Center for Advanced Intelligence Project
The University of Hong Kong
NUS National University of Singapore
UNSW SYDNEY
東京大学 THE UNIVERSITY OF TOKYO

Graph Neural Networks Provably Benefit from Structural Information: A Feature Learning Perspective

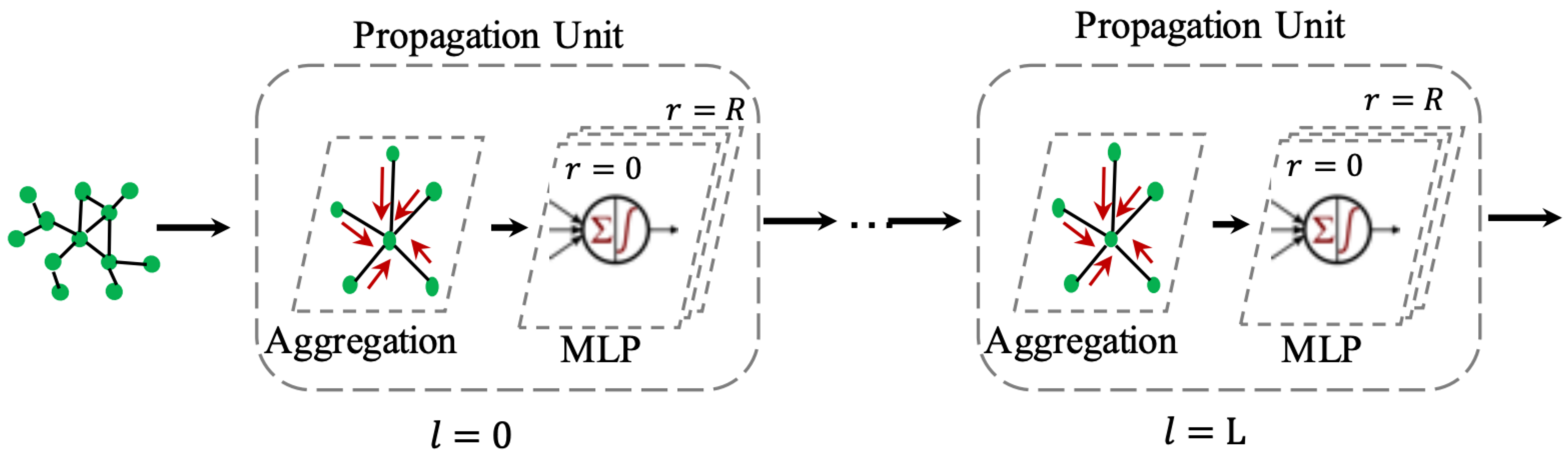
Wei Huang
RIKEN AIP, Japan

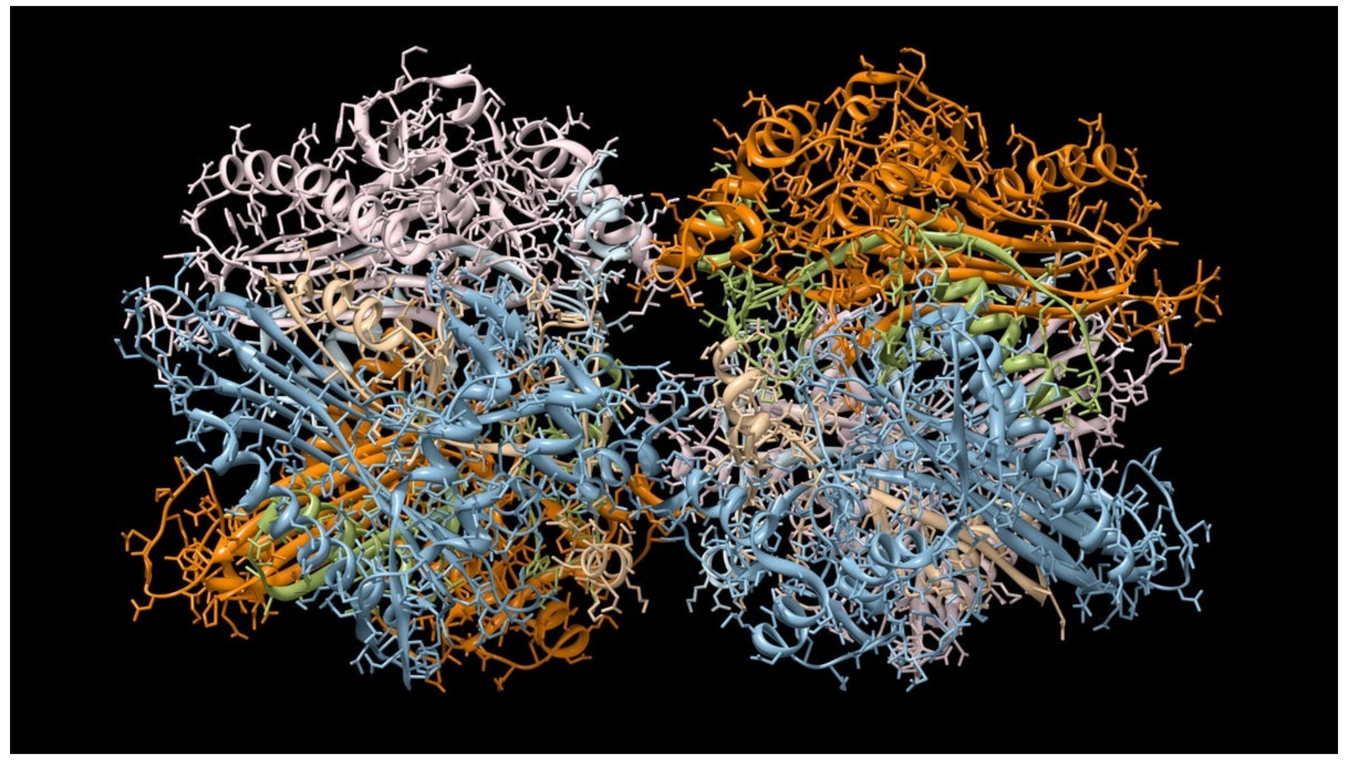
with Yuan Cao, Haonan Wang, Xin Cao, and Taiji Suzuki

ICML 40 Years International Conference On Machine Learning

HiLD: High-dimensional Learning Dynamics Workshop

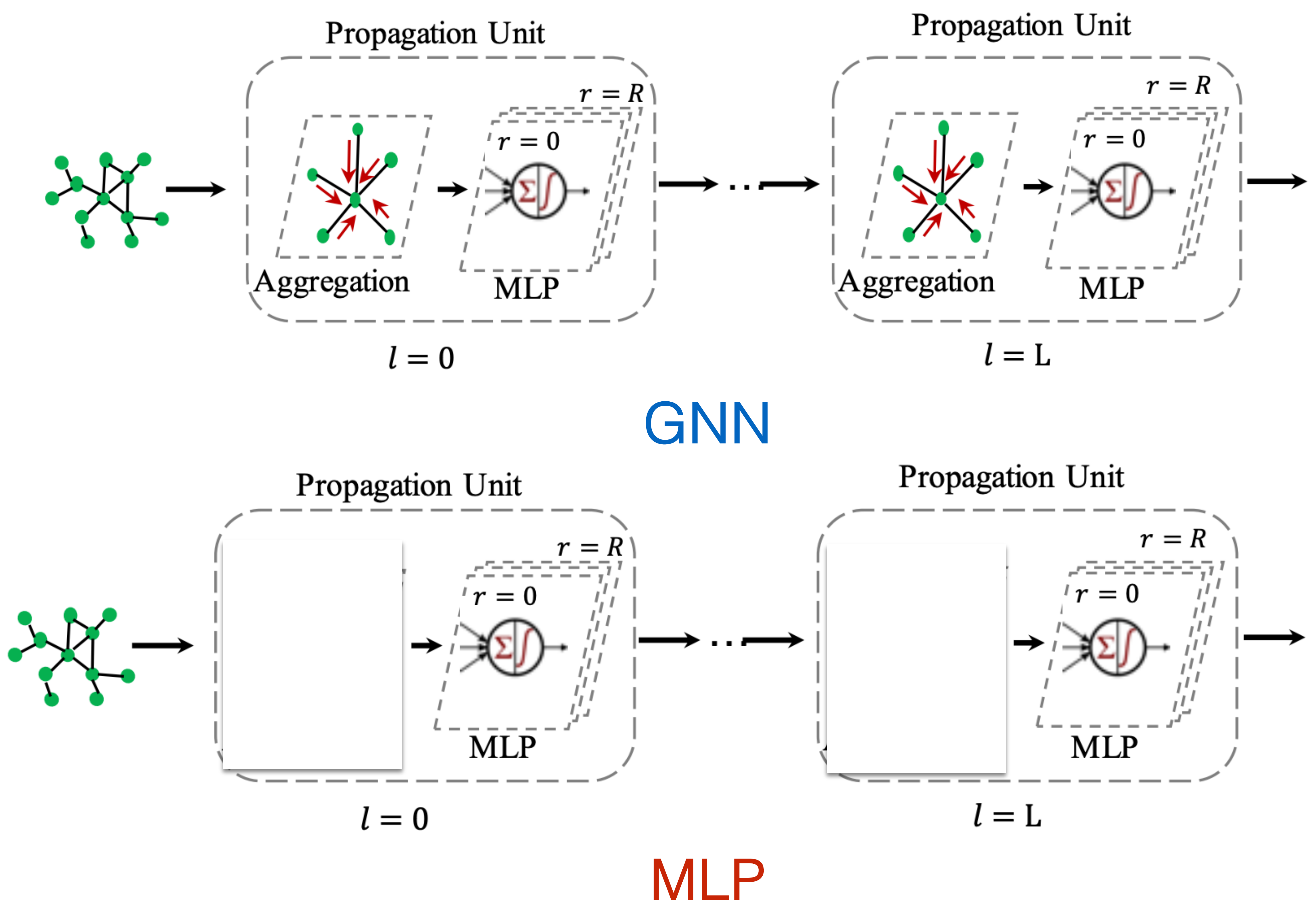
Graph Neural Network Powers Diverse Research Areas

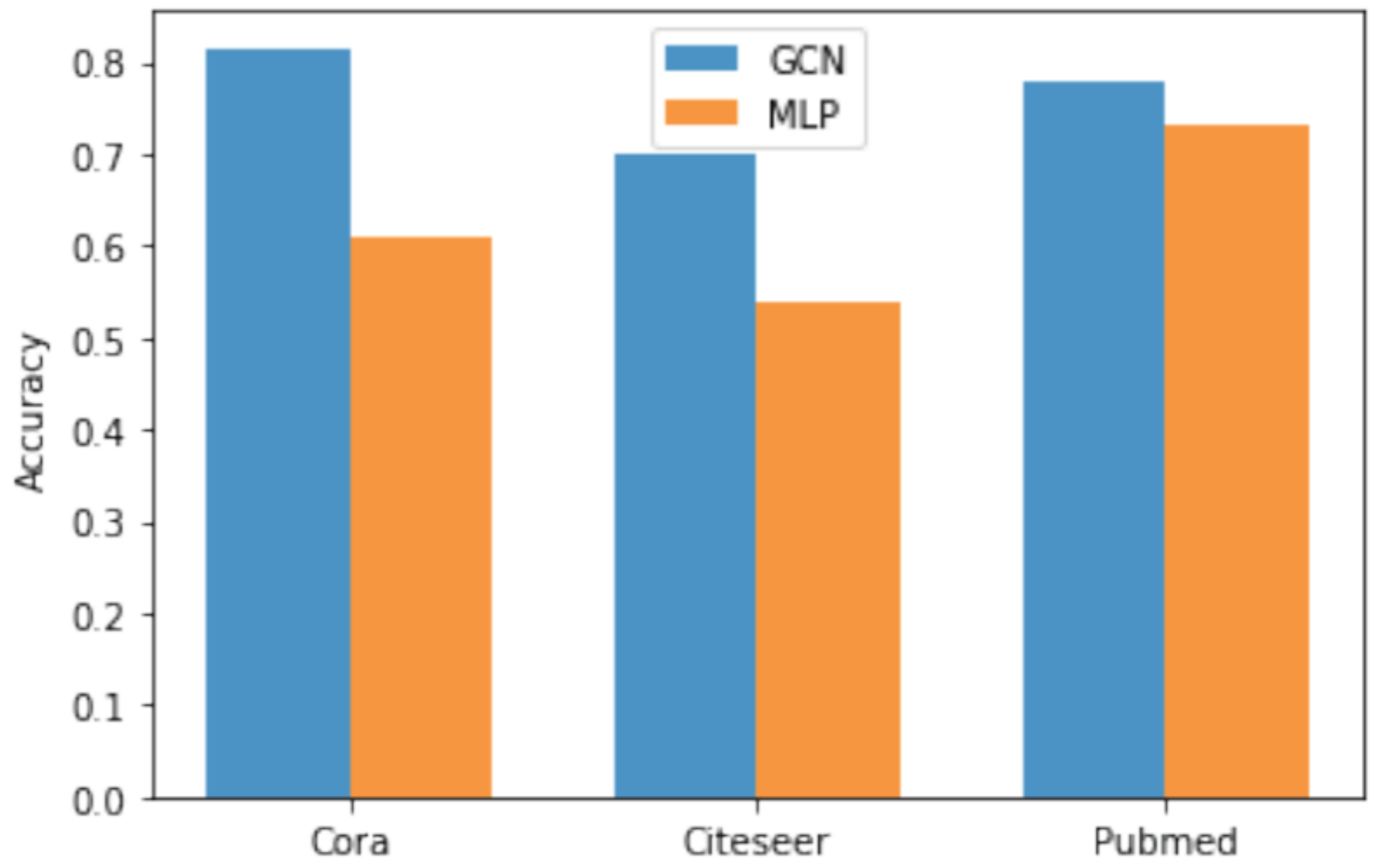




GNN>MLP?

- Empirical evidence from three node classification tasks, suggests GCNs outperform MLPs.



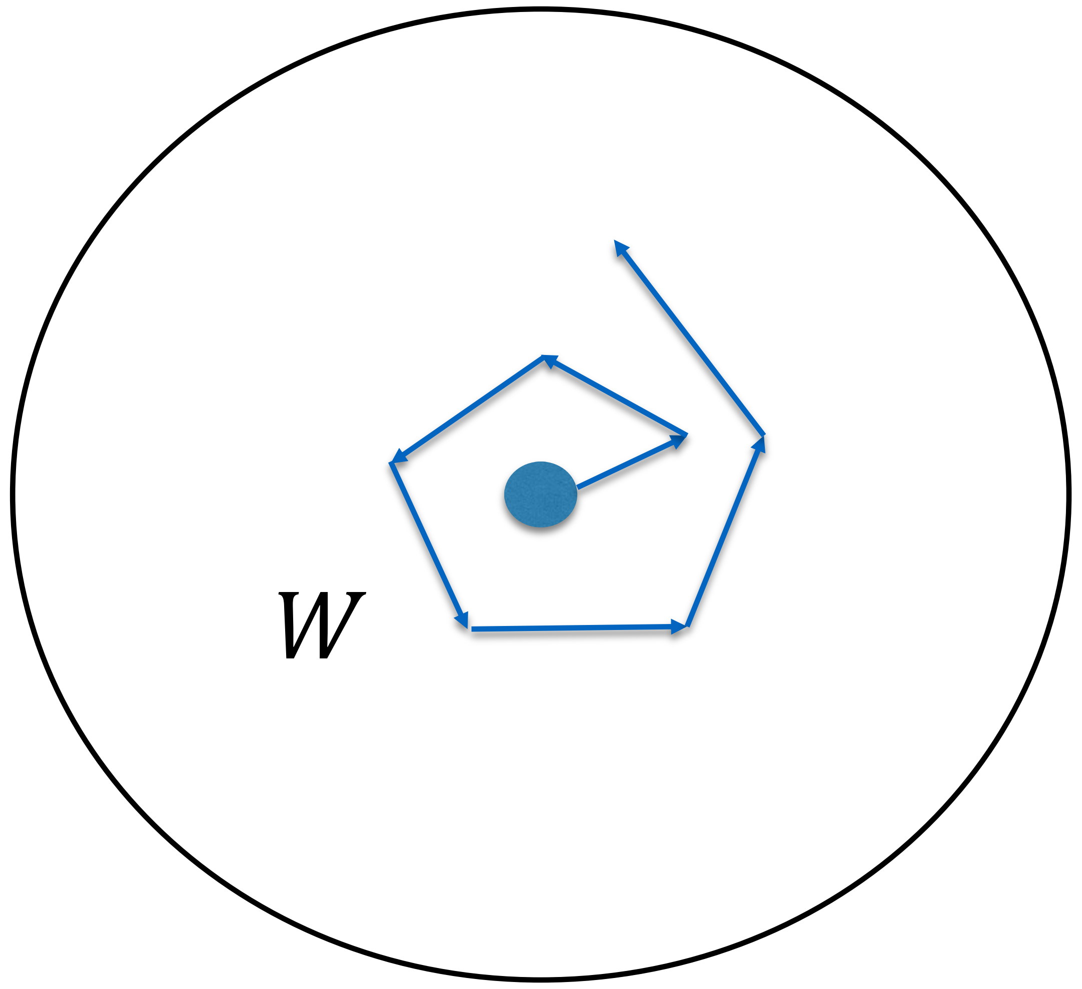


🤔

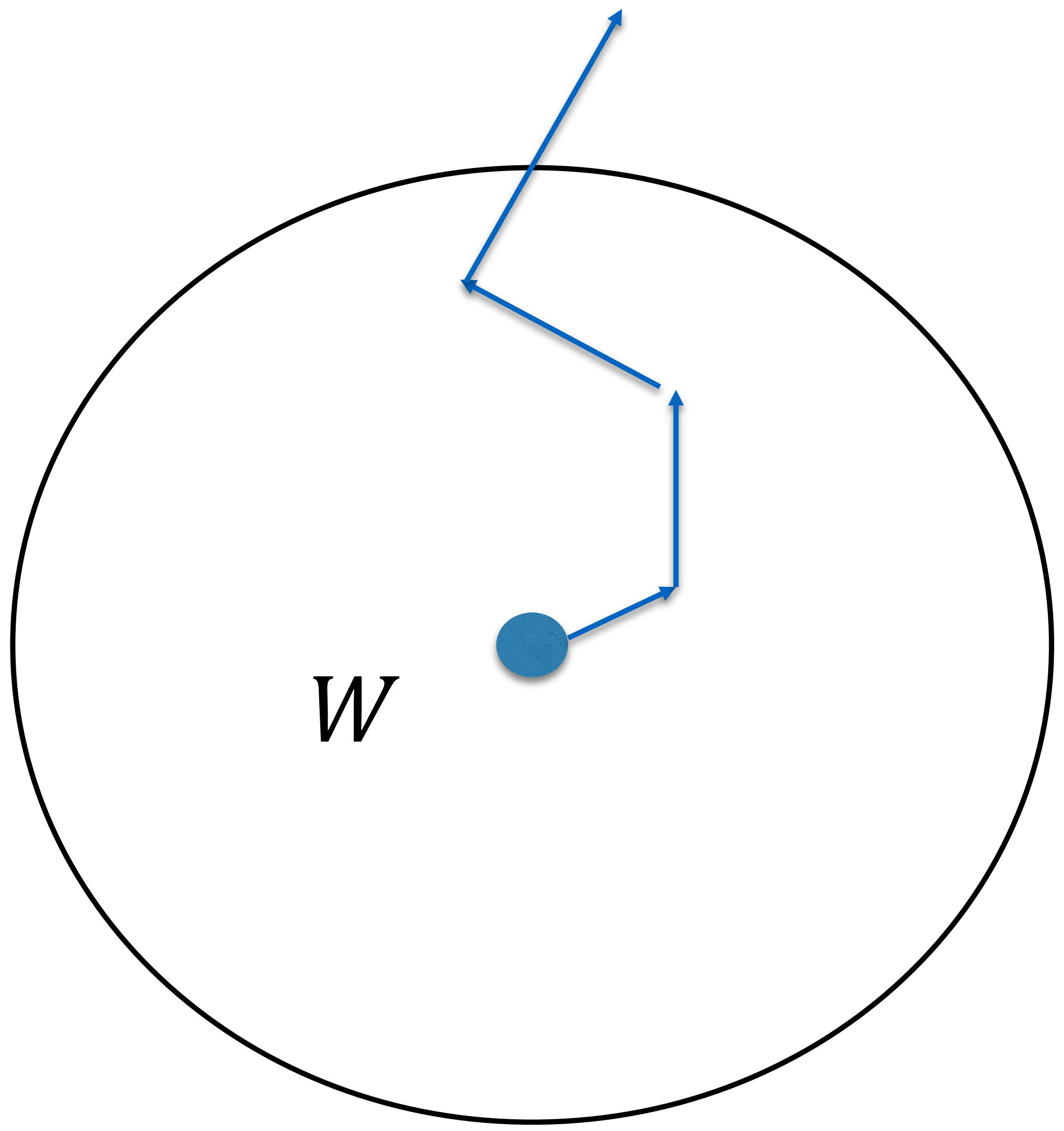
What role does graph convolution play during gradient descent training?

Feature Learning

- It is widely believed in literature that the NTK analyses cannot fully explain the success of deep learning, as the neural networks in the NTK regime are almost “linearized”
- Feature learning theory states that the weights can escape the ball and align to the feature in data.



Lazy training



Feature learning

Data Model

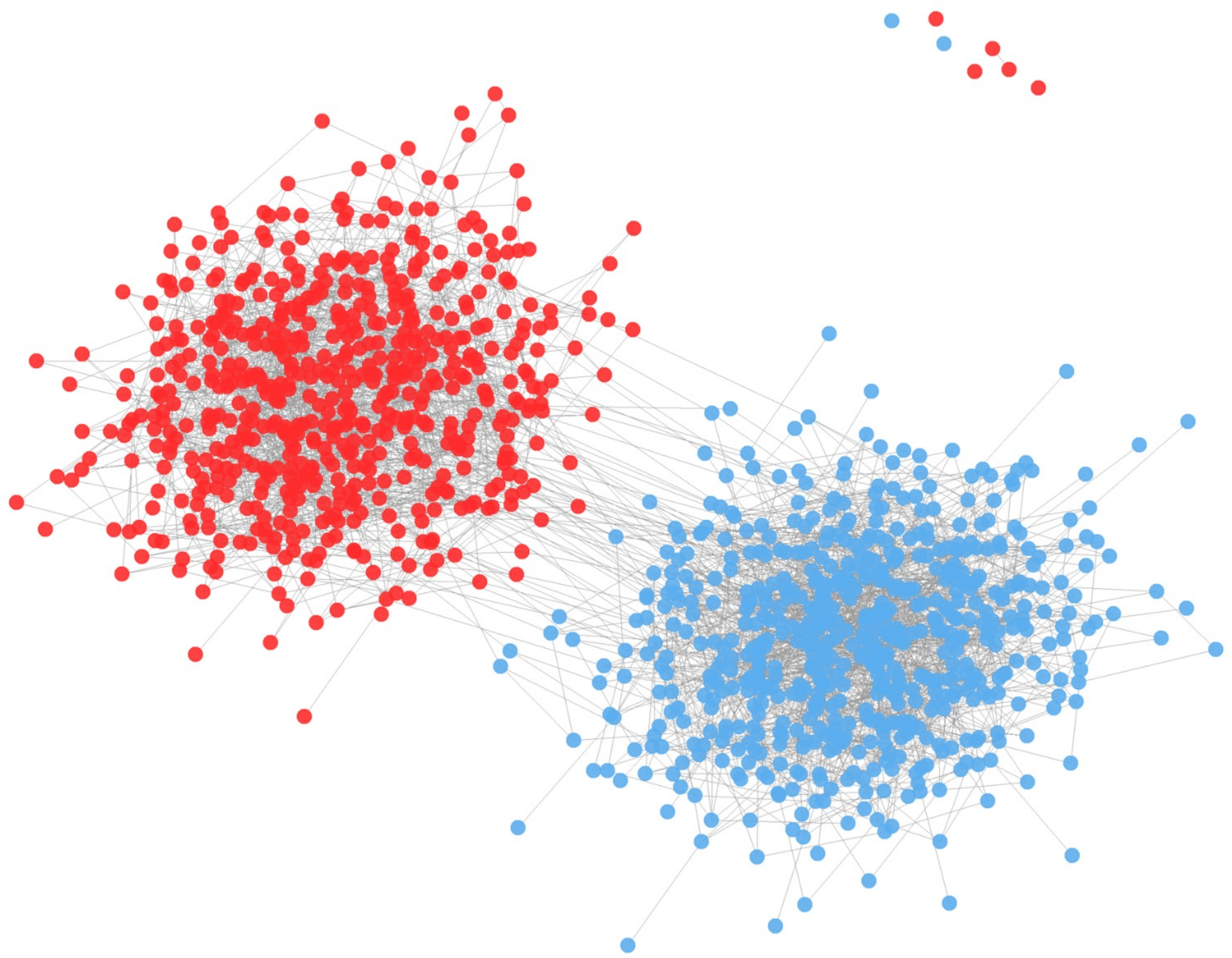
- Two classes $y \in \{-1, 1\}$
- The input $\mathbf{x} \in \mathbb{R}^{2d}$ is composed of a signal patch and noise path:

$$\mathbf{x} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}] = [y \cdot \boldsymbol{\mu}, \boldsymbol{\xi}],$$

Signal patch

$$\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma_p^2 \cdot (\mathbf{I} - \boldsymbol{\mu}\boldsymbol{\mu}^\top \cdot \|\boldsymbol{\mu}\|_2^{-2}))$$

- Stochastic Block Model for graph structure



$$\mathbf{A} = (a_{ij})n \times n$$

$$a_{ij} \sim \mathrm{Ber}(p) \qquad y_i = y_j$$

$$a_{ij} \sim \mathrm{Ber}(s) \qquad y_i = -y_j$$

Neural Network Model

- CNN

$$f(\mathbf{W},\mathbf{x}) = F_{+1}(\mathbf{W}_{+1},\boxed{\mathbf{x})} - F_{-1}(\boxed{\mathbf{W}_{-1},\mathbf{x})}$$

$$F_j(\mathbf{W}_j,\mathbf{x}) = \frac{1}{m}\sum_{r=1}^{m}\Big[\sigma(\mathbf{w}_{j,r}^{\top}\mathbf{x}^{(1)}) + \sigma(\mathbf{w}_{j,r}^{\top}\mathbf{x}^{(2)})\Big],$$

- GNN

$$f(\mathbf{W},\tilde{\mathbf{x}}) = F_{+1}(\boxed{\mathbf{W}_{+1},\tilde{\mathbf{x}})} - F_{-1}(\boxed{\mathbf{W}_{-1},\tilde{\mathbf{x}})}$$

$$F_j(\mathbf{W}_j,\tilde{\mathbf{x}}) = \frac{1}{m}\sum_{r=1}^{m}\Big[\sigma(\mathbf{w}_{j,r}^{\top}\tilde{\mathbf{x}}^{(1)}) + \sigma(\mathbf{w}_{j,r}^{\top}\tilde{\mathbf{x}}^{(2)})\Big].$$

$$\tilde{\mathbf{X}} \triangleq [\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2,\cdots,\tilde{\mathbf{x}}_n]^{\top} = \boxed{\tilde{\mathbf{D}}^{-1}\tilde{\mathbf{A}}}\mathbf{X} \in \mathbb{R}^{n\times 2d} \qquad \tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_n$$

Graph Convolution

Gradient descent training

- Gradient descent training

$$\begin{aligned}\mathbf{w}_{j,r}^{(t+1)} &= \mathbf{w}_{j,r}^{(t)} - \eta \cdot \nabla_{\mathbf{w}_{j,r}} L_S^{\mathrm{GCN}}(\mathbf{W}^{(t)}) \\ &= \mathbf{w}_{j,r}^{(t)} - \frac{\eta}{nm}\sum_{i=1}^{n} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \tilde{y}_i \boldsymbol{\mu}\rangle) \cdot j\tilde{y}_i\boldsymbol{\mu} - \frac{\eta}{nm}\sum_{i=1}^{n} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \tilde{\boldsymbol{\xi}}_i\rangle) \cdot j y_i \tilde{\boldsymbol{\xi}}_i\end{aligned}$$

- Weight decomposition

$$\tilde{\boldsymbol{\xi}}_i = D_i^{-1} \textstyle\sum_{k \in \mathcal{N}(i)} \boldsymbol{\xi}_k$$

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + \gamma_{j,r}^{(t)} \cdot \|\boldsymbol{\mu}\|_2^{-2} \cdot \boldsymbol{\mu} + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i.$$

Iterative analysis of the signal-noise decomposition

- To analyze the feature learning process of graph neural networks during gradient descent training, we introduce an iterative methodology, based on the signal-noise decomposition.

Lemma 5.1. *The coefficients $\gamma_{j,r}^{(t)}, \overline{\rho}_{j,r,i}^{(t)}, \underline{\rho}_{j,r,i}^{(t)}$ in decomposition ($\boxed{10}$) adhere to the following equations:*

$$\gamma_{j,r}^{(0)}, \overline{\rho}_{j,r,i}^{(0)}, \underline{\rho}_{j,r,i}^{(0)} = 0, \tag{11}$$

$$\gamma_{j,r}^{(t+1)} = \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \sum_{i=1}^{n} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \tilde{y}_i \boldsymbol{\mu}_i \rangle) y_i \tilde{y}_i \|\boldsymbol{\mu}\|_2^2, \tag{12}$$

$\overline{\rho}_{j,r,i}^{(t)} \triangleq \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \geq 0)$

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \cdot \sum_{k \in \mathcal{N}(i)} D_k^{-1} \cdot \ell_k'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \tilde{\boldsymbol{\xi}}_k \rangle) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}(y_k = j), \tag{13}$$

$\underline{\rho}_{j,r,i}^{(t)} \triangleq \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \leq 0)$

$$\underline{\rho}_{j,r,i}^{(t+1)} = \underline{\rho}_{j,r,i}^{(t)} + \frac{\eta}{nm} \cdot \sum_{k \in \mathcal{N}(i)} D_k^{-1} \cdot \ell_k'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \tilde{\boldsymbol{\xi}}_k \rangle) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}(y_k = -j). \tag{14}$$

Two-stage dynamics

large

small

- Stage one: feature learning

Lemma 5.3. *Under the same conditions as Theorem 4.3, there exists* $T_1 = \tilde{O}(\eta^{-1} m \sigma_0^{2-q} \Xi^{-q} \|\boldsymbol{\mu}\|_2^{-q})$ *such that*

- $\max_r \gamma_{j,r}^{(T_1)} = \Omega(1)$ *for* $j \in \{\pm 1\}$.
- $|\rho_{j,r,i}^{(t)}| = O\left(\sigma_0 \sigma_p \sqrt{d} / \sqrt{n(p+s)}\right)$ *for all* $j \in \{\pm 1\}$, $r \in [m]$, $i \in [n]$ *and* $0 \le t \le T_1$.

- Stage two: convergence analysis

Lemma 5.4. *Let* T, T_1 *be defined in Theorem 4.3 and Lemma 5.3 respectively and* $\mathbf{W}^*$ *be the collection of GCN parameters* $\mathbf{w}_{j,r}^* = \mathbf{w}_{j,r}^{(0)} + 2qm \log(2q/\epsilon) \cdot j \cdot \|\boldsymbol{\mu}\|_2^{-2} \cdot \boldsymbol{\mu}$. *Then under the same conditions as Theorem 4.3, for any* $t \in [T_1, T]$, *it holds that:*

- $\max_r \gamma_{j,r}^{(T_1)} \ge 2, \forall j \in \{\pm 1\}$ *and* $|\rho_{j,r,i}^{(t)}| \le \sigma_0 \sigma_p \sqrt{d/(n(p+s))}$ *for all* $j \in \{\pm 1\}$, $r \in [m]$ *and* $i \in [n]$.
- $\frac{1}{t-T_1+1} \sum_{s=T_1}^{t} L_{\mathcal{S}}^{\mathrm{GCN}}(\mathbf{W}^{(s)}) \le \frac{\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F^2}{(2q-1)\eta(t-T_1+1)} + \frac{\epsilon}{(2q-1)}$.

Here we denote $\|\mathbf{W}\|_F \triangleq \sqrt{\|\mathbf{W}_{+1}\|_F^2 + \|\mathbf{W}_{-1}\|_F^2}$.

Main Result

large

small

Main condition

Theorem 4.3. *Suppose* $\epsilon > 0$, *and let* $T = \tilde{\Theta}(\eta^{-1} m \sigma_0^{-(q-2)} \Xi^{-q} \|\boldsymbol{\mu}\|_2^{-q} + \eta^{-1}\epsilon^{-1} m^3 \|\boldsymbol{\mu}\|_2^{-2})$. *Under Assumption* $\boxed{4.1}$, *if* $n \cdot \mathrm{SNR}^q \cdot \sqrt{n(p+s)}^{q-2} = \tilde{\Omega}(1)$, *where* $\mathrm{SNR} \triangleq \|\boldsymbol{\mu}\|_2 / (\sigma_p \sqrt{d})$ *is the signal-to-noise ratio, then with probability at least* $1 - d^{-1}$, *there exists a* $0 \le t \le T$ *such that:*

- *The GCN learns the signal:* $\max_r \gamma_{j,r}^{(t)} = \Omega(1)$ *for* $j \in \{\pm 1\}$.
- *The GCN does not memorize the noises in the training data:* $\max_{j,r,i} |\rho_{j,r,i}^{(T)}| = \tilde{O}(\sigma_0 \sigma_p \sqrt{d/n(p+s)})$.
- *The training loss converges to* ϵ, *i.e.,* $L_{\mathcal{S}}^{\mathrm{GCN}}(\mathbf{W}^{(t)}) \le \epsilon$.
- *The trained GCN achieves a small test loss:* $L_{\mathcal{D}}^{\mathrm{GCN}}(\mathbf{W}^{(t)}) \le c_1 \epsilon + \exp(-c_2 n^2)$.

where c_1 *and* c_2 *are positive constants.*

GNN vs CNN

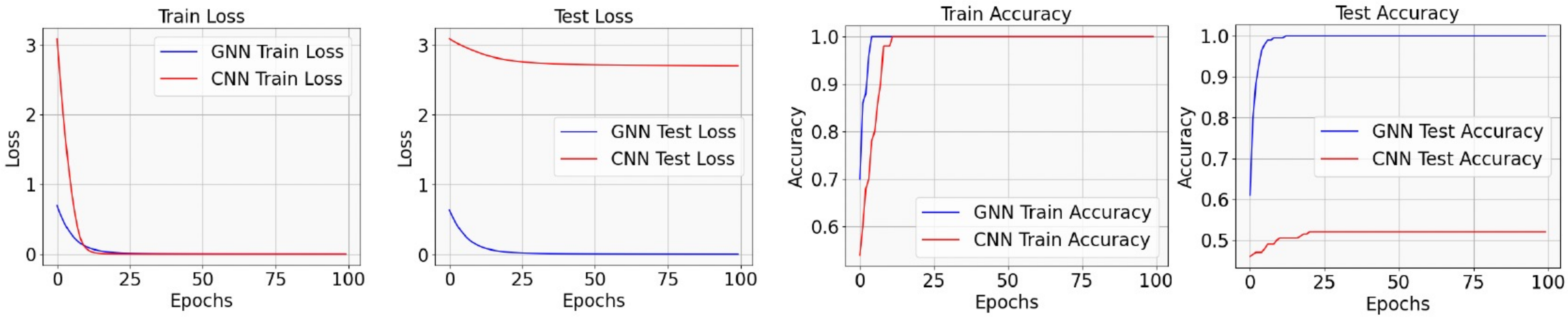
Corollary 4.4 (Informal). *Under assumption 4.1, if* $n \cdot \mathrm{SNR}^q \cdot \sqrt{n(p+s)}^{q-2} = \tilde{\Omega}(1)$ *and* $n^{-1} \cdot \mathrm{SNR}^{-q} = \tilde{\Omega}(1)$, *then with probability at least* $1 - d^{-1}$, *then there exists a* t *such that:*

- *The trained GNN achieves a small test loss:* $L_{\mathcal{D}}^{\mathrm{GCN}}(\mathbf{W}^{(t)}) \leq c_1 \epsilon + \exp(-c_2 n^2)$.
- *The trained CNN has a constant order test loss:* $L_{\mathcal{D}}^{\mathrm{CNN}}(\mathbf{W}^{(t)}) = \Theta(1)$.

Post-training GNN can achieve a small test loss
Post-training CNN has a constant order test loss

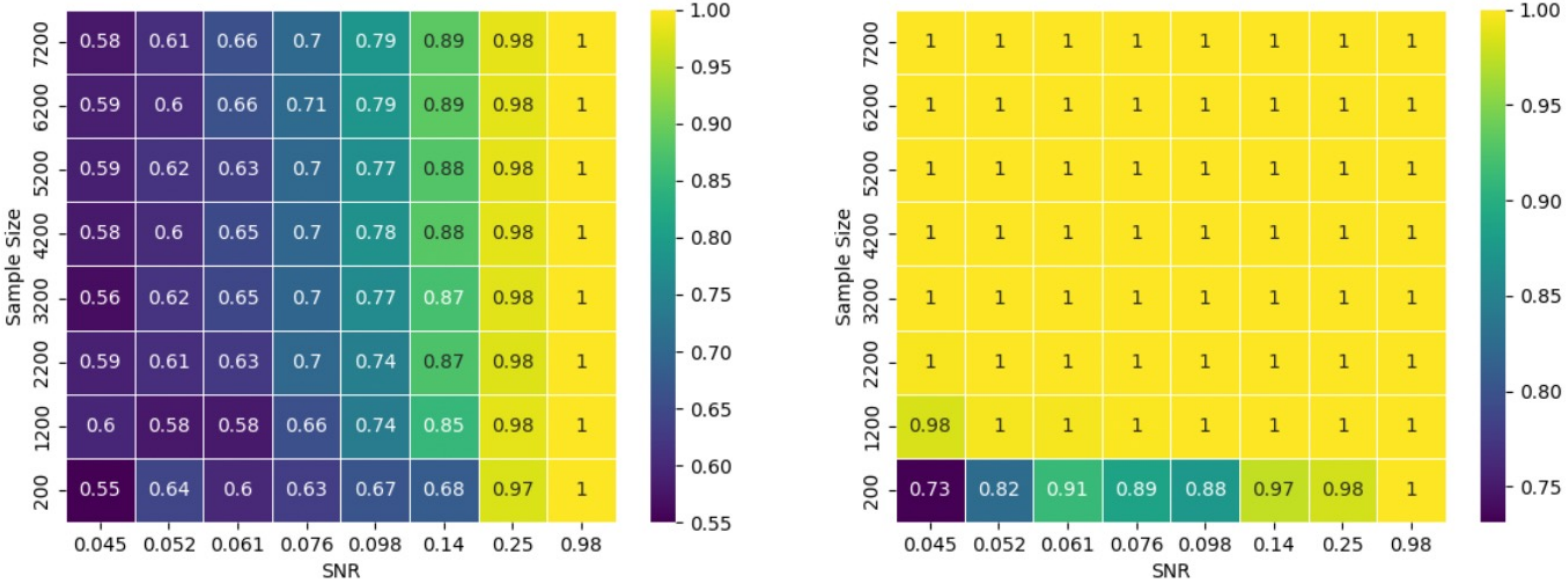
(Cao et al. Benign Overfitting in Two-layer Convolutional Neural Networks, NeurIPS 2022)

GNN vs CNN



Training loss, testing loss, training accuracy, and testing accuracy for both CNN and GNN over a span of 100 training epochs.

GNN vs CNN



Test accuracy heatmap for CNN and GNN after training.

Summary

This paper utilizes a signal-noise decomposition to study the signal learning and noise memorization process in training a two-layer GCN.

We provide specific conditions under which a GNN will primarily concentrate on signal learning, thereby achieving low training and testing errors.

Our results theoretically demonstrate that GCNs, by leveraging structural information, outperform CNNs in terms of generalization ability across a broader benign overfitting regime.

Scan me!

Thank you!

Contact: wei.huang.vr@riken.jp