Effects of graph convolutions in multi-layer networks

Aseem Baranwal, Kimon Fountoulakis, Aukosh Jagannath

UNIVERSITY OF WATERLOO | DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

Contributions

- Complete characterization of up to two graph convolutions (GCs) in networks with up to 3 layers
  - Improvement in the classification threshold
  - Comparison of various placement choices for convolutions
- Theoretical analysis on contextual stochastic block model (CSBM) modelled after XOR data
- Extensive experiments in various settings to illustrate our results

Graph Convolutions

- Dataset of n nodes, each node has d-dimensional features
- $X_i \in \mathbb{R}^d$ denotes features of node i
- Undirected edges between nodes denoted by adjacency matrix A
- D denotes the diagonal degree matrix

Convolved feature matrix: $\tilde{X} = D^{-1}AX$

$$\tilde{X}_i = \frac{1}{D_{ii}} \sum_{j \in [n]} a_{ij} X_j$$

Architecture

- Two sources of information: $(\mathbf{A}, \mathbf{X})$

$$\left.\begin{aligned} \mathbf{H}^{(0)} &= \mathbf{X}, \\ f^{(l)}(\mathbf{X}) &= (\mathbf{D}^{-1}\mathbf{A})^{k_l}\mathbf{H}^{(l-1)}\mathbf{W}^{(l)} + \mathbf{b}^{(l)} \\ \mathbf{H}^{(l)} &= \mathrm{ReLU}(f^{(l)}(\mathbf{X})) \end{aligned}\right\} \text{ for } l \in [L],$$

$$\hat{\mathbf{y}} = \varphi(f^{(L)}(\mathbf{X})).$$

- $\mathbf{X} \in \mathbb{R}^{n \times d} \rightarrow$ input data
- $\varphi \rightarrow$ sigmoid function
- $\hat{\mathbf{y}} \rightarrow$ output of the network
- $k_l \rightarrow$ number of GCs in layer l

- A generalization of Kipf and Welling's GCN with variable GCs at each layer
- Similar models analyzed previously with power iterations in the last layer (APPNP) or first layer (SIGN):
  - Empirically known to have comparable performance to SOTA

Data model

- Linear classifiers can be realized using one-layer NNs
- Class of one-layer NNs is too simple to capture the extent of GC effects
- Need to look at multi-layer NNs for placement questions
- Identification of the relevant SNR in the data

Data model

- Four-component XOR-based Gaussian Mixture Model (GMM) coupled with a Stochastic Block Model (SBM)

- Two classes C_0, C_1
  n data points with features $(X_i)_{i=1}^{n} \in \mathbb{R}^d$

- $X_i \sim \mathcal{N}(\pm\mu, \sigma^2 I)$ if $i \in C_0$
  $X_i \sim \mathcal{N}(\pm\nu, \sigma^2 I)$ if $i \in C_1$

- $A \sim SBM(p, q)$

$$\mathbb{P}(A_{ij} = 1) = \begin{cases} p & \text{if } i, j \text{ are in the same class} \\ q & \text{otherwise} \end{cases}$$

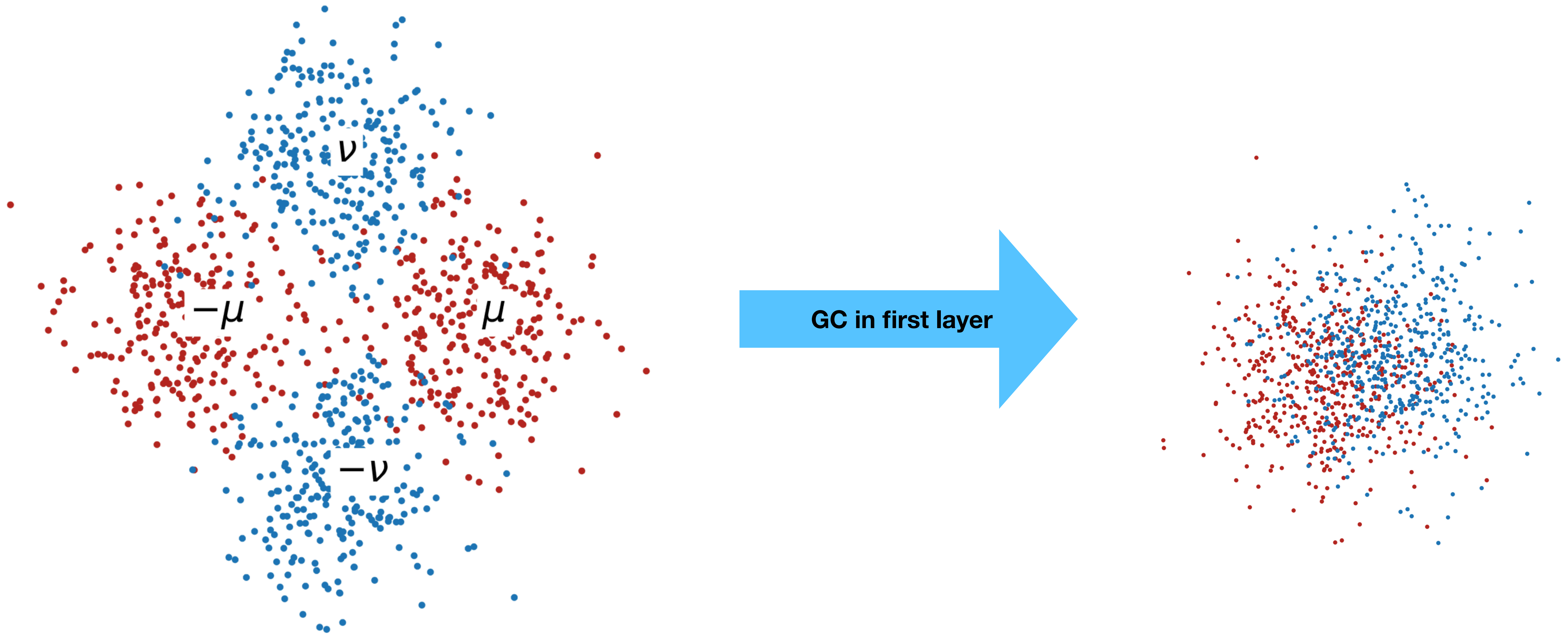
Data model

- $X_i \sim \mathcal{N}(\pm\mu, \sigma^2 I)$ if $i \in C_0$
  $X_i \sim \mathcal{N}(\pm\nu, \sigma^2 I)$ if $i \in C_1$
- $A \sim SBM(p, q)$
  $$\mathbb{P}(A_{ij} = 1) = \begin{cases} p & \text{if } i, j \text{ in same class} \\ q & \text{otherwise} \end{cases}$$

- Identified signals in data

$$\zeta = \frac{\|\mu - \nu\|}{\sigma}, \qquad \Gamma = \frac{|p - q|}{p + q}$$

Data model



Original input node features

Features after GC at the first layer

A typical two-layer GCN (one GC in each layer) performs poorly on this data

Baseline — No graph

- Characterize fraction of misclassifications in terms of GMM signal

(Fraction of errors) $f = 2\Phi_c(\zeta/2)^2$

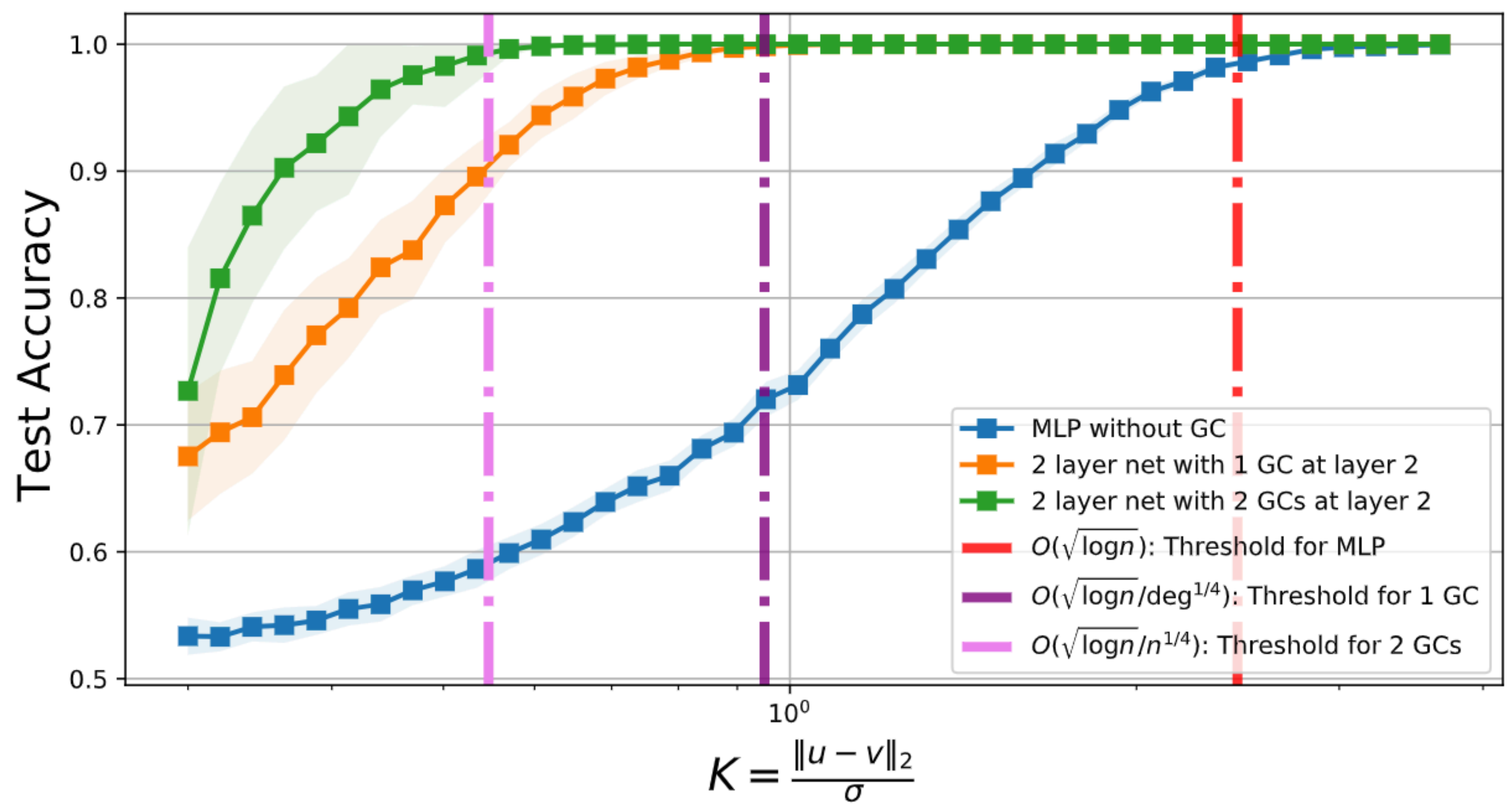
- $\zeta \to \infty \implies f \to 0$
- $\zeta \to 0 \implies f \to 1/2$
- $\zeta \to O(\sqrt{\log n}) \implies n \cdot f \to \Omega(1)$

- Conclusion: $\zeta = O(\sqrt{\log n})$ makes a constant number of mistakes. So the threshold for perfect classification should be $\zeta = \omega(\sqrt{\log n})$

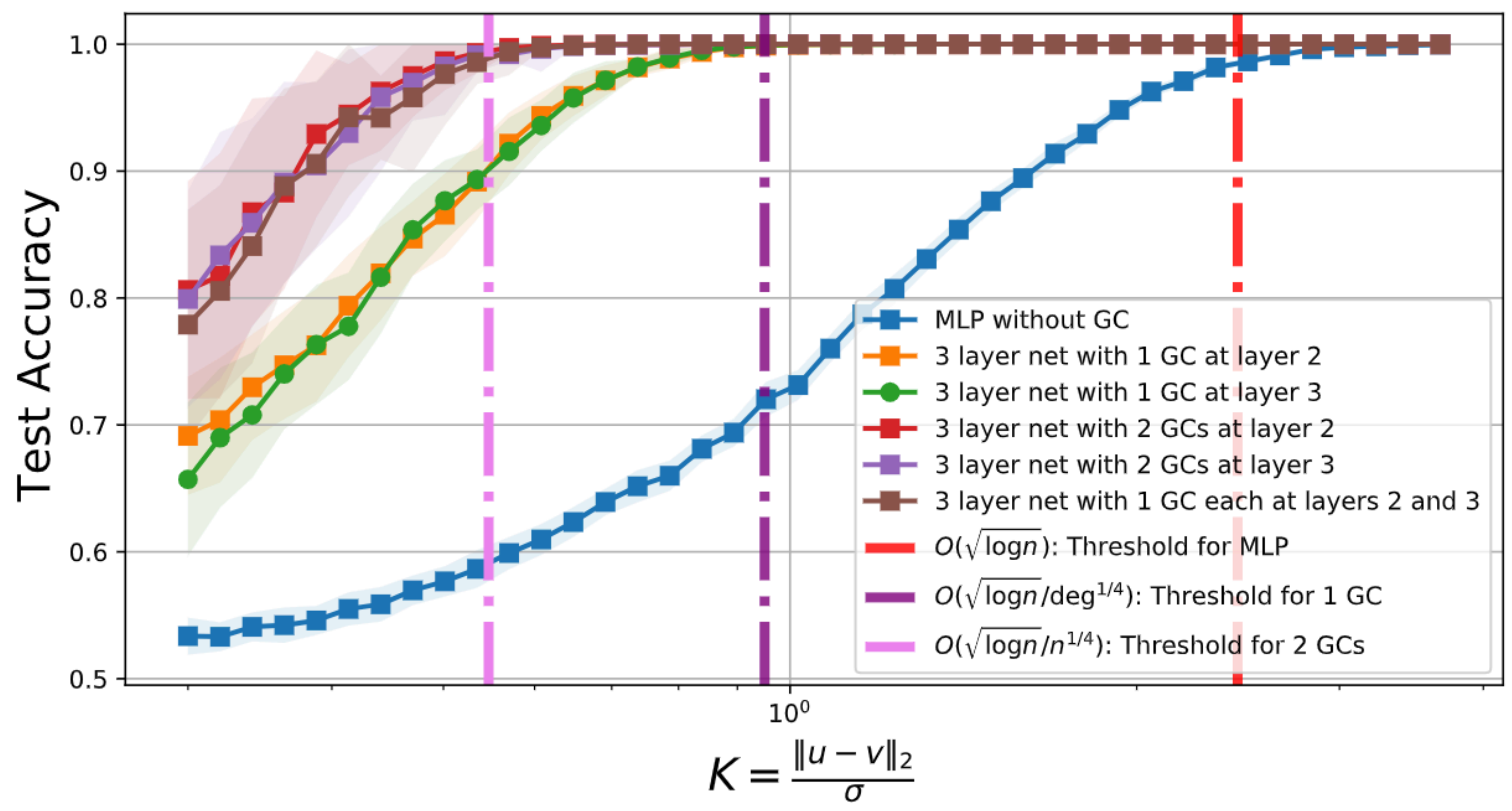
Main Result — With Graph

Number of GCs	Perfect Classification Threshold	Reference
0	$\zeta = \omega(\sqrt{\log n})$	Theorem 1
1	$\Gamma \cdot \zeta = \omega\left(\sqrt{\frac{\log n}{n(p+q)}}\right)$	Theorem 2 part 1
2	$\Gamma^2 \cdot \zeta = \omega\left(\sqrt{\frac{\log n}{n}}\right)$	Theorem 2 part 2

Main result



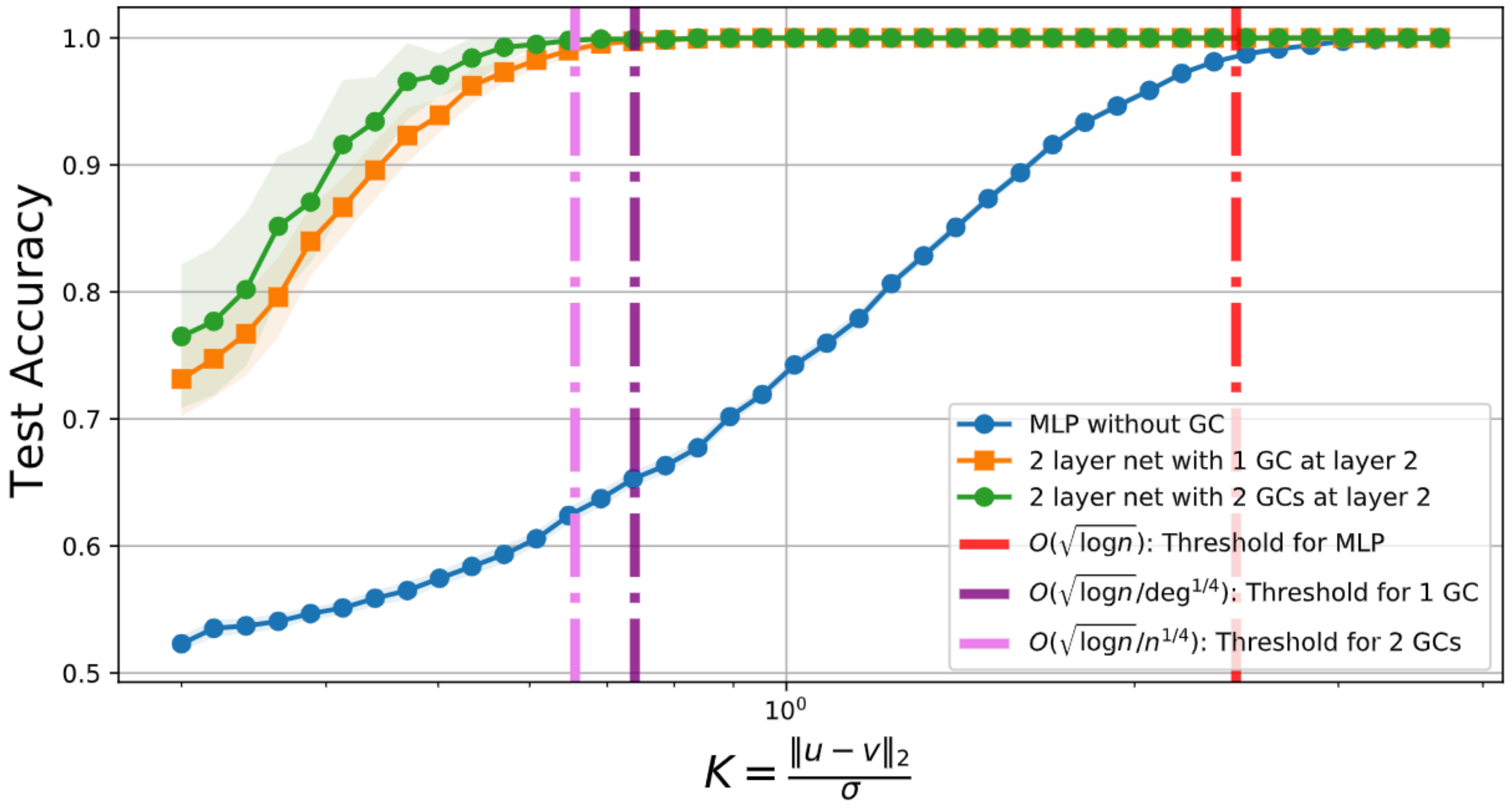
(a) Two-layer networks with $(p, q) = (0.2, 0.02)$.



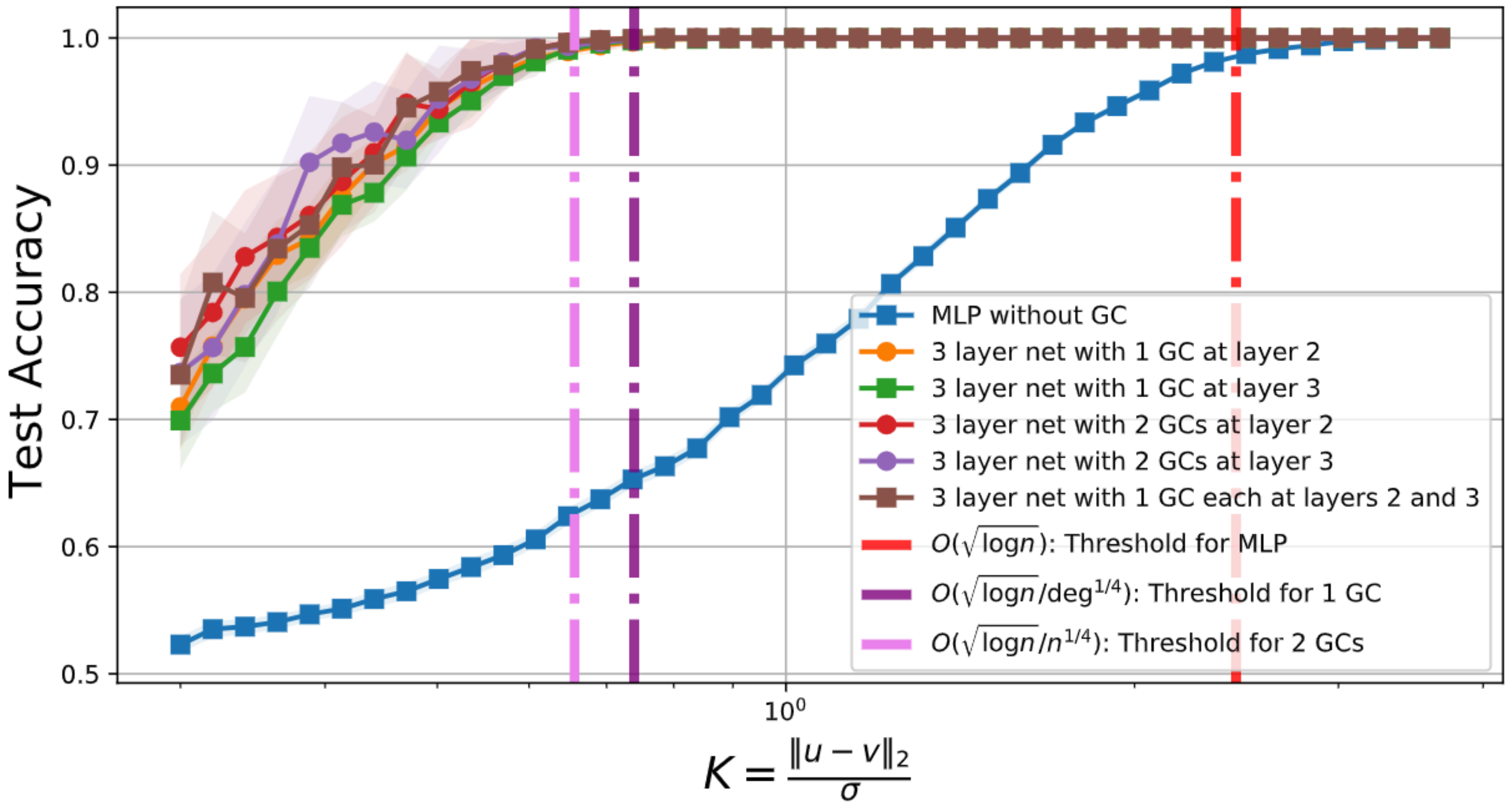
(b) Three-layer networks with $(p, q) = (0.2, 0.02)$.

Comparison of the performance of models with 1 GC vs 2 GCs

Main result



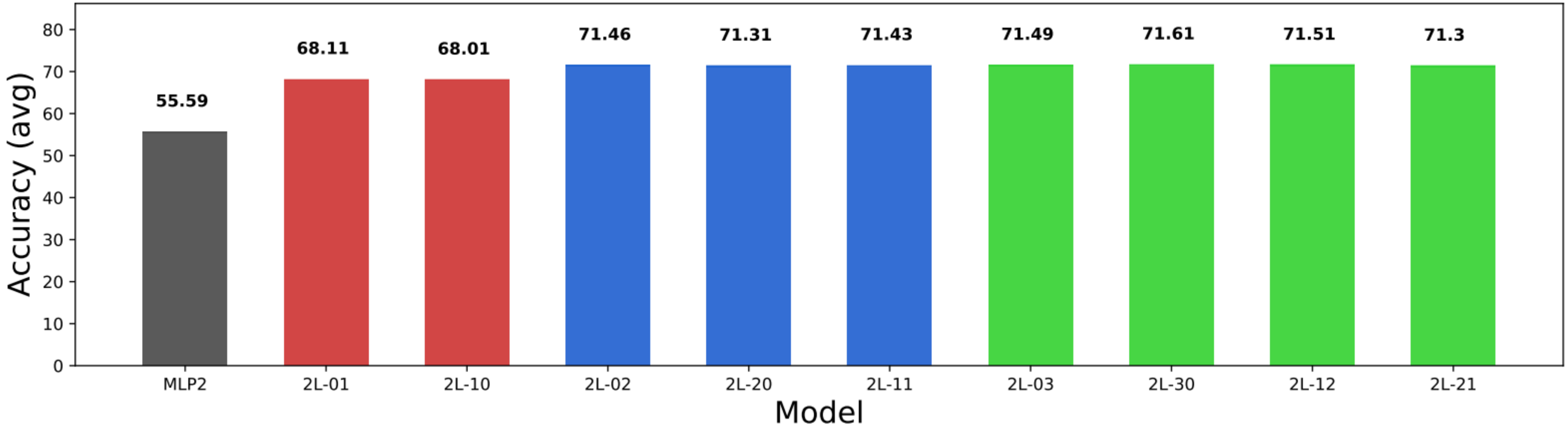
(c) Two-layer networks with $(p, q) = (0.5, 0.1)$.



(d) Three-layer networks with $(p, q) = (0.5, 0.1)$.

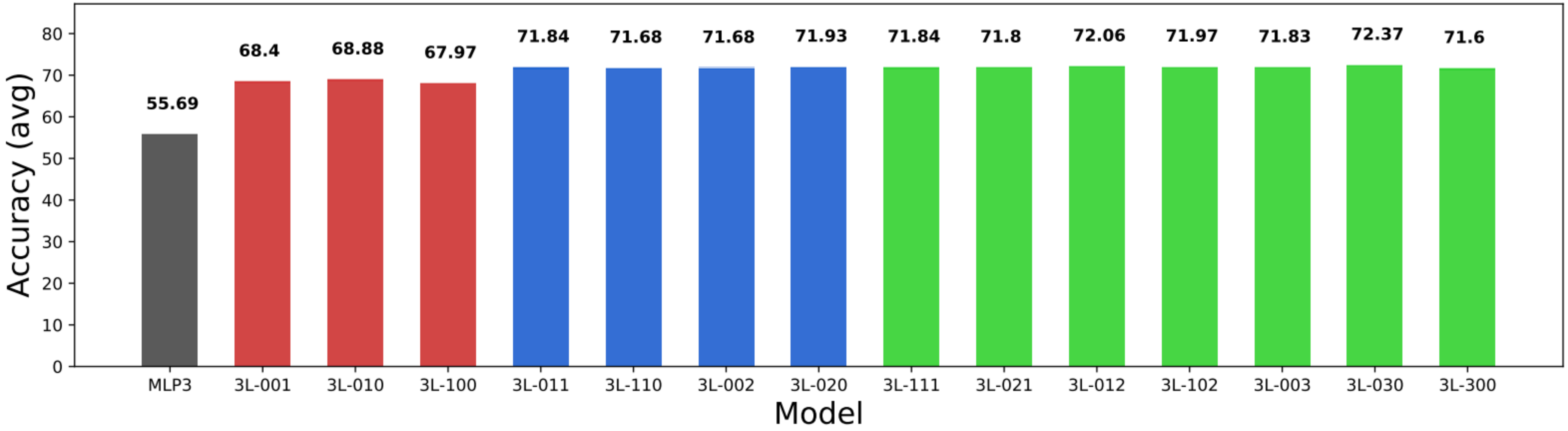
Comparison of the performance of models with 1 GC vs 2 GCs

Experiments on real data



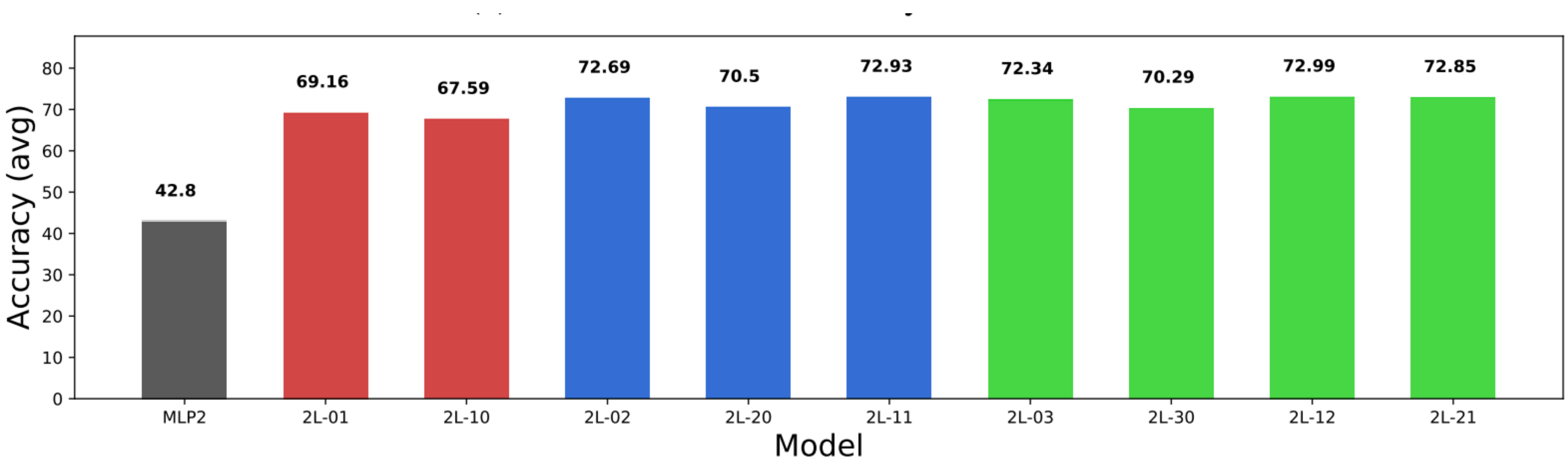
2-layer models on OGBN-ARXIV

Experiments on real data



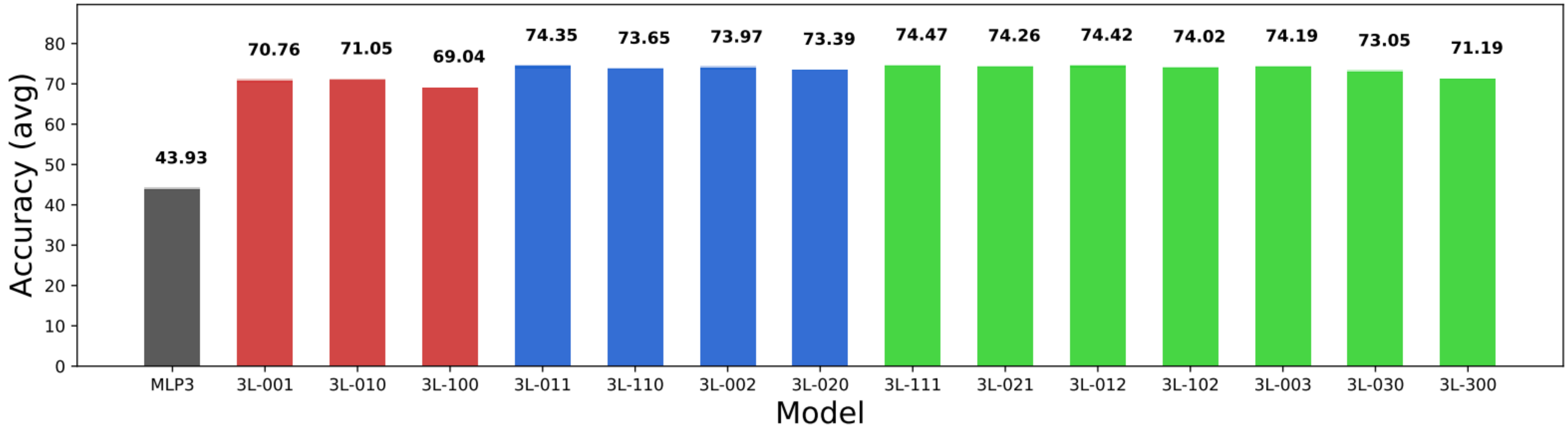
3-layer models on OGBN-ARXIV

Experiments on real data



2-layer models on OGBN-PRODUCTS

Experiments on real data



3-layer models on OGBN-PRODUCTS

Conclusions

- Theoretical characterization of the capacity of GCs placed across different layers of an MLP
- High-probability classification guarantees in terms of signals in the data
- Any combination of the placement of GCs in an MLP achieves similar performance if number of GCs is the same
- 2 GCs are better than 1 GC only when the graph is relatively sparser