FAQ – EFFECTS OF GRAPH CONVOLUTIONS IN MULTI-LAYER NETWORKS

Anonymous authors
Paper under double-blind review

1 HOW DOES THIS WORK COMPARE TO OVERSMOOTHING RESULTS?

Our results informally align with oversmoothing results, however, we do not provide rigorous theorems in this regard because our focus is on regimes where GNNs work better compared to an MLP, rather than regimes where the oversmoothing phenomenon is observed. A theoretical study of the oversmoothing problem in our framework requires computing the distribution of the feature representations of nodes at each layer, along with the effects of graph convolutions on this distribution. For the XOR-CSBM data model, an analysis on multi-layer networks with non-linearities requires non-trivial ideas, because the random variables (representing node features) become highly correlated when convolutions are applied in deeper layers. We leave this analysis for future work. However, for intuition, let us describe this effect for a simpler data model: the binary CSBM with means $\mu, -\mu$ for the two components, and a network with only one layer. Points from the component with mean $-\mu$ are in one class and points from the other component with mean μ are in the other class. Using Lemma A.3, we observe that in expectation, for any $K \geq 2$, the value of $\rho(K)$ is $\frac{1}{n}(1 + \Gamma(p,q)^{2K})$, where $\Gamma(p,q) = |p-q|/(p+q)$. We also note that for this simple model, a single graph convolution moves the means from $\mu, -\mu$ to $\Gamma(p,q)\mu, -\Gamma(p,q)\mu$. Thus the distance between the means is reduced by a factor of $\Gamma(p,q)^K$ after K convolutions. One can now compare the reduction in this distance with the reduction in the variance ($\rho(K)$) to obtain a condition on K in terms of n, p, q. When the distance between the means is small compared to the standard deviation, there is a large overlap between the two distributions, causing misclassification. Even for the simple binary CSBM case, this result requires stronger assumptions on the graph density to show that the value of $\rho(K)$ is close to its expectation. Since oversmoothing is not the focus of our results, we defer the complete analysis of oversmoothing in the XOR-CSBM data model for future work.

2 WHAT IS THE INTUITIVE INTERPRETATION OF THEOREM 2?

Theorem 2 identifies a critical threshold involving the product of two quantities: Γ and ζ, where $\Gamma = \frac{|p-q|}{p+q}$ represents the signal of the graph (relational data), and ζ is a function of $\frac{\|\mu-\nu\|_2}{\sigma}$ representing the signal of the Gaussian mixture (feature data). We obtain lower bounds on the product $\Gamma\zeta$ for perfect classification of all nodes. This condition encapsulates how the regime of perfect classification varies with $\frac{\|\mu-\nu\|}{\sigma}$ and $\frac{|p-q|}{p+q}$. We have also added section A.8 in the appendix, where we show how to obtain the previously stated result from the general theorem statement. In Fig. 6 (g-h), the accuracy is poor because p and q are large and very close to each other, i.e., Γ (which represents the signal in the graph) is very small. Thus, a convolution averages the features of a roughly equivalent number of nodes from both the classes. Fig. 6 (a-h) are arranged in decreasing order of Γ, which explains the degrading performance as Γ becomes smaller.

3 HOW DO THE RESULTS HOLD FOR HETEROGENOUS GRAPHS?

Our analysis holds for arbitrary p and q. Note that the ansatz we construct in Proposition A.7 consists of the factor $\mathrm{sgn}(p-q)$. The GCN architecture works the same way for the case $q > p$ as it does for $p > q$ and obtains the same variance reduction as expected. The intuition behind why our results hold for heterophilous case as well is that graph convolution is an averaging operation, and hence, to compute the feature representation of a node in class C_0, it gathers more information from nodes in C_0 than in C_1 for $p > q$, and more information from nodes in C_1 than in C_0 for $p < q$. In both cases, it performs a variance reduction on the data which is the key to the improvement in the threshold.

4 WHY ARE EXPERIMENTS ONLY FOR NETWORKS WITH UP TO 3 LAYERS?

There are two reasons for choosing to restrict ourselves to networks with up to three layers.

1. Our work is theoretical in nature, hence, from a theoretical standpoint if is sufficient to consider three-layer networks to show the effects of graph convolutions at various layers and draw a subsequent comparison.
2. The real-world datasets that we work with to demonstrate our theoretical results are known to have state-of-the-art models consisting of at most three layer networks. In fact, we show for the OGB datasets that the number of graph convolutions is what matters, instead of the number of layers.