A Random Matrix Approach to Graph Convolutional Networks

Zhenyu Liao, Ming Li

Huazhong University of Science & Tech, Zhejiang Normal University

Date: February 10, 2023

1 Introduction and system model

Graph convolutional networks (GCNs) proposed by Kipf and Welling (2017) is an efficient way to mine graph-type data. Its theoretical properties, however, remain less clear.

Consider an undirected graph G having n nodes, with symmetric adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$. A GCN is constructed by considering the following layer-wise propagation rule: The output feature $\mathbf{\Sigma}_{\ell+1} \in \mathbb{R}^{d_{\ell+1}\times n}$ of layer $\ell+1$ is given by

$$\mathbf{\Sigma}_{\ell+1} = \sigma\left(\mathbf{W}_\ell \mathbf{\Sigma}_\ell \mathbf{D}_\gamma^{-\alpha} \mathbf{A}_\gamma \mathbf{D}_\gamma^{-\alpha}\right) \equiv \sigma\left(\mathbf{W}_\ell \mathbf{\Sigma}_\ell \mathbf{L}_\alpha\right), \tag{1}$$

with $\mathbf{\Sigma}_\ell \in \mathbb{R}^{d_\ell \times n}$ the input feature of layer $\ell+1$ (with the convention that $\mathbf{\Sigma}_0 = \mathbf{X} \in \mathbb{R}^{d_0\times n}$), (trainable) output weights $\mathbf{W}_\ell \in \mathbb{R}^{d_{\ell+1}\times d_\ell}$ of layer ℓ,

$$\mathbf{L}_{\alpha,\gamma} \equiv \mathbf{D}_\gamma^{-\alpha} \mathbf{A}_\gamma \mathbf{D}_\gamma^{-\alpha}, \tag{2}$$

the so-called α-normalized Laplacian of $\mathbf{A}$ Ali and Couillet (2018), and

$$\mathbf{A}_\gamma \equiv \mathbf{A} + \gamma \mathbf{I}_n \in \mathbb{R}^{n\times n}, \tag{3}$$

the adjacency with added self-connections of strength $\gamma \in \mathbb{R}$, diagonal $\mathbf{D}_\gamma \equiv \operatorname{diag}(\mathbf{d}_\gamma) \in \mathbb{R}^{n\times n}$ that contains the degrees $\mathbf{d}_\gamma \equiv \mathbf{A}_\gamma \mathbf{1}_n \in \mathbb{R}^n$ of $\mathbf{A}_\gamma$ so that $\mathbf{D}_\gamma = \mathbf{D} + \gamma \mathbf{I}_n$, for $\mathbf{D} \equiv \operatorname{diag}(\mathbf{d})$ containing the degree of $\mathbf{A}$.

Under the propagation rule in (1), the n-dimensional output of the layer ℓ of all n nodes is given by $\boldsymbol{\beta}^\mathsf{T} \mathbf{\Sigma}_\ell \mathbf{L}_{\alpha,\gamma} \in \mathbb{R}^n$, for $\boldsymbol{\beta} \in \mathbb{R}^{d_\ell}$ the output weight vector. We consider the ridge regressor $\boldsymbol{\beta}_r \in \mathbb{R}^{d_\ell}$ that minimizes the following MSE

$$L(\boldsymbol{\beta}) = \|\mathbf{y}^\mathsf{T} \operatorname{diag}(\mathbf{1}_{\rm train}) - \boldsymbol{\beta}^\mathsf{T} \mathbf{\Sigma}_\ell \mathbf{L}_{\alpha,\gamma} \operatorname{diag}(\mathbf{1}_{\rm train})\|^2 + r\|\boldsymbol{\beta}\|^2, \tag{4}$$

with target/label vector $\mathbf{y} \in \mathbb{R}^n$ associated to $\mathbf{X} \in \mathbb{R}^{d_0\times n}$ composed of both training $\mathbf{y}_{\rm train} \in \mathbb{R}^{n_{\rm train}}$ and test (unknown) $\mathbf{y}_{\rm test} \in \mathbb{R}^{n_{\rm test}}$ as

$$\mathbf{y}^\mathsf{T} = \begin{bmatrix} \mathbf{y}_{\rm train}^\mathsf{T} & \mathbf{y}_{\rm test}^\mathsf{T} \end{bmatrix}, \tag{5}$$

and $\mathbf{1}_{\rm train} \in \mathbb{R}^n$ the indicator vector of labelled nodes, with its ith entry given by $[\mathbf{1}_{\rm train}]_i = \delta_{\mathbf{x}_i \in \mathbf{X}_{\rm train}}$. The solution to (4) is explicitly given by

$$\boldsymbol{\beta}_r = \left(\boldsymbol{\Sigma}_\ell \mathbf{L}_{\alpha,\gamma} \operatorname{diag}(\mathbf{1}_{\text{train}}) \mathbf{L}_{\alpha,\gamma} \boldsymbol{\Sigma}_\ell^\mathsf{T} + r \mathbf{I}_{d_\ell} \right)^{-1} \boldsymbol{\Sigma}_\ell \mathbf{L}_{\alpha,\gamma} \operatorname{diag}(\mathbf{1}_{\text{train}}) \mathbf{y}. \tag{6}$$

2 Research objective

We aim to characterize the asymptotic behavior of the following quantities

1. Training MSE $E_{\text{train}} = \|\mathbf{y}^\mathsf{T} \operatorname{diag}(\mathbf{1}_{\text{train}}) - \boldsymbol{\beta}_r^\mathsf{T} \boldsymbol{\Sigma}_\ell \mathbf{L}_{\alpha,\gamma} \operatorname{diag}(\mathbf{1}_{\text{train}})\|^2$ with $\boldsymbol{\beta}_r$ given in (6).
2. Test MSE $E_{\text{test}} = \|\mathbf{y}^\mathsf{T} \operatorname{diag}(\mathbf{1}_{\text{test}}) - \boldsymbol{\beta}_r^\mathsf{T} \boldsymbol{\Sigma}_\ell \mathbf{L}_{\alpha,\gamma} \operatorname{diag}(\mathbf{1}_{\text{test}})\|^2$ with $\boldsymbol{\beta}_r$ given in (6).
3. And possibly the misclassification error rate on a test datum if we put additional statistical assumption on the data.

We propose to perform theoretical studies of the above GCN model under the following assumptions:

1. High-dimensional asymptotics: the graph size n and the dimension d_ℓ, $\ell = 0, \ldots$, all go to infinity.

2.1 The underlying graph follows a Contextual Stochastic Block Model (CSBM) in the sense that, or

2.2 The underlying graph follows a Degree-Correlated Stochastic Block Model (DC-SBM), or the hierarchical stochastic block model, or the mixed-membership block model, and the data at each node is deterministically given and satisfy $\|\mathbf{X}\| < \infty$.

We wish the proposed theoretical results could shed some novel light on

1. double descent on GCN
2. optimal choice of α-normalized Laplacian in GCN
3. impact of nonlinear activation in nonlinear GCN
4. heterophily, homophily versus positive and negative self-loops in GCNs
5. over-squashing and over-smoothing in deep GCNs
6. invariance and/or equivalence properties of GCNs Keriven and Peyré (2019)

We would like to cover both the shallow nonlinear case and the deep linear case with random weights, as well as a possible extension to deep nonlinear setting via the NTK approach Du et al. (2019).

3 Related reading

1. Kipf and Welling (2017): the original paper that introduces GCNs;
2. Shi et al. (2022) on a statistical mechanics approach to large-dimensional analysis of GCNs;
3. Esser et al. (2021); Keriven et al. (2020); Li et al. (2018); Verma and Zhang (2019); Zhou and Wang (2021) for some statistical learning theory of GCN model
4. Cai and Wang (2020); Wu et al. (2022) on theoretical and/or empirical perspective of over-smoothing, and Di Giovanni et al. (2023); Giraldo et al. (2022); Topping et al. (2022) on over-squashing in deep GCNs

5. Ali and Couillet (2018); Couillet et al. (2016) and (Couillet and Liao, 2022, Chapter 2) for related works as well as technical tools on the topic of random matrix theory.

References

Ali, H. T. and Couillet, R. (2018). Improved spectral community detection in large heterogeneous networks. *Journal of Machine Learning Research*, 18(225):1–49.

Cai, C. and Wang, Y. (2020). A Note on Over-Smoothing for Graph Neural Networks.

Couillet, R. and Liao, Z. (2022). *Random Matrix Methods for Machine Learning*. Cambridge University Press.

Couillet, R., Wainrib, G., Sevi, H., and Ali, H. T. (2016). The Asymptotic Performance of Linear Echo State Neural Networks. *Journal of Machine Learning Research*, 17(178):1–35.

Di Giovanni, F., Giusti, L., Barbero, F., Luise, G., Lio', P., and Bronstein, M. (2023). On Over-Squashing in Message Passing Neural Networks: The Impact of Width, Depth, and Topology.

Du, S. S., Hou, K., Salakhutdinov, R. R., Poczos, B., Wang, R., and Xu, K. (2019). Graph Neural Tangent Kernel: Fusing Graph Neural Networks with Graph Kernels. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Esser, P., Chennuru Vankadara, L., and Ghoshdastidar, D. (2021). Learning Theory Can (Sometimes) Explain Generalisation in Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 27043–27056. Curran Associates, Inc.

Giraldo, J. H., Malliaros, F. D., and Bouwmans, T. (2022). Understanding the Relationship between Over-smoothing and Over-squashing in Graph Neural Networks.

Keriven, N., Bietti, A., and Vaiter, S. (2020). Convergence and Stability of Graph Convolutional Networks on Large Random Graphs. In *Advances in Neural Information Processing Systems*, volume 33, pages 21512–21523. Curran Associates, Inc.

Keriven, N. and Peyré, G. (2019). Universal Invariant and Equivariant Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kipf, T. N. and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

Li, Q., Han, Z., and Wu, X.-m. (2018). Deeper Insights Into Graph Convolutional Networks for Semi-Supervised Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Shi, C., Pan, L., Hu, H., and Dokmanić, I. (2022). Statistical Mechanics of Generalization In Graph Convolution Networks.

Topping, J., Giovanni, F. D., Chamberlain, B. P., Dong, X., and Bronstein, M. M. (2022). Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations*.

Verma, S. and Zhang, Z.-L. (2019). Stability and Generalization of Graph Convolutional Neural Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 1539–1548, New York, NY, USA. Association for Computing Machinery.

Wu, X., Chen, Z., Wang, W., and Jadbabaie, A. (2022). A Non-Asymptotic Analysis of Oversmoothing in Graph Neural Networks.

Zhou, X. and Wang, H. (2021). The generalization error of graph convolutional networks may enlarge with more layers. *Neurocomputing*, 424:97–106.