptotic convergence analysis for two-layer ReLU networks without restrictions on size of data/network.

PROBLEM SETTING

Problem: Training two-layer ReLU network for binary classification on orthogonally separable data

- Data with two classes:
  $\{x_i, y_i\}_{i=1}^n$: input $x_i \in \mathbb{R}^D$, label $y_i \in \{+1, -1\}$
- Two-layer ReLU Network:
  $$f(x;\theta) = \Sigma_{j=1}^h v_j \mathrm{ReLU}(w_j^\top x), \theta := \{w_j, v_j\}_{j=1}^h$$
- Classification Loss:
  $\mathcal{L}(\theta) = \sum_{i=1}^n \ell(y_i, f(x_i;\theta))$, ℓ is exp or logistic loss
- Gradient flow training: $\dot{\theta} = -\nabla_\theta \mathcal{L}(\theta), \theta(0) = \theta_0$

Assumptions:

- (critical) Small initialization: $\|\theta(0)\|_F = \mathcal{O}(\epsilon)$
- (technical) Balanced initialization: $\|w_j(0)\|_F^2 = v_j^2(0)$
- (critical) μ-orthogonally separable data ($\mu > 0$)
  $$\cos(x_i, x_j) \begin{cases} \geq \mu & , y_i = y_j \\ \leq \mu & , y_i \neq y_j \end{cases}$$

CONVERGENCE OF TWO-LAYER RELU NETWORKS WITH SMALL INITIALIZATION

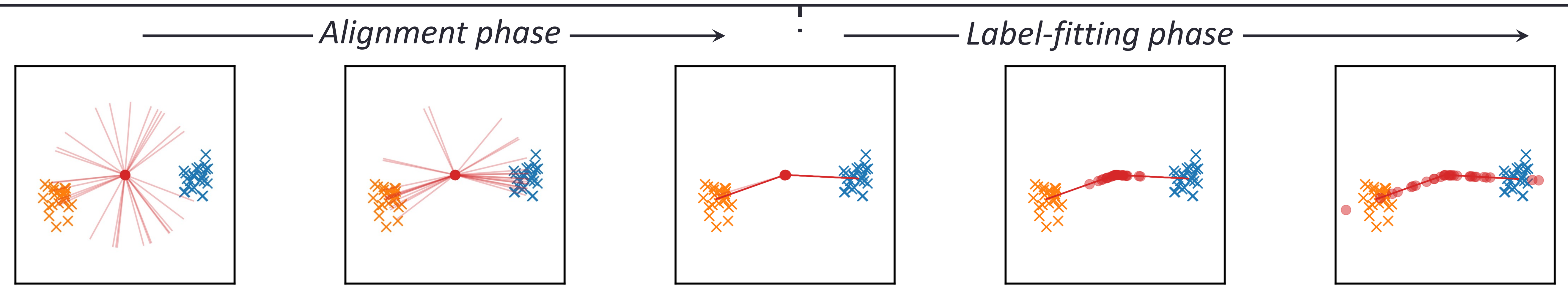
Small initialization leads to two training phases. (1) Neurons align with class centers. (2) Neurons grow their norms to fit the labels.

x: Positive data $\{x_i : y_i = +1\}$

x: Negative data $\{x_i : y_i = -1\}$

●: Neurons $\{w_j\}$

–: Neuron directions $\left\{\frac{w_j}{\|w_j\|}\right\}$



Initialization: Neurons have small norms, and random directions

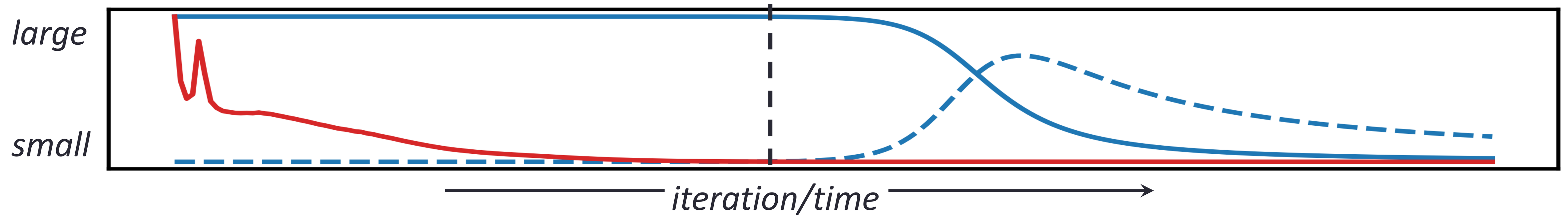
Alignment: Neuron directions align with class centers

Fitting: Neurons grow their norms to minimize loss

– : Loss $\mathcal{L}$

- - : change in norm $\sum_j \frac{d}{dt}\|w_j\|^2$

– : change in direction $\sum_j \left\|\frac{d}{dt}\frac{w_j}{\|w_j\|}\right\|$



Theoretical Results

Complete (from alignment to convergence), **quantitative** (bounds on scale, time, & rate), and **non-asymptotic** (finite init. scale) analysis of GF

With sufficiently small init. scale

$$\epsilon = \mathcal{O}\left(\frac{1}{\sqrt{h}}\exp\left(-\frac{n\log n}{\sqrt{\mu}}\right)\right)$$

Early alignment

"Good alignment with data" requires **at most** $\mathcal{O}\left(\frac{\log n}{\sqrt{\mu}}\right)$ time

Phase transition

Alignment phase ends at $\Theta\left(\frac{1}{n}\log\frac{1}{\sqrt{h}\epsilon}\right)$ time

Convergence rate

During fitting phase, loss converges at $\mathcal{O}\left(\frac{1}{t}\right)$ rate

Low-rank bias

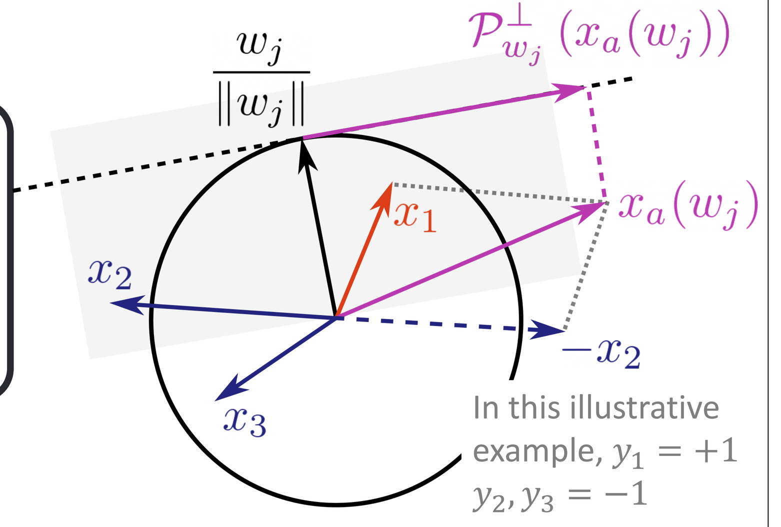
Weight matrix $W = [w_1, \cdots, w_h]$ asymptotically has rank at most 2

EARLY NEURON ALIGNMENT

Key dynamics in alignment phase

Neuron angular dynamics $\quad \mathcal{P}_{w_j}^\perp := (I - w_j w_j^\top / \|w_j\|^2)$

$$\frac{d}{dt}\frac{w_j}{\|w_j\|} = \mathcal{P}_{w_j}^\perp(x_a(w_j)), x_a(w_j) = \sum_{i:\langle x_i, w_j\rangle > 0} x_i y_i$$



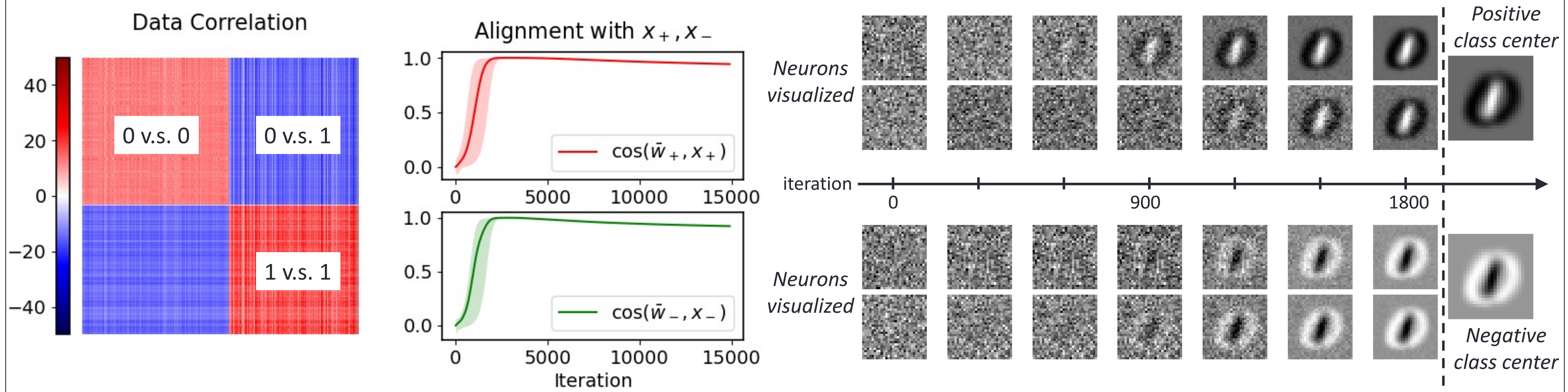
Challenges:

- Non-linear, non-smooth dynamics
- Heavily depends on activation patterns $\{\mathrm{sign}(\langle x_i, w_j\rangle)\}$

Good news:

- Move-to-centroid ($x_a(w_j)$) interpretation under fixed activation pattern
- Tracking "monotone" evolution of activation patterns under orthogonally separable data

NUMERICAL EXPERIMENT



- Images of two MNIST digits (centered by mean image) are approximately orthogonally separable
- Starting from random initialization, all neurons align with either the positive (x_+) or negative class center (x_-)
- Label fitting: refer to our main paper