Polynomial Graph Neural Networks and the Effect of Graph Noise

Ethan Young

University of Washington

AMATH 590: Topics in Random Matrix Theory and High-Dimensional Probability
Spring 2025

Motivation

Background

- Graph neural networks (GNNs) are state-of-the-art for graph learning
- Empirical evidence suggests deeper networks $\neq$ "better"
- Lack theoretical understanding of the role of network depth

Key questions

- Do deeper networks $=$ better performance for GNNs?
- What is the fundamental limit of GNN performance?
- How does graph noise affect classification?

Motivation

- Vinas and Amini attempt to answer some of these questions and their work[1] is the focus of this presentation
- Emphasis on the results that use arguments from matrix perturbation theory and random matrix theory

[1] Vinas and Amini, *Sharp Bounds for Poly-GNNs and the Effect of Graph Noise*, 2024

SSNC with GNNs

In the task of semi-supervised node classification (SSNC) one is given an adjacency matrix $A \in \{0,1\}^{n \times n}$ and is asked to make predictions using a partially observed set of labels

SSNC with GNNs

Input:

- Graph: $A \in \{0,1\}^{n \times n}$
- Node features: $X \in \mathbb{R}^{n \times d}$ where the i-th row is $x_i^\top$ (i.e., the feature vector of node i)
- Partial labels: y_i where $i \in \mathcal{O} \subset [n]$

Goal:

- Predict the unseen labels y_i, $i \in \mathcal{O}^c$

SSNC with GNNs

The prototypical GNN is defined (layer-wise) where, for $Z^{(0)} = X$, the intermediate feature $Z^{(l+1)}$ is

$$Z^{(l+1)} = \varphi\left(AZ^{(l)}W^{(l)}\right)$$

Here,

- $l = 0, 1, ..., k-1$ denotes the layer index,
- $\varphi : \mathbb{R} \to \mathbb{R}$ is a non-linear function applied elementwise, and
- $W^{(l)} \in \mathbb{R}^{d_l} \times \mathbb{R}^{d_{l-1}}$ is the weight matrix for layer l

SSNC with GNNs

- If we take φ to be the identity map, then we obtain $Z^{(k)} = A^k X W^{(0)} \dots W^{(k-1)}$
- We reparameterize and obtain the **poly-GNN**

$$Z^{(k)} = A^k X W \tag{1}$$

SSNC with GNNs

To train a classifier for (1), we form the graph-aggregated features $\phi^{(k)} := A^k X \in \mathbb{R}^{n\times d}$ and then train a linear classifier on the observed pairs

$$\left((\phi^{(k)})_{i\star}, y_i\right), \quad i \in \mathcal{O},$$

where $(\cdot)_{i\star}$ denotes the operator that extracts the i-th row of a matrix

SSNC with GNNs

To explain the performance of $\phi^{(k)}$, we use the **signal-to-noise ratio** (SNR):

$$\frac{1}{\rho^{(k)}} := \min_{i,j:y_i \neq y_j} \frac{\left\| \mathbb{E}\left[\phi_i^{(k)}\right] - \mathbb{E}\left[\phi_j^{(k)}\right] \right\|_2}{\left(\frac{1}{n} \sum_i \left\| \phi_i^{(k)} - \mathbb{E}\left[\phi_i^{(k)}\right] \right\|_2^2 \right)^{1/2}}, \tag{2}$$

where $\phi_i^{(k)}$ is the i-th row of $\phi^{(k)}$ viewed as a column vector

Contextual Stochastic Block Model

- A suitable model is the **contextual stochastic block model** (CSBM)
- Network data (A, X) is CSBM-generated if, for some cluster centers $\mu_1, ..., \mu_L \in \mathbb{R}^d$ and a connectivity matrix $B \in \mathbb{R}^{L \times L}$, the data follows

$$x_i \,|\, y_i \sim \mu_{y_i} + \epsilon_i \quad \text{and} \quad A_{ij} \,|\, y_i, y_j \sim \mathrm{Bern}(B_{y_i y_j}),$$

with $\epsilon_i \sim \mathrm{SG}(\sigma)$ being a zero-mean, sub-Gaussian random variable with parameter σ

Noise Decomposition

- We want to control the noise deviation in (2):

$$\bar{D} := \left(\frac{1}{n} \sum_i \left\| \phi_i^{(k)} - \mathbb{E}\left[\phi_i^{(k)}\right] \right\|_2^2 \right)^{1/2} = \left(\frac{1}{n} \sum_{i,m} D_{im}^2 \right)^{1/2}$$

- For k-hop aggregated features $\phi_{im}^{(k)} = (A^k X)_{im}$, we have

$$D_{im} = \underbrace{\sum_j \left(\left(A^k\right)_{ij} - \mathbb{E}\left[A^k\right]_{ij} \right) x_{jm}}_{=:\Delta_{im}\ (\textbf{graph noise})} + \underbrace{\sum_j \mathbb{E}\left[A^k\right]_{ij} \epsilon_{jm}}_{=:\Delta^{\epsilon}_{im}\ (\textbf{feature noise})}$$

Notation

- Define $\tilde{\mu}_l^{(k)} := \mathbb{E}\phi_l^{(k)}$ to be the ideal center of $\mathcal{C}_l = \{j : y_j = l\}$ (i.e., the set of indices corresponding to the l-th class)
- Then, (2) becomes

$$\frac{1}{\rho^{(k)}} := \frac{\min_{l \neq l'} \left\| \tilde{\mu}_l^{(k)} - \tilde{\mu}_{l'}^{(k)} \right\|_2}{\left(\frac{1}{n} \sum_i \left\| \phi_i^{(k)} - \mathbb{E}\left[\phi_i^{(k)} \right] \right\|_2^2 \right)^{1/2}} = \min_{l \neq l'} \frac{S(l, l')}{\bar{D}},$$

where $S(l, l') := \left\| \tilde{\mu}_l^{(k)} - \tilde{\mu}_{l'}^{(k)} \right\|_2$, $\bar{D}$ defined as before

Notation

- Let $p_{ij} = \mathbb{E}[A_{ij}]$ and let $\nu_n := np_{\max}$, where $p_{\max} := \max_{i,j} p_{ij} = \max_{l,l'} B_{ll'}$
- For cluster $\mathcal{C}_l$ let $\pi_l = \frac{|\mathcal{C}_l|}{n}$ and $\pi = (\pi_1, \ldots, \pi_L)$
- Let $\mu = [\mu_1, \ldots, \mu_L] \in \mathbb{R}^{d \times L}$ where $\mu_l = \mathbb{E}[x_i]$ for class l
- Define $\bar{\xi}_l^{(k)} = \mu \left(\Pi \frac{nB}{\nu_n} \right)^k e_l$ where $\Pi := \mathrm{diag}(\pi) \in \mathbb{R}^{L \times L}$

Assumptions

Sparsity structure

- (A1) For every class l, we have $nB_{ll'} \geq c_B \nu_n$ and $nB_{ll'} \leq C_B \nu_n^{1-\delta}$ where $c_B, C_B > 0$ and $\delta \in (0, \infty]$
- (A2) $\nu_n \leq (1 - c_\nu)n$, $c_\nu \in (0, 1)$

Class balance

- (A3) $L\pi_l \geq c_\pi$ and $\sqrt{L}\|\pi\|_2 \leq C_\pi$

Feature separation

- (A4) $\|\mu\| \leq C_\mu \sqrt{d}$
- (A5) $\left\|\bar{\xi}_l^{(k)} - \bar{\xi}_{l'}^{(k)}\right\|_2 \geq c_\xi \sqrt{d}$, where $c_\xi \leq 1$

Additional Conditions

Sparsity growth

- (C1) $\nu_n \gtrsim \max\left\{\log n, \frac{LC_\mu^2 C_k^2}{c_\pi c_\xi^2}\right\}$

Sample size

- (C2) $\min\left\{\frac{n}{k}, \frac{\nu_n^\delta}{C_B}\right\} \geq \frac{4C_\mu L}{c_\pi c_\xi}$
- (C3) $\min\left\{(2k-1)^{-2}n, \nu_n^\epsilon\right\} \geq \frac{\kappa_1}{2\|\mu\|_{\max}^2}$

Moment control

- $r_n(\epsilon) \geq 4$ where r_n controls moment growth

Main Result (Formal)

Theorem 1 (Signal bound)

Assume (A3)–(A5), and (C1). Then, for $l \neq l'$,

$$\frac{c_\xi}{2}\sqrt{d}\,\nu_n^k \leq S(l,l') \leq \sqrt{8d}\,C_\mu C_\pi^k \nu_n^k.$$

In other words, the signal $S(l,l')$ grows precisely at the rate ν_n^k

Main Result (Formal)

Theorem 2 (Noise upper bound)

Assume $\nu_n \geq k e^{2(k-1)}$ *and* $r_n(\epsilon) \geq 2$. *Then, for all real* $r \in [2, r_n(\epsilon)]$,

$$\mathbb{E}\left[|\bar{D}|^r\right] \leq \left(\kappa_3 \sqrt{8dr}\, \nu_n^{k-1/2}\right)^r .$$

Moreover, for $u \geq 8de$,

$$\mathbb{P}\left(\bar{D} \geq \kappa_3 \nu_n^{k-1/2} \sqrt{u}\right) \leq \exp\left(-\frac{1}{2}\min\left\{\frac{u}{4de}, r_n\right\}\right) .$$

Main Result (Formal)

Theorem 3 (Noise lower bound)

Assume (A1)–(A3), (A5), and (C2)–(C3). Then, for any $\eta \in (0,1)$,

$$\mathbb{P}\left(\bar{D} \geq \sqrt{\eta\kappa_1 d}\,\nu_n^{k-1/2}\right) \geq (1-\eta)^2 \frac{\kappa_1^2}{\kappa_2^2}.$$

Combined with Theorem 2, we see that the noise $\bar{D}$ grows precisely at the rate $\nu_n^{k-1/2}$

Main Result (Informal)

Combining Theorems 1–3, the SNR grows at the precise rate $\nu_n^{1/2}$ independent of k

Theorem 4 (SNR bound)

Let (A, X) be generated from an L-class CSBM satisfying (A1)–(A5) with $\nu_n \gtrsim \log n$ and n sufficiently large. Then, for any $k \geq 1$, with high probability,

$$\sqrt{\nu_n}\,\rho^{(k)} \leq C\,c_\xi^{-1}$$

for a constant C independent of n and k. Furthermore, with probability bounded away from zero,

$$\sqrt{\nu_n}\,\rho^{(k)} \geq c\,c_\xi$$

for a constant $c > 0$ independent of n and k.

Signal Analysis

We introduce some notation:

- $P := ZBZ^\top$, where $Z \in \{0,1\}^{n \times L}$ is the cluster membership matrix for y
- $M := \mu Z^\top \in \mathbb{R}^{d \times n}$
- $\tilde{S}(l, l')$ where $\tilde{S}(l, l') \asymp \nu_n^k$ (signal proxy)

Signal Analysis

- We now show that $\tilde{S}(l, l')$ is close to the signal deviation $\left\| \tilde{\mu}_l^{(k)} - \tilde{\mu}_{l'}^{(k)} \right\|_2$
- Using a Banach-valued mean value theorem, we obtain the following lemmas:

Lemma 5

$\left\| \mathbb{E}[A]^k - P^k \right\| \leq \frac{k \nu_n^k}{n}$.

Signal Analysis

Lemma 6 (Concentration inequality for A^k)

Suppose that $\nu_n \geq c'_\nu \log n \geq 1$ for some constant $c'_\nu > 0$. Then, for any integer $k \geq 1$, the spectrum of A concentrates as

$$\mathbb{E}\left\|A^k - \mathbb{E}[A]^k\right\| \leq C_k\, \nu_n^{k-1/2},$$

where $C_k = k\, 2^k \left(C + \sqrt{\left(\frac{c}{c'_\nu}\right)(k+1)}\right)^k$ for some universal constants $C > 1$ and $c > 0$.

Signal Analysis

Proof sketch of Theorem 1.

By Lemmas 5 and 6, $\|\mathbb{E}[A^k] - P^k\| \le 2C_k \nu_n^{k-1/2}$. Then, we have

$$\left| \left\| \tilde{\mu}_l^{(k)} - \tilde{\mu}_{l'}^{(k)} \right\| - \tilde{S}(l,l') \right| \le \left\| M \left(\mathbb{E}\left[A^k\right] - P^k \right) w \right\|_2$$

$$\vdots$$

$$\le \sqrt{8dL/c_\pi}\, C_\mu C_k \nu_n^{k-1/2}$$

By (C1) we obtain $1 \ge \frac{c_\xi}{2} \ge \sqrt{8L/c_\pi}\, C_\mu C_k \nu_n^{-1/2}$, hence

$$\frac{c_\xi}{2}\sqrt{d}\,\nu_n^k \le S(l,l') \le \sqrt{8d}\, C_\mu C_\pi^k \nu_n^k$$

as desired. □

Noise Analysis

There are two main ingredients for proving Theorems 2 and 3:

1. Walk analysis (walk sequences, trees)
2. High-order moment bounds

Noise Analysis

Recall that $D_{im} =: \Delta_{im} + \Delta_{im}^{\epsilon}$

- We can upper-bound the r-th moment of the noise as

$$\mathbb{E}(\bar{D})^r \leq \frac{d^{r/2-1}}{n} \sum_{i,m} \mathbb{E} D_{im}^r \tag{3}$$

- We have for $r \in 2\mathbb{N}$,

$$D_{im}^r \leq 2^{r-1} \left(\Delta_{im}^r + (\Delta_{im}^{\epsilon})^r \right) \tag{4}$$

- We first control $\mathbb{E} \left(\Delta_{im}^{\epsilon} \right)^r$

Noise Analysis

- Recall $\epsilon_{im} \sim \mathrm{SG}(\sigma)$
- It follows that $\Delta^{\epsilon}_{im} \sim \mathrm{SG}\left(\sqrt{\sigma^2 \sum_j \mathbb{E}[A^k]^2_{ij}}\right)$
- We have also the following lemma[2]:

Lemma 7

If Z is sub-Gaussian with parameter σ, then $\mathbb{E}|Z|^r \leq \left(C_1 \sigma r^{1/2}\right)^r$ where C_1 is a numerical constant.

[2] Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, 2018

Noise Analysis

We have

$$\left(\mathbb{E}[A^k]_{ij}^2\right)^{1/2} \le \mathbb{E}[A^k]_{ii} + \sqrt{n}\max_{j\neq i}\mathbb{E}[A^k]_{ij}$$
$$\vdots$$
$$\le 4\nu_n^{k-1/2}$$

Applying Lemma 7 gives

$$\mathbb{E}\left(\Delta_{im}^{\epsilon}\right)^r \le \left(4C_1\sigma\nu_n^{k-1/2}r^{1/2}\right)^r, \qquad (5)$$

hence Δ_{im}^{ϵ} is sub-Gaussian with parameter $\lesssim \sigma\nu_n^{k-1/2}$

Noise Upper Bound

We have the following lemma:

Lemma 8

Let $\eta > 0$ *and* $r_0 \in 2\mathbb{R} \cup \{\infty\}$. *Assume that for all even integers* $r \leq r_0$, *we have*

$$\mathbb{E}\left[|\Delta|^r\right] \leq \left(K(C\eta r)^\eta\right)^r. \tag{6}$$

Then, (6) holds for all real $r \in [2, r_0]$ *with* C *replaced with* $2C$. *Moreover, if* $x \geq 4\eta Ce$, *then*

$$\mathbb{P}\left(|\Delta| \geq Kx^\eta\right) \leq \exp\left(-\min\left\{\frac{x}{2Ce}, \eta r_0\right\}\right).$$

Noise Upper Bound

Proof sketch of Theorem 2.

Combining (3) and (4), we have

$$\mathbb{E}(\bar{D})^r \leq \frac{d^{r/2-1}}{n} \sum_{i,m} 2^{r-1} \left(\mathbb{E}\left[\Delta_{im}^r\right] + \mathbb{E}(\Delta_{im}^{\epsilon})^r\right)$$

We use a walk analysis to bound the first term and (5) to bound the second. Applying Lemma 8 with specific choices of constants, we obtain

$$\mathbb{E}\left[|\bar{D}|^r\right] \leq \left(\kappa_3 \sqrt{8dr}\, \nu_n^{k-1/2}\right)^r .$$

□

Noise Lower Bound

We have the following lemmas to help prove Theorem 3:

Lemma 9

Assume (C2)–(C3) and $r_n \geq 2$*. Then,*

$$\mathbb{E}(\bar{D})^2 \geq \kappa_1 d \nu_n^{2k-1}.$$

Lemma 10

Under the assumptions of Lemma 9, further assume $r_n \geq 4$*. Then,*

$$\frac{(\mathbb{E}\bar{D}^2)^2}{\mathbb{E}\bar{D}^4} \geq \frac{\kappa_1^2}{\kappa_2}.$$

Noise Lower Bound

Proof of Theorem 3.

Applying the Paley-Zygmund inequality to the non-negative quantity $\bar{D}^2$ yields

$$\mathbb{P}\left(\bar{D}^2 \geq \eta\,\mathbb{E}\bar{D}^2\right) \geq (1-\eta^2)\frac{(\mathbb{E}\bar{D}^2)^2}{\mathbb{E}\bar{D}^4}.$$

Using Lemma 9 on the LHS and Lemma 10 on the RHS, we obtain

$$\mathbb{P}\left(\bar{D} \geq \sqrt{\eta\kappa_1 d}\,\nu_n^{k-1/2}\right) \geq (1-\eta)^2\frac{\kappa_1^2}{\kappa_2^2}.$$

$\square$

Consequences

- $\rho^{(k)}$ is rate invariant to the Poly-GNN depth k
- Graph aggregation by GNNs does help at the rate $\sqrt{\nu_n}$
- Oversmoothing does not affect SNR rate
- The rate optimal choice for SNR is obtained at $k = 1$

Future Work: A CLT for GNNs

- Consider 1-D node features $x \in \mathbb{R}^n$
- Define $\xi_i^{(k)} = \nu_n^{1/2-k}\left((A^k x)_i - \mathbb{E}[(A^k x)_i]\right)$
- We look at the empirical distribution of $\{\xi_i^{(k)}\}_{i=1}^n$:

$$\mathbb{P}_n^{(k)} := \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i^{(k)}}$$

- One can show $\mathbb{P}_n^{(k)}$ leads to

$$\mathbb{G} := \sum_{l=1}^L \pi_l N(0, \sigma_l^2),$$

i.e., a scale-mixture of Gaussians

References I

Vershynin, Roman. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018, pp. 11–37.

Vinas, Luciano and Arash A. Amini. *Sharp Bounds for Poly-GNNs and the Effect of Graph Noise*. 2024. arXiv: 2407.19567 [cs.LG]. URL: https://arxiv.org/abs/2407.19567.