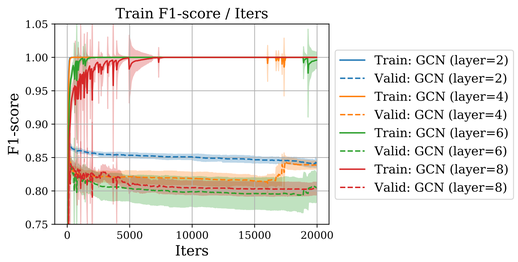
# On Provable Benefits of Depth in Training Graph Convolutional Networks

Weilin Cong, Morteza Ramezani, Mehrdad Mahdavi

## Motivation

GNNs are known to suffer from performance degradation issue as the number of layers increases, which is usually attributed to over-smoothing. However, we argue that over-smoothing does not necessarily happens in practice, a deeper model can still achieve very high training accuracy if properly trained, but generalize poorly during the evaluation stage.




### Q1: Does increasing the depth really impair the expressive power of GCNS?

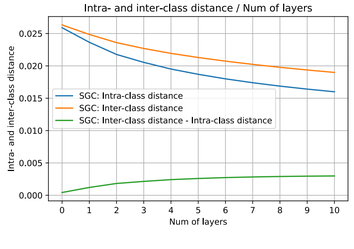
We start by reviewing existing results on over-smoothing.

Over-smoothing [1] is defined as a phenomenon where all node embeddings converge to a single vector after applying multiple graph convolution operations to the node features. However, they only consider graph convolution without non-linearity and per-layer weight matrices.

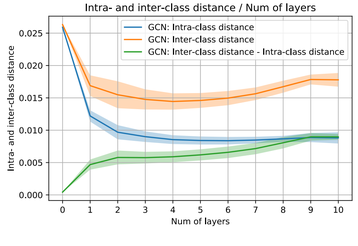
We measure the pairwise distance between node embeddings, we observe that:

- From left figure, without non-linearity and weight matrices, the pairwise distance is indeed decrease as the number of layers increases.
- However, from the right figure, if considering the weight matrices and non-linearity, the pairwise distance is increasing after a certain depth, which contradict the definition of over-smoothing.

$\mathbf{H}^{(\ell)} = \mathbf{L}\mathbf{H}^{(\ell-1)}, \mathbf{H}^{(0)} = \mathbf{X}$




$\mathbf{H}^{(\ell)} = \sigma(\mathbf{L}\mathbf{H}^{(\ell-1)}\mathbf{W}^{(\ell)}), \mathbf{H}^{(0)} = \mathbf{X}$




[2] generalize the idea of over-smoothing by takes non-linearity and weight matrices into consideration, under the notation of expressive power:

- Expressive power $d_{\mathcal{M}}(\mathbf{H}^{(\ell)})$ is measure by the distance of node embeddings $\mathbf{H}^{(\ell)}$ to a subspace $\mathcal{M}$ that only has node degree information.
- Let $\lambda_L$ as the second largest eigenvalue of Laplacian, $\lambda_W$ as the largest singular value of weight matrices.
- They show $d_{\mathcal{M}}\big(\mathbf{H}^{(\ell)}\big) \leq (\lambda_L\lambda_W)^{\ell}\, d_{\mathcal{M}}\big(\mathbf{H}^{(0)}\big)$, i.e., the expressive power will be exponentially decreasing as the number of layers increases under the assumption that $\lambda_L\lambda_W < 1$ holds.

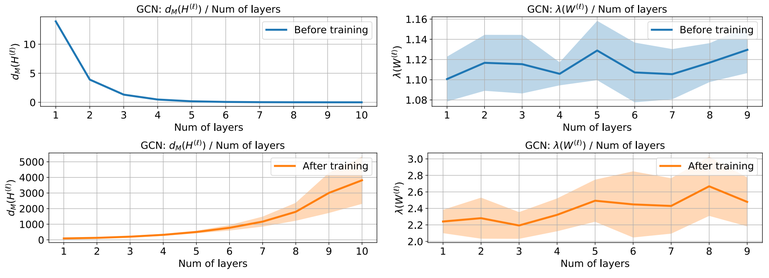
However, we argue that the assumption $\lambda_L\lambda_W < 1$ not always hold.

From the theoretical perspective:

- Let assume weight matrices $W^{(\ell)} \in \mathbb{R}^{d_{\ell-1}\times d_\ell}$ is initialized by uniform distribution $\mathcal{N}(0, \sqrt{1/d_{\ell-1}})$. By the Gordon's theorem for Gaussian matrices, we know that the expected largest singular value is bounded by $\mathbb{E}[\lambda_W] \leq 1 + \sqrt{d_\ell/d_{\ell-1}}$. This also hold for other initializations. The above discussion also hold for other initialization methods.
- Real-world graphs are sparse, $\lambda_L$ is close to 1. For example, Cora $\lambda_L$=0.9964, Citeseer $\lambda_L$=0.9987, PubMed $\lambda_L$=0.9905

Empirical test on real world dataset:

- On the untrained model, as the number of layers increases, the distance $d_{\mathcal{M}}(\mathbf{H}^{(\ell)})$ is indeed decreasing.
- However, on the trained model, as the number of layers increases, the distance $d_{\mathcal{M}}(\mathbf{H}^{(\ell)})$ is increasing, which indicate the expressive power of GNN is not always decreasing as stated in the existing theoretical analysis.




We argue that a well-trained deep GCN is at least as powerful as a shallow one:

- By leveraging the connection between GCN and WL-test, in Theorem 1, we can show deeper GCNs have stronger expressive power than the shallow GCNs.
- Furthermore, we also provides the global convergence of GCNs in Theorem 2, which shows that deeper GCNs can still converge to its global optimal with linear convergence rate.

However, it is still unclear why a deeper GCN has worse performance than a shallow one during evaluation phase.

### Q2: If GCN is expressive, why then deep GCNs generalize poorly?

To answer this question, we provide a different view by analyzing the impact of GCN strictures on the generalization.

Interestingly, we observe that

- (Theorem 4) The existing methods that originally designed to alleviate the over-smoothing issue (e.g., SGC, APPNP, GCNII, DropEdge, PairNorm) all enjoys a better generalization power than classical GCN
- (Appendix E.3 and E.4) Besides, according to our empirical results, adding DropEdge/PairNorm is actually hurting the training accuracy (i.e., not solving over-smoothing) but reduce the generalization gap, therefore leads to a better results during evaluation.

Based on our generalization analysis, we propose Decoupled GCN, with the following forward propagation rule.

$$\mathbf{Z} = \sum_{\ell=1}^{L} \alpha_\ell f^{(\ell)}(\mathbf{X}),\ \ f^{(\ell)}(\mathbf{X}) = \mathbf{P}^\ell \mathbf{X}\left(\beta_\ell \mathbf{W}^{(\ell)} + (1-\beta_\ell)\mathbf{I}\right)$$

- $\alpha_\ell, \beta_\ell$ are trainable parameters
- $\mathbf{P} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ and $\mathbf{P}^\ell$ stands for $\mathbf{P}$ to the power of $\ell$

## References

[1] Li, Qimai, Zhichao Han, and Xiao-Ming Wu. "Deeper insights into graph convolutional networks for semi-supervised learning." Thirty-Second AAAI conference on artificial intelligence. 2018.

[2] Oono, Kenta, and Taiji Suzuki. "Graph Neural Networks Exponentially Lose Expressive Power for Node Classification." International Conference on Learning Representations. 2019.

[3] Morris, Christopher, et al. "Weisfeiler and leman go neural: Higher-order graph neural networks." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.