Neural Collapse under Gradient Flow on Shallow ReLU Networks for Orthogonally Separable Data

Hancheng Min — *INS&SMS, SJTU*
Zhihui Zhu — *CSE, OSU*
René Vidal — *IDEAS, UPenn*

NEURAL INFORMATION PROCESSING SYSTEMS

INTRODUCTION

- **Neural Collapse** (NC) is a phenomenon where last-layer features and classifiers exhibit a highly structured, symmetric pattern
- Prior work on the **theory of NC** focuses on the **unconstrained feature model**: dynamics are simplified by treating all feature layers as one
- **We prove the emergence of NC for ReLU nets**:
  - With input data, only directional collapse is achieved, instead of singleton collapse
  - Gradient flow with small initialization provably converges to NC solution

PROBLEM

- **Orthogonally separable data from K classes:**

$$\langle x_i, x_j\rangle > 0 \quad \text{if} \quad y_i = y_j \quad \text{where } x_i \in \mathbb{R}^D,$$
$$\langle x_i, x_j\rangle < 0 \quad \text{if} \quad y_i \neq y_j \quad y_i \text{ 1-hot vector}$$

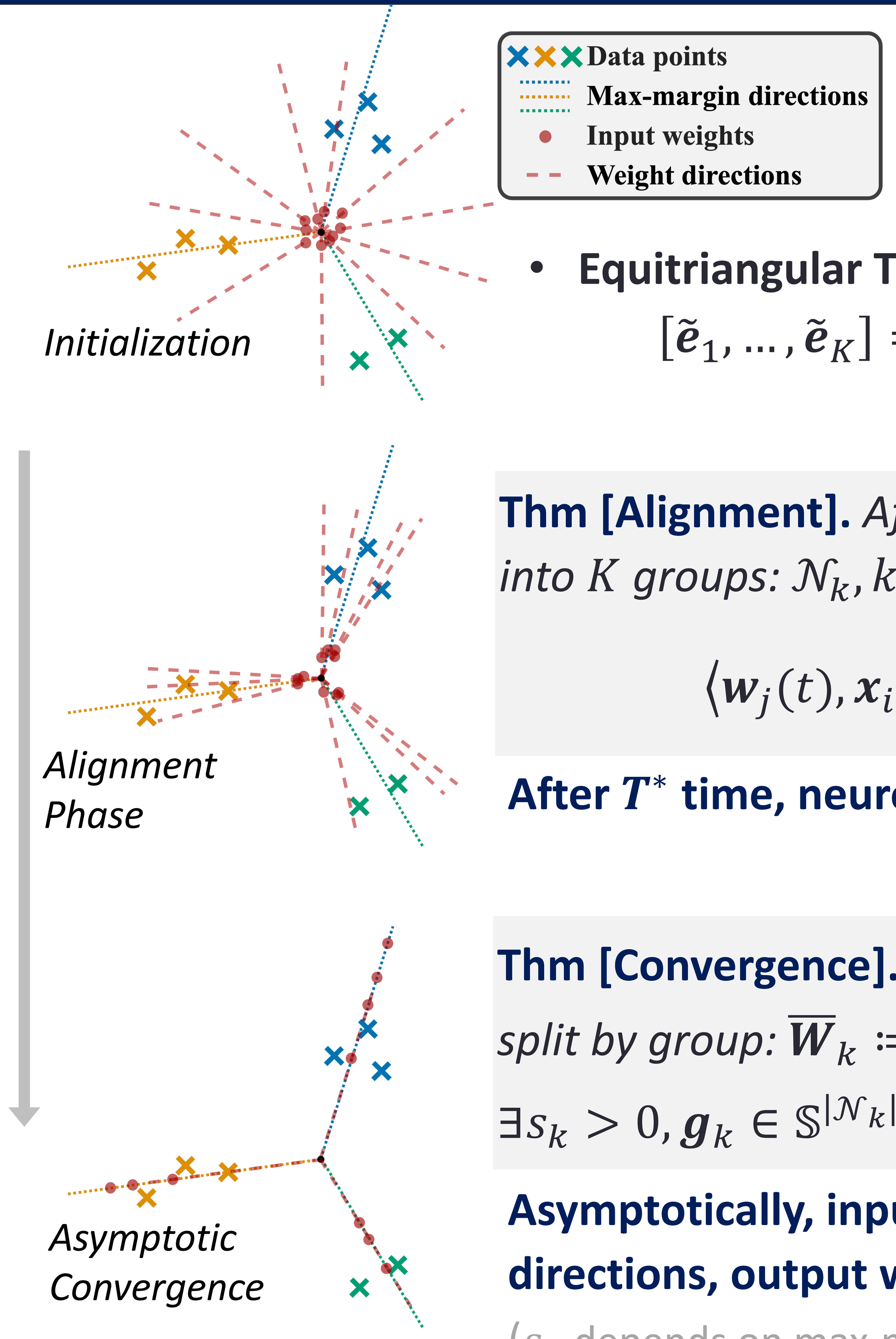
- **ReLU Network:** $f: \mathbb{R}^D \times \Theta \to \mathbb{R}^K, \theta = \{V, W\}$

$$f(x;\theta) = V\sigma(W^\top x) = \sum_{j=1}^h v_j \sigma(\langle w_j, x\rangle), \sigma\text{: ReLU}$$

  - **Input weights**: w_j ; **Output weights**: v_j
  - **Neurons**: $(w_j, v_j), j = 1, \dots, h > K$
  - **Last-layer Feature**: $\phi_\theta(x) = \sigma(W^\top x)$
  - **Last-layer Classifier**: V
- **Cross-Entropy Loss**: $\mathcal{L} = \sum_{i=1}^n \ell_{\mathrm{CE}}(y_i, f(x_i;\theta))$
- **Gradient flow (GF)** with small initialization:

$$\dot{\theta} = -\nabla_\theta \mathcal{L}, \qquad \|\theta(0)\| \ll 1$$

CONVERGENCE OF GRADIENT FLOW



- **Max-margin directions**:

$$u_k = \arg\max_{u\in\mathbb{S}^{D-1}} \min_{i \text{ in class } k} \langle x_i, u\rangle$$

- **Equitriangular Tight Frame (ETF) directions**:

$$[\tilde{e}_1, \dots, \tilde{e}_K] = \sqrt{K/(K-1)}\,(I - \mathbb{1}\mathbb{1}^\top/K)$$

Thm [Alignment]. *After some T^*, the h neurons are split into K groups: $\mathcal{N}_k, k\in[K]$, s.t. $\forall t \geq T^*, \forall k, \forall j \in \mathcal{N}_k$,*

$$\langle w_j(t), x_i\rangle \begin{cases} > 0, & \forall i \text{ in class } k \\ \leq 0, & \text{otherwise}\end{cases}$$

After T^* time, neurons in $\mathcal{N}_k$ classify data from class k

Thm [Convergence]. $\|\theta\| \to \infty$; *For* $\bar{\theta} := \lim_{t\to\infty} \theta/\|\theta\|$, *split by group:* $\bar{W}_k := [\bar{w}_j]_{j\in\mathcal{N}_k}, \bar{V}_k := [\bar{v}_j]_{j\in\mathcal{N}_k}$, *then* $\exists s_k > 0, g_k \in \mathbb{S}^{|\mathcal{N}_k|-1}$, *s.t.* $\bar{W}_k = s_k u_k g_k^\top, \bar{V}_k = s_k \tilde{e}_k g_k^\top$

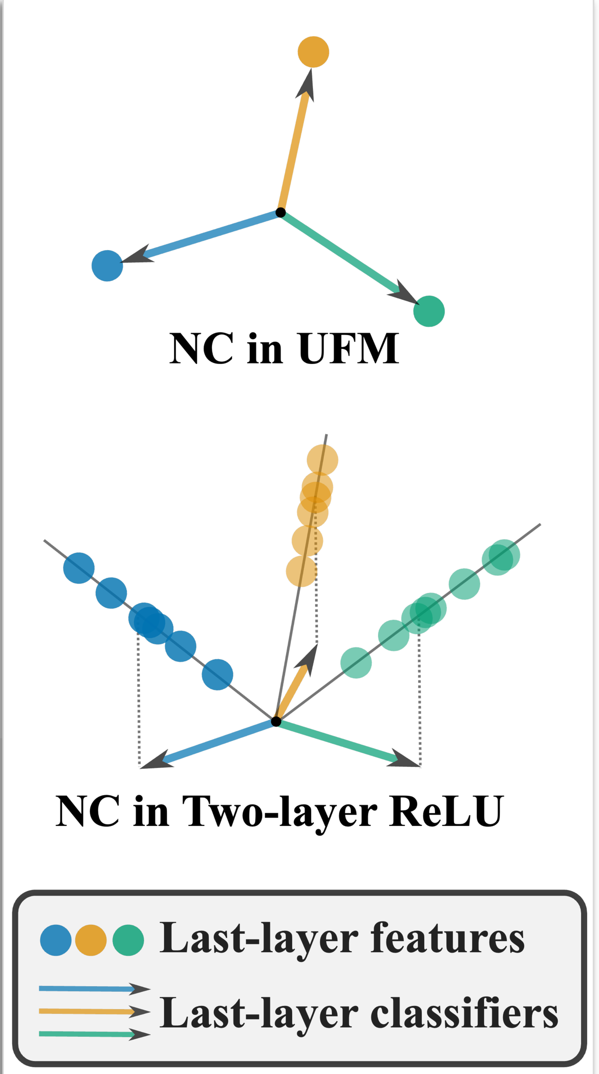
Asymptotically, input weights align with max-margin directions, output weights align with ETF directions

(s_k depends on max-margins $\gamma_k = \max_{u\in\mathbb{S}^{D-1}} \min_{i \text{ in class } k} \langle x_i, u\rangle$)

NEURAL COLLAPSE IN SHALLOW RELU NETWORKS

Recall the **NC for the unconstrained feature model**:

- *Intra-class variability collapse*: Last-layer features of same class collapse into a singleton
- *Maximal class separation*: Class means form an ETF
- *Self duality*: classifiers align with class means

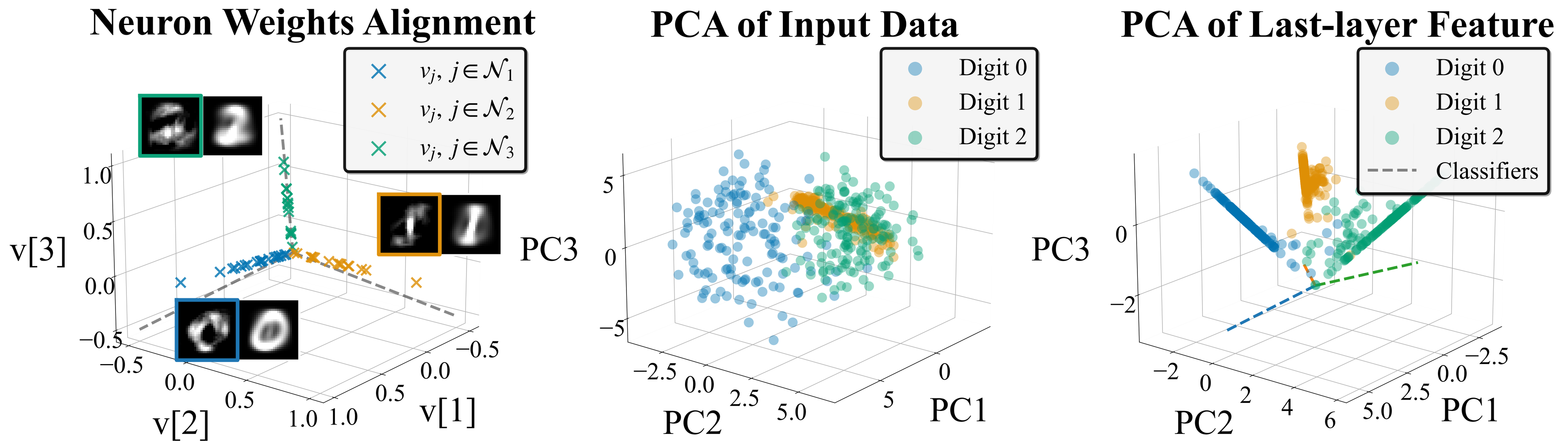


NC for shallow ReLU with orthogonally separable data:

Thm [NC]. *For* $\bar{\theta} := \lim_{t\to\infty} \theta/\|\theta\|$, $\exists \bar{\phi}_k \in \mathbb{S}^{h-1}$, *s.t.*

1. ***(Intra-class directional collapse)***
   $\phi_{\bar{\theta}}(x_i) = \langle s_k u_k, x_k\rangle \cdot \bar{\phi}_k, \forall i \text{ in class } k, \forall k \in [K]$
   Last-layer features collapse into 1-d subspaces
2. ***(Orthogonal class means)***
   $\bar{\phi}_k \geq 0,\ \langle \bar{\phi}_k, \bar{\phi}_{k'}\rangle = 0, \forall k, k' \in [K], k \neq k'$
   Class means form a non-negative orthogonal frame (when normalized)
3. ***(Projected self-duality)*** $\bar{V} = \sqrt{K/(K-1)}\,(I - \mathbb{1}\mathbb{1}^\top/K)[s_1\bar{\phi}_1, \dots, s_K\bar{\phi}_K]^\top$
   classifiers align with centered class means

EXPERIMENT ON MNIST DIGITS



Training a shallow ReLU net for classifying digits 0, 1, 2

- Output weights align with ETF directions (determines neurons' group)
- Average input weight per group aligns with average digit image
- PCA of raw data X has appr. error of 61%
- PCA of collapsed features $\phi_\theta(X)$ has appr. error of 0.2%

Full Paper