POLYNOMIAL GNNS AND THE EFFECT OF GRAPH NOISE

ETHAN YOUNG

Department of Applied Mathematics, University of Washington, Seattle, WA
ethanjy@uw.edu

1. Introduction

Graph neural networks (GNNs) have been shown to be the state-of-the-art for graph learning [6]. Although empirical evidence suggests that deeper networks are not necessarily better due to the phenomenon of over-smoothing [2], we lack theoretical understanding of the role that network depth plays. Vinas and Amini [4] explore the implications of GNN depth in semi-supervised node classification (SSNC) and their work is the focus of this report. They consider GNNs with polynomial features and derive a misclassification rate that is sharp and invariant to network depth. We give a brief overview of some of the ideas, particularly from random matrix theory, that they use to derive this rate.

1.1. **SSNC and GNNs.** In the task of 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. More formally, we observe the graph A, the 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), and a subset of the labels y_i where $i \in \mathcal{O} \subset [n]$. The goal is to predict the unseen labels y_i, $i \in \mathcal{O}^c$.

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. Recent empirical work [5] suggests that one may replace φ with the identity function without noticeably changing the performance on various SSNC benchmarks. Thus, if we take φ to be the identity map, then we obtain $Z^{(k)} = A^k X W^{(0)} \cdots W^{(k-1)}$. We reparameterize the product of weight matrices into a single weight matrix W and obtain

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

which we refer to as the *poly-GNN*.

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

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

Date: June 9, 2025.

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

1.2. **CSBM and Noise Decompositions.** A suitable theoretical model for SSNC is the contextual stochastic block model (CSBM) [1]. We say 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 $\epsilon_i \sim \mathrm{SG}(\sigma)$ denoting a zero-mean, sub-Gaussian random variable with parameter σ.

We control the noise deviation in (2) using the following decomposition. Let

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

where

$$\begin{aligned} D_{im} &= \sum_j \left(\left(A^k\right)_{ij} - \mathbb{E}\left[A^k\right]_{ij}\right) x_{jm} + \sum_j \mathbb{E}\left[A^k\right]_{ij} \epsilon_{jm} \\ &=: \Delta_{im} + \Delta_{im}^{\epsilon}. \end{aligned}$$

We refer to Δ_{im} as graph noise and Δ_{im}^{ϵ} as feature noise.

2. Main Results

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

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

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

We now introduce some notation. Let $p_{ij} = \mathbb{E}[A_{ij}]$ and let $\nu_n := np_{\max}$, where $p_{\max} := \max_{i,j} p_{ij} = \max_{l,l'} B_{ll'}$. The parameter ν_n captures the sparsity of the graph. For cluster $\mathcal{C}_l$ let $\pi_l = |\mathcal{C}_l|/n$ and $\pi = (\pi_1, ..., \pi_L)$. Let $\mu = [\mu_1, ..., \mu_L] \in \mathbb{R}^{d\times L}$ where $\mu_l = \mathbb{E}[x_i]$ for class l. Define $\bar{\xi}_l^{(k)} = \mu\left(\Pi \cdot nb/\nu_n\right)^k e_l$ where $\Pi := \mathrm{diag}(\pi) \in \mathbb{R}^{L\times L}$.

We have the following assumptions:

- (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) $\nu_n \leq (1 - c_\nu)n$, $c_\nu \in (0, 1)$;
- (A3) $L\pi_l \geq c_\pi$ and $\sqrt{L}\|\pi\|_2 \leq C_\pi$;
- (A4) $\|\mu\| \leq C_\mu\sqrt{d}$; and
- (A5) $\left\|\bar{\xi}_l^{(k)} - \bar{\xi}_{l'}^{(k)}\right\|_2 \geq c_\xi\sqrt{d}$, where $c_\xi \leq 1$.

(A1)–(A2) are assumptions on sparsity structure, (A3) on class balance, and (A4)–(A5) on feature separation. We also have the following growth conditions:

- (C1) $\nu_n \gtrsim \max\left\{\log n, \frac{LC_\mu^2 C_k^2}{c_\pi c_\xi^2}\right\}$;
- (C2) $\min\left\{\frac{n}{k}, \frac{\nu_n^\delta}{C_B}\right\} \geq \frac{4C_\mu L}{c_\pi c_\xi}$; and
- (C3) $\min\left\{(2k-1)^{-2}n, \nu_n^\epsilon\right\} \geq \frac{\kappa_1}{2\|\mu\|_{\max}^2}$.

(C1) is a condition on sparsity growth and (C2)–(C3) on sample size. Next, we have $r_n(\epsilon) \geq 4$ where r_n controls moment growth of sub-Gaussian random variables. Finally, let κ_1, κ_2, and κ_3 be constants.

2.1. **Informal Statement.** The main result shows that the SNR in k-hop aggregated features has strong invariance to the depth k as n grows:

Theorem 1 (Informal)**.** *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.

Theorem 1 states that there is a fundamental rate of separation, $\sqrt{\nu_n}$, which is the same for any k-hop aggregated features for n sufficiently large. Moreover, $\rho^{(k)}$ is said to be rate invariant to the poly-GNN depth k. There are two important implications of Theorem 1.

(1) Oversmoothing does not affect the SNR rate.
(2) The rate optimal choice for SNR is obtained at $k = 1$.

These insights increase our theoretical understanding of GNNs and can inform how we design GNN architectures for semi-supervised classification problems.

2.2. **Formal Statement.** At a high level, the results are summarized as follows.

- The signal $S(l, l')$ grows precisely at the rate ν_n^k (Theorem 2).
- The noise $\bar{D}$ grows precisely at the rate $\nu_n^{k-1/2}$ (Theorems 3 and 4).

Combining Theorems 2–4 we see that the SNR grows at the precise rate $\nu_n^{1/2}$ independent of k. We state these results formally. We have the following bound on the signal growth:

Theorem 2 (Signal bound)**.** *Assume (A3)–(A5), and (C1). Then, for $l \neq l'$,*

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

We have the following bounds on the noise growth:

Theorem 3 (Noise upper bound)**.** *Assume $\nu_n \geq ke^{2(k-1)}$ 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 $u \geq 8de$,

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

Theorem 4 (Noise lower bound)**.** *Assume (A1)–(A3), (A5), and (C2)–(C3). Then, for any* $\eta \in (0,1)$*,*

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

Theorem 5 (Main result)**.** *Assume (A1)–(A5), (C1)–(C3), and* $r_n(\epsilon) \geq 4$*. Then, for any* $\alpha \geq \sqrt{2}$*, with probability at least* $1-\exp\left(\frac{1}{2}\min\{\alpha^2, r_n(\epsilon)\}\right)$*, we have*

$$\sqrt{\nu_n}\,\rho^{(k)} \leq \sqrt{e}\,\alpha\left(\frac{\kappa_3}{c_\xi}\right). \tag{3}$$

Moreover, for any $\eta \in (0,1)$*, with probability at least* $(1-\eta)^2\frac{\kappa_1^2}{\kappa_2}$*, we have*

$$\sqrt{\nu_n}\,\rho^{(k)} \geq \sqrt{\frac{\eta}{8}}\cdot\frac{\sqrt{\kappa_1}}{C_\mu C_\pi^k}.$$

Proof. Note that the condition $\nu_n \geq ke^{2(k-1)}$ of Theorem 3 is automatically satisfied by $r_n(\epsilon \geq 4)$. Take $u = \alpha^2 4de$ for $\alpha^2 \geq 2$ in Theorem 3. Then, with probability at least $1-\exp\left(\frac{1}{2}\min\{\alpha^2, r_n(\epsilon)\}\right)$, we have $\bar{D} \leq \kappa_3\nu_n^{k-1/2}\sqrt{4de}\,\alpha$. Combined with the lower bound in Theorem 2, we obtain the claimed upper bound with the same probability

$$\rho^{(k)} \leq \frac{\kappa_3\nu_n^{k-1/2}\sqrt{4de}\,\alpha}{c_\xi\sqrt{d}\,\nu_n^k/2} = \left(\frac{\kappa_3}{c_\xi}\right)\sqrt{e}\,\alpha\,\nu_n^{-1/2}.$$

For the lower bound, we combine Theorem 4 with the lower bound in Theorem 2. □

We remark that in order for the upper bound (3) to hold with high probability, we require $r_n(\epsilon) \to \infty$ as $n \to \infty$, i.e., when $\nu_n \to \infty$.

3. Signal Analysis

We first introduce some notation. Let $P := ZBZ^\top$ where $Z \in \{0,1\}^{n\times L}$ is the cluster membership matrix for y, $M := \mu Z^\top \in \mathbb{R}^{d\times n}$, and $\bar{\mathbf{1}}_{\mathcal{C}_l} := \mathbf{1}_{\mathcal{C}_l}/n_l$. In order to show $\left\|\tilde{\mu}_l^{(k)} - \tilde{\mu}_{l'}^{(k)}\right\| \asymp \nu_n^k$, we construct a proxy $\tilde{S}(l,l')$ where $\tilde{S}(l,l') \asymp \nu_n^k$. Let $w := \bar{\mathbf{1}}_{\mathcal{C}_l} - \bar{\mathbf{1}}_{\mathcal{C}_{l'}}$. Then, after some analysis, we have

$$\tilde{\mu}_l^{(k)} - \tilde{\mu}_{l'}^{(k)} = M\mathbb{E}[A^k]w \quad\text{and}\quad \tilde{S}(l,l') = MP^kw.$$

Using a Banach-valued variant to the mean value theorem, we have the following lemma:

Lemma 6. $\left\|\mathbb{E}[A]^k - p^k\right\| \leq \frac{k\nu_n^k}{n}$.

Using the same Banach-valued mean value theorem and sharp concentrations on $\|A-\mathbb{E}[A]\|$, we obtain the following concentration inequality for A^k:

Lemma 7. *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* $c > 0$*.*

We are now ready to prove the signal bound.

Proof sketch of Theorem 2. By Lemmas 6 and 7, $\|\mathbb{E}[A^k] - P^k\| \leq 2C_k\nu_n^{k-1/2}$. Then, we have

$$\begin{aligned} \left|\left\|\tilde{\mu}_l^{(k)} - \tilde{\mu}_{l'}^{(k)}\right\| - \tilde{S}(l,l')\right| &\leq \left\|M\left(\mathbb{E}\left[A^k\right] - P^k\right)w\right\|_2 \\ &\vdots \\ &\leq \sqrt{8dL/c_\pi}\, C_\mu C_k \nu_n^{k-1/2} \end{aligned}$$

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

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

as desired. ☐

4. Noise Analysis

There are two main ingredients in proving Theorems 3 and 4: a walk analysis and high-order moment bounds. For the purposes of this report, we will sidestep the discussion on the walk analysis, which uses ideas from combinatorics and graph theory.

First, recall that $D_{im} =: \Delta_{im} + \Delta_{im}^\epsilon$. We can upper-bound the r-th moment of the noise as

$$\mathbb{E}(\bar{D})^r \leq \frac{d^{r/2-1}}{n}\sum_{i,m}\mathbb{E}D_{im}^r, \tag{4}$$

where the right hand side follows from the Jensen inequality with expectation operator $\frac{1}{nd}\sum_{i,m}$. We will control the moments $\mathbb{E}D_{im}^r$. For $r \in 2\mathbb{N}$, by the convexity of $x \mapsto x^r$, we have

$$D_{im}^r \leq 2^{r-1}\left(\Delta_{im}^r + (\Delta_{im}^\epsilon)^r\right). \tag{5}$$

We focus on $\mathbb{E}\left(\Delta_{im}^\epsilon\right)^r$. Recall $\epsilon_{im} \sim \mathrm{SG}(\sigma)$. It follows that $\Delta_{im}^\epsilon \sim \mathrm{SG}\left(\sqrt{\sigma^2\sum_j \mathbb{E}[A^k]_{ij}^2}\right)$. We also have the following lemma from [3]:

Lemma 8. *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.*

After a walk analysis, we have

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

Applying Lemma 8 gives

$$\mathbb{E}\left(\Delta_{im}^\epsilon\right)^r \leq \left(4C_1\sigma\nu_n^{k-1/2}r^{1/2}\right)^r. \tag{6}$$

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

We also have the following lemma:

Lemma 9. *Let $\eta > 0$ and $r_0 \in 2\mathbb{R}\cup\{\infty\}$. Assume that for all even integers $r \leq r_0$, we have*

$$\mathbb{E}\left[|\Delta|^r\right] \leq \left(K(C\eta r)^\eta\right)^r. \tag{7}$$

Then, (7) holds for all real $r \in [2, r_0]$ with C replaced with $2C$. Moreover, if $x \geq 4\eta Ce$, then

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

We are now ready to prove Theorem 3.

Proof sketch of Theorem 3. Combining (4) and (5), 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 (6) to bound the second. Applying Lemma 9 with specific choices of constants, we obtain

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

□

We have the following lemmas to help prove Theorem 4:

Lemma 10. *Assume (C2)–(C3) and $r_n \geq 2$. Then,*

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

Lemma 11. *Under the assumptions of Lemma 10, further assume $r_n \geq 4$. Then,*

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

We are now ready to prove the noise lower bound.

Proof of Theorem 4. 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 10 on the LHS and Lemma 11 on the RHS, we obtain

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

□

5. Conclusion

In this report, we give a brief overview of sharp bounds on the signal-to-noise ratio for graph aggregated features derived in [4]. These features have a fundamental misclassification rate that is invariant to the network depth. We focus on the ideas from random matrix theory that are used to obtain these results.

References

[1] Y. Deshpande, A. Montanari, E. Mossel, and S. Sen. Contextual stochastic block models, 2018.
[2] T. K. Rusch, M. M. Bronstein, and S. Mishra. A survey on oversmoothing in graph neural networks, 2023.
[3] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.
[4] L. Vinas and A. A. Amini. Sharp bounds for poly-gnns and the effect of graph noise, 2024.
[5] X. Wang and M. Zhang. How powerful are spectral graph neural networks, 2022.
[6] Z. Yang, W. Cohen, and R. Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 40–48, New York, New York, USA, 20–22 Jun 2016. PMLR.