Robustness & Generalization in Uncertainty-Aware Message Passing Neural Networks

Alesia Chernikova, Moritz Laber★, Narayan G. Sabhahit, and Tina Eliassi-Rad

Northeastern University, Boston, MA, United States

★ laber.m@northeastern.edu

Introduction

Problem Statement

Message Passing Neural Networks (MPNNs) are widely used in graph machine learning. However, existing formulations often ignore uncertainty in node features arising from noise or finite measurement precision [1, 2].

Contributions

1. **Uncertainty Quantification:** We use moment propagation to approximate node embedding distributions in Graph Convolutional Networks (GCNs) [3] and validate this experimentally.
2. **Robustness Certification:** We provide robustness guarantees for node classification with GCNs against L_2 -norm feature perturbations, based on the moments of the node embedding distributions.
3. **Feature Convolution Distance:** We introduce *Feature Convolution Distance* FCD_p, a pseudo-metric on the input space of the Simple Graph Convolution (SGC) [4] architecture, which has the same discriminative power as SGC, show that SGC is Lipschitz continuous in this distance.
4. **Generalization Bound:** We present the first covering number-based generalization bound for node classification with SGC in the presence of node feature uncertainty.

Setup

- **Node features** are random vectors $\xi_i \in \mathbb{R}^f$ at each node, $i \in \{1, \dots, n\}$, drawn from a joint Gaussian distribution $\mathbb{P}$ on $\mathbb{R}^{f \times \dots \times f}$ that allows for feature correlations at a node and across nodes.
- **Graph structure** is fixed and given by the adjacency matrix $A \in \mathbb{R}^{n \times n}$.
- **Architectures** that we consider are GCNs, $X^{(l+1)} = \sigma\big(SX^{(l)}W^{(l)}\big)$ and SGCs $X^{(L)} = \sigma\big(S^L X W^{(0)}\big)$. $X^{(l)} \in \mathbb{R}^{n \times f_l}$ and $W^{(l)} \in \mathbb{R}^{f_l \times f_{l+1}}$ are embeddings and weights at layer $l \in \{0, \dots, L\}$, respectively. The structural update matrix is $S = (I + D)^{-1/2}(I + A)(I + D)^{-1/2}$.

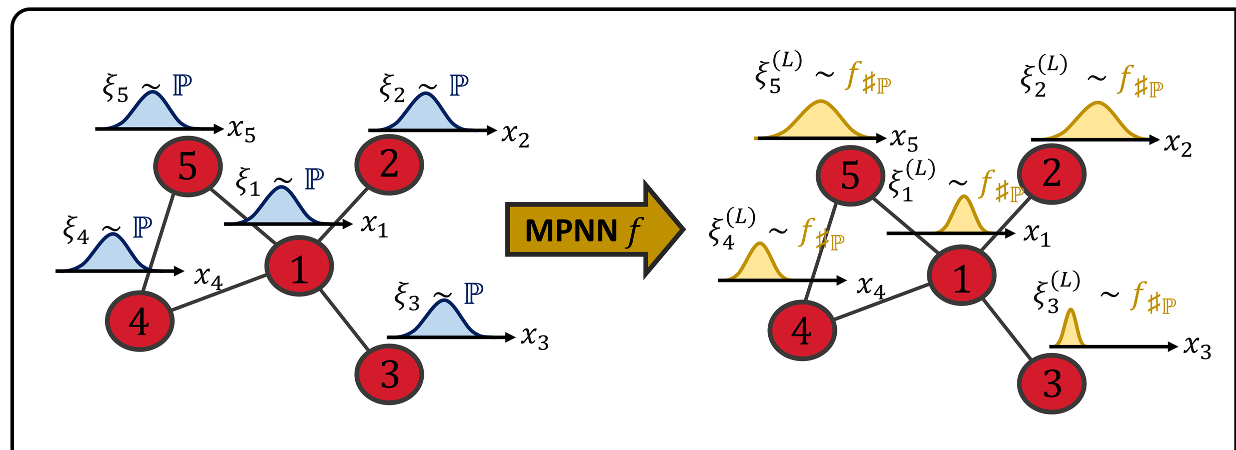

Figure 1: Schematic depiction of how a MPNN f transforms the input distribution $\mathbb{P}$ into the push-forward distribution $f_{\sharp\mathbb{P}}$.

Acknowledgements

A.C. was supported by Northeastern Universities' Network Science Institute. M.L. and T.E.R. were supported by the Inaugural Joseph E. Aoun Endowment. N.S. was supported by NSF Grant No. CCF-2311160.

Uncertainty Quantification

We develop **moment propagation** techniques based on the Taylor expansion of the nonlinearity to first order (1d-T1), to second order with moment closure (1d-T2-GC), or without it (1d-T2-Tr). We compare these techniques to the multivariate Taylor expansion of the entire architecture to first (T1) or second (T2-Tr) order, and to pseudo-Taylor polynomial expansion (PTPE) [5].

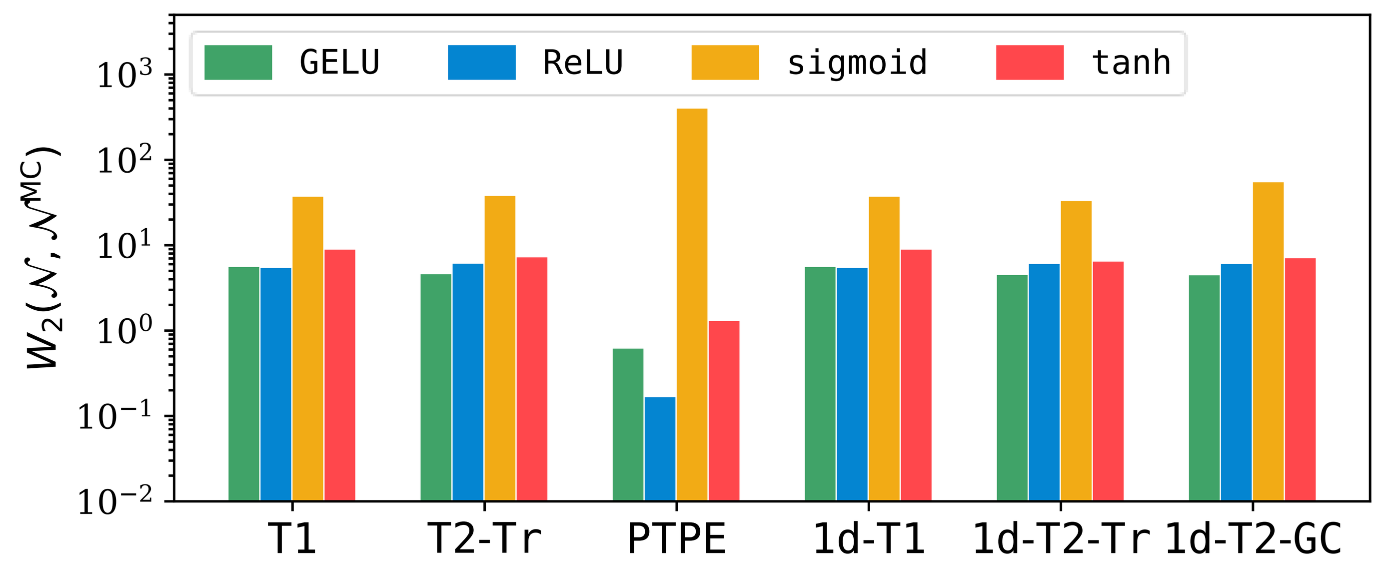

Figure 2: On the Cora dataset, the Wasserstein distance between Gaussian distributions with ground-truth mean and covariance obtained via Monte Carlo sampling and approximate mean and variance obtained through different moment propagation techniques varies depending on the non-linearity and the moment propagation technique used.
➔ *Layer-wise techniques (*1d-T1, 1d-T2-Tr, 1d-T2-GC, PTPE*) offer good performance and are often computationally cheaper than full Taylor expansion.*

Robustness Certification

Theorem: ***Robustness to node feature perturbations***

Let $\mu_{i,y}^z$ and $\Sigma_{i,y}^z$ be *the mean and variance of the random margin between the logits of true class $y^\star$ and any other class y, then a MPNN with Lipschitz constant $\bar{C}$ for node classification is, with probability $1-\delta$, robust to feature* perturbations with L_2-norm ε if

$$\varepsilon < \min_{y \neq y^\star}\left(\mu_{i,y}^z - \sqrt{\Sigma_{i,y}^z}\sqrt{(1-\delta)/\delta}\right)\Big/\left(2\sqrt{\bar{C}}\right),$$

with $\delta = \sum_y \delta_y$ denoting the sum of class-wise misclassification probabilities δ_y.

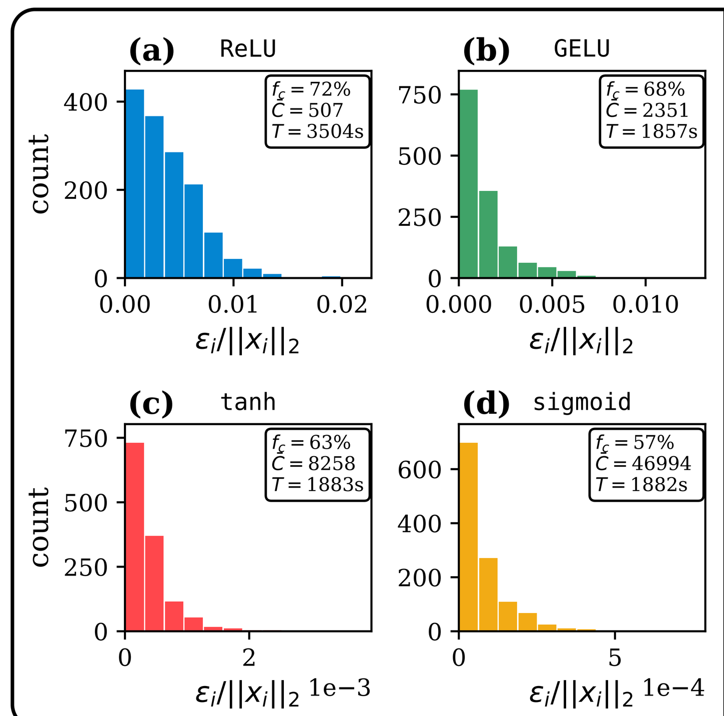

Figure 3: The certified radius ε_i of node i relative to the magnitude of the node features x_i depends on the nonlinearity for the Cora dataset with **(a)** ReLU and **(b)** GELU showing higher robustness than **(c)** tanh and **(d)** sigmoid. The former also show a higher fraction f_c of nodes for which robustness can be certified. These trends are partially explained by differences in the L_2-norm Lipschitz constant $\bar{C}$ between models. The time T to certify all nodes, including moment propagation with 1d-T2-GC is of the same order of magnitude in all cases.
➔ *Nonlinearities strongly influence robustness to L2-norm feature perturbations.*

Feature Convolution Distance

Definition: ***Feature Convolution Distance***

Consider the random matrix of node features $\xi \in \mathbb{R}^{n \times f}$ and the L-layer structural update matrix $S^L \in \mathbb{R}^{n \times n}$. Then the *Feature Convolution Distance* FCD_p between nodes i and j is

$$\mathrm{FCD}_p(i,j) = W_p\big([S^L\xi]_i, [S^L\xi]_j\big)$$

where W_p is the Wasserstein-p distance between measures on $\mathbb{R}^f$.

Properties of FCD_p

- FCD_p is a pseudo-metric on the space of nodes with random features.
- FCD_p has the same discriminative power as the SGC architecture.
- The SGC architecture is Lipschitz continuous in FCD_p.

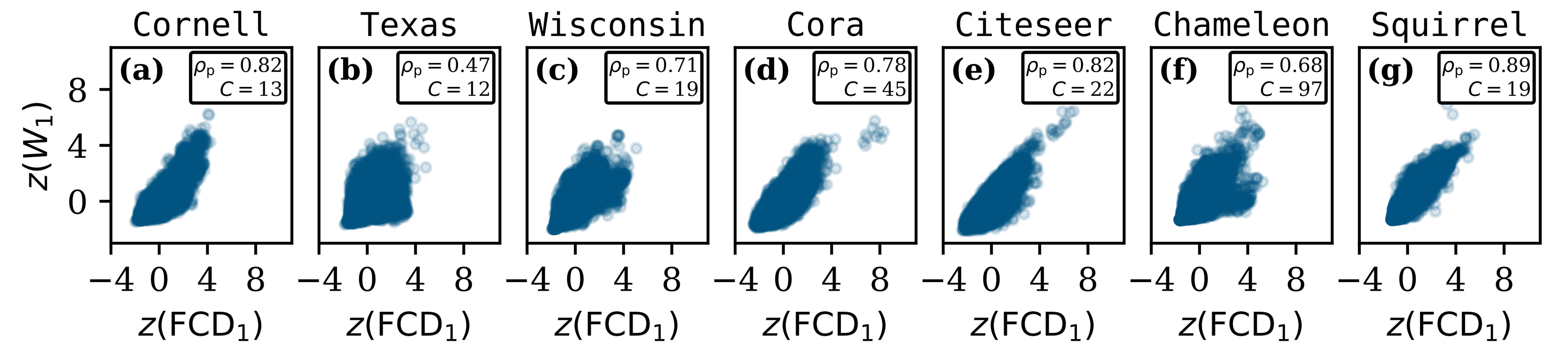

Figure 4: The FCD_p in the input space shows a high Pearson correlation ρ_p with the Wasserstein-1 distance W_1 of node embeddings from a SGC architecture with $L = 3$ layers. We display z-scores (i.e., the value minus the mean divided by the standard deviation) of both distances for seven different real-world graphs with Gaussian noise added to their features. The Lipschitz constants C are moderate.
➔ *Small changes in the input measured by FCD_p lead to small changes in the output.*

Generalization Bounds

Theorem: ***Generalization*** (informally)

Consider the SGC input space with FCD_1 and a node classification task, then the Wasserstein-1 distance W_1 between the empirical loss distribution $\widehat{\mathbb{P}}_\ell$ and the population loss distribution $\mathbb{P}_\ell$ is, with probability $1-\delta$, bounded by

$$W_1\big(\widehat{\mathbb{P}}_\ell, \mathbb{P}_\ell\big) \le 2C\varepsilon + M\sqrt{\chi/N\,((2K_\varepsilon + 1)\log 2 + 2\log 1/\delta)},$$

where C is the architectures Lipschitz constant in FCD_1, K_ε is the covering number for balls of radius ε, M is the diameter of the space of loss distributions, χ is the chromatic number of the dependency graph, and N the sample size.

References

[1] Wang, F. et al. *Uncertainty in graph neural networks.* TMLR (2024)
[2] Ju, W. et al. *A survey of graph neural networks in the real world: Imbalance, noise, privacy and ood challenges.* IEEE TIPAMI (2025)
[3] Kipf, T. & Welling, M. *Semi-supervised classification with graph convolution networks.* ICLR (2017)
[4] Wu, F. et al. *Simplifying graph convolutional networks.* ICML (2019)
[5] Zhang, S. & Ching, S. *A stochastic polynomial expansion for uncertainty propagation through networks.* TMLR (2025)