Optimality of message-passing architectures for sparse graphs

Aseem Baranwal, Kimon Fountoulakis, Aukosh Jagannath

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

Contributions

- Study of node classification on sparse feature-decorated graphs on a fairly general statistical data model
- Define a notion of asymptotically local Bayes optimality
- Optimal classifier is realizable via a message-passing GNN architecture
- Generalization error bounds in terms of recognizable SNR in the data
- Empirical demonstration and comparison with other architectures

Data Model

n = # of nodes

d = # of features per node

$\{y_u\}_{u\in[n]}$ = class labels

C = # of classes

<u>Graph Component</u>

$A = (a_{uv})_{u,v\in[n]} \sim \mathrm{SBM}(n, Q)$

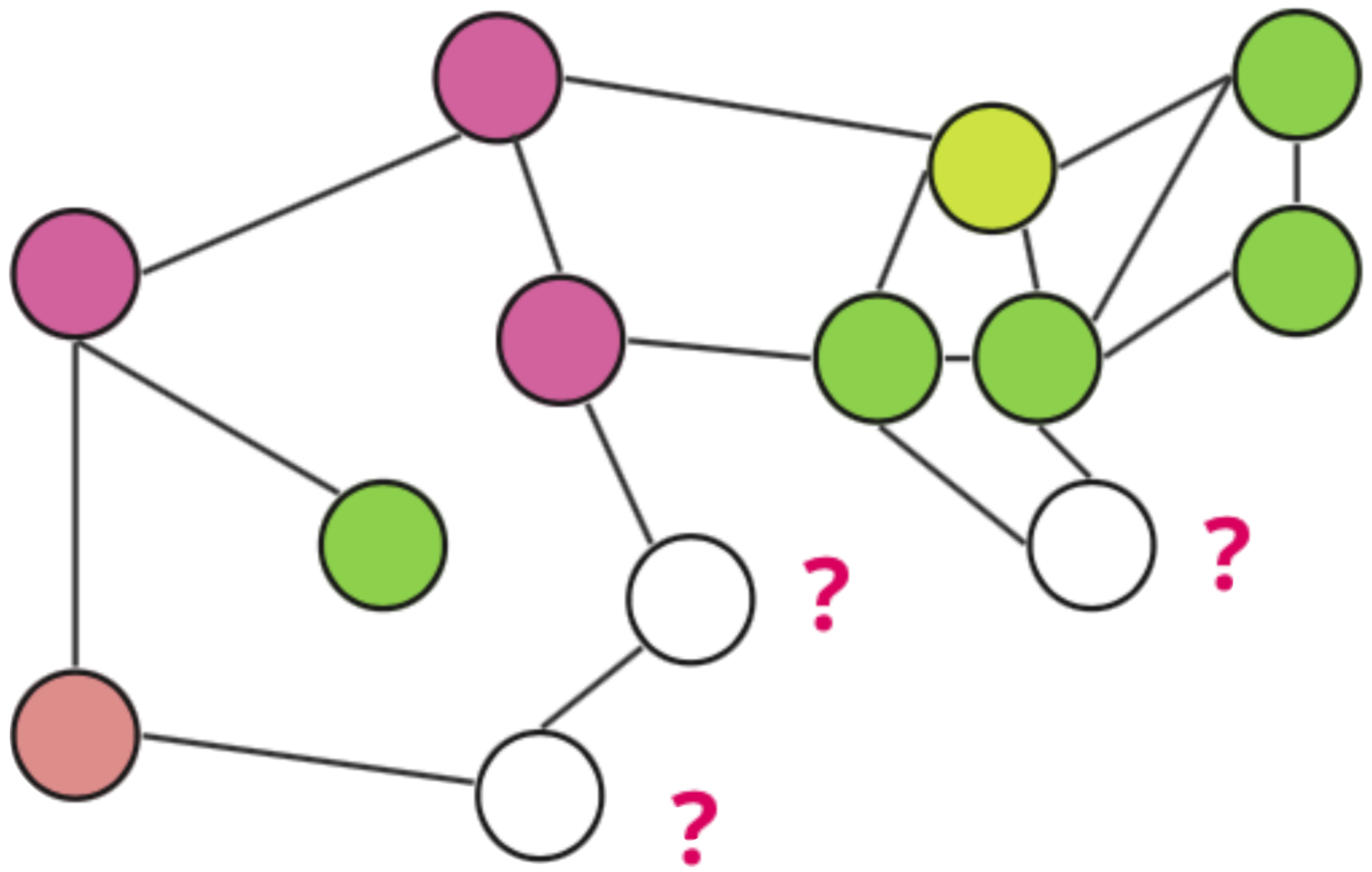
$\Pr(a_{uv} = 1 \mid y_u, y_v) = q_{y_u y_v}$

$q_{ij} = O(1/n)$

<u>Node features</u>

$X_u \sim \mathbb{P}_{y_u} \in \mathbb{R}^d$

$\mathbb{P}_c$ = Feature distribution for class c



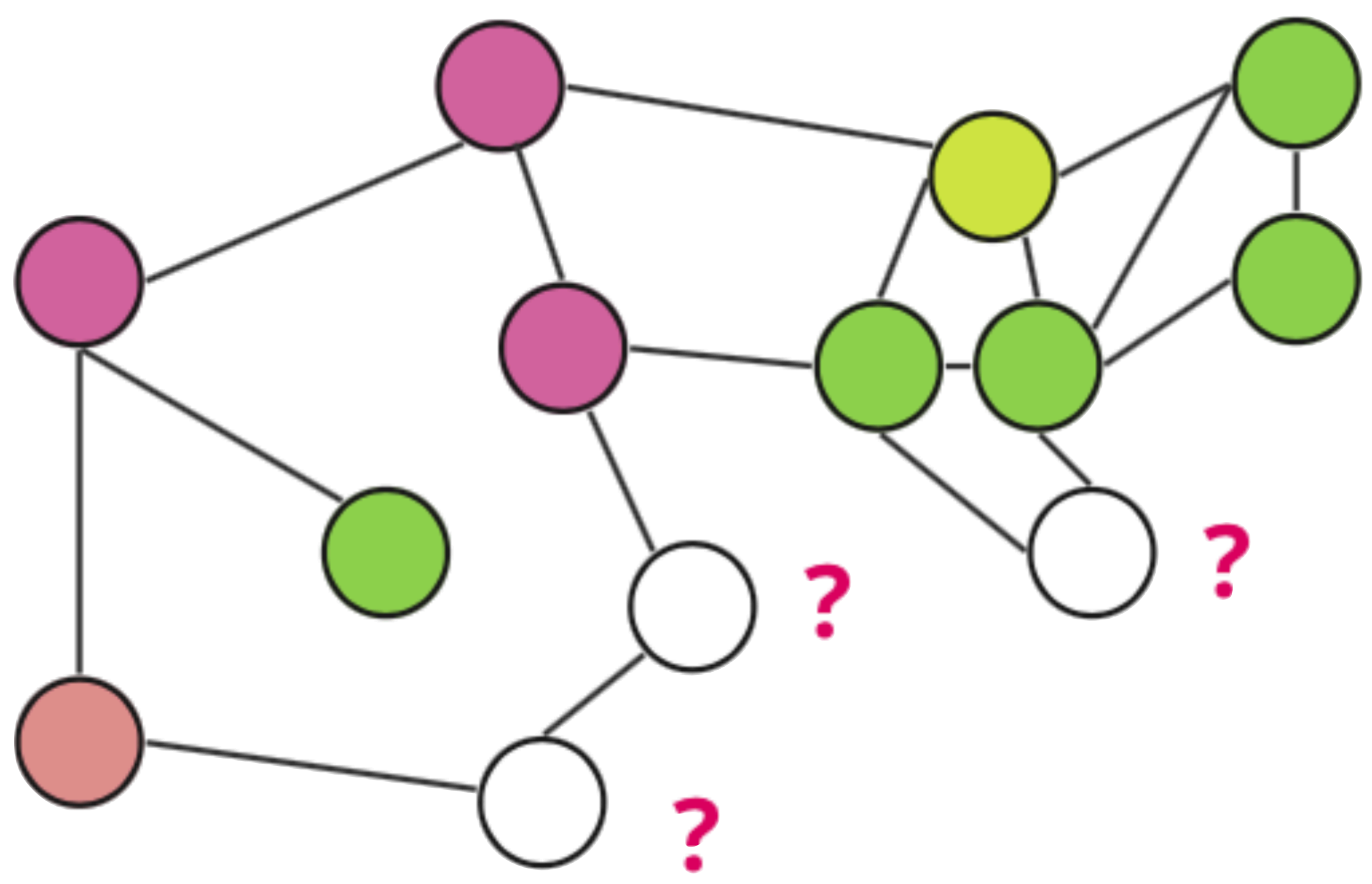
Data Model

$G_n \sim \text{CSBM}(n, \mathbb{P}, Q)$ denotes a feature-decorated graph from this model with:

- Adjacency matrix $A \in \{0,1\}^{n \times n}$
- Node features $X \in \mathbb{R}^{n \times d}$

u_n denotes a uniform at random node in G_n

(G_n, u_n) denotes a graph rooted at node u_n



Question: Given $\mathbb{P}, Q$ and a root $u_n \in V(G_n)$ along with its local neighbourhood information, how to define the notion of an "optimal classifier" for the model?

ℓ-local Classifier

Denoted $f(u, \eta_\ell(u), \{X_v\}_{v \in \eta_\ell(u)})$

Input:

- A root node u
- Subgraph induced by $\eta_\ell(u)$, the ℓ-hop neighbourhood of u
- Features $\{X_v\}\ \forall v \in \eta_\ell(u)$

Output: a class label prediction $\hat{y}_u$ for u

$\mathscr{C}_\ell$ denotes the class of all ℓ-local classifiers.

Local Weak Convergence

For a uniform at random root node u_n, the sequence of rooted graphs from this model converges locally weakly:

$$(G_n, u_n) \xrightarrow{LWC} (G, u).$$

- (G, u) is a feature-decorated Poisson Galton-Watson tree.
- Roughly speaking, in the limit $n \to \infty$ the local neighbourhood of a uniform at random node behaves like the local neighbourhood of a Poisson Galton-Watson tree.

Optimal Classifier

We say h_ℓ^* is the asymptotically ℓ-locally Bayes optimal classifier of the root of the sequence $\{(G_n, u_n)\}$ if it minimizes the misclassification probability of the root of the local weak limit (G, u) over $\mathscr{C}_\ell$.

Theorem

$$h_\ell^*(u, \eta_\ell(u), \{X_v\}_{v \in \eta_\ell(u)}) = \underset{i \in [C]}{\operatorname{argmax}} \left\{ \log \mathbb{P}_i(X_u) + \sum_{v \in \eta_\ell(u) \setminus \{u\}} M_{i\, d(u,v)}(X_v) \right\}$$

$$M_{ik}(x) = \max_{j \in [C]} \left\{ \log \mathbb{P}_j(x) + \log q_{ij}^k \right\}$$

GNN Architecture

$$H^{(0)} = X,$$

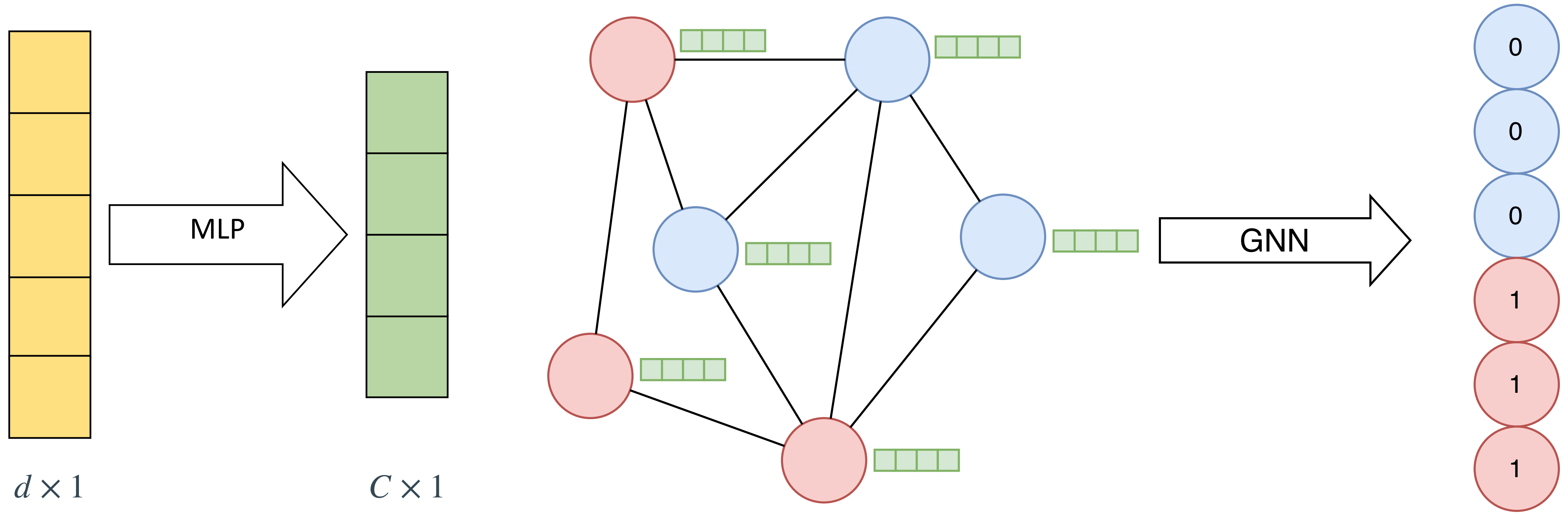
$$H^{(l)} = \sigma_l(H^{(l-1)}W^{(l)} + \mathbf{1}_n b^{(l)}) \text{ for } l \in [L],$$

$$Q = \text{sigmoid}(Z),$$

$$M^{(k)}_{u,i} = \max_{j \in [C]} \left\{ H^{(L)}_{u,j} + \log(Q^k_{i,j}) \right\} \text{ for } k \in [\ell], u \in [n], i \in [C]$$

$W^{(l)}$, $b^{(l)}$ for $l \in [L]$ and Z are learnable parameters of the model.

$$\hat{y}_u = \underset{i \in [C]}{\text{argmax}} \left(H^{(L)}_{u,c} + \sum_{k=1}^{\ell} \tilde{A}^{(k)}_{u,:} M^{(k)}_{:,i} \right)$$



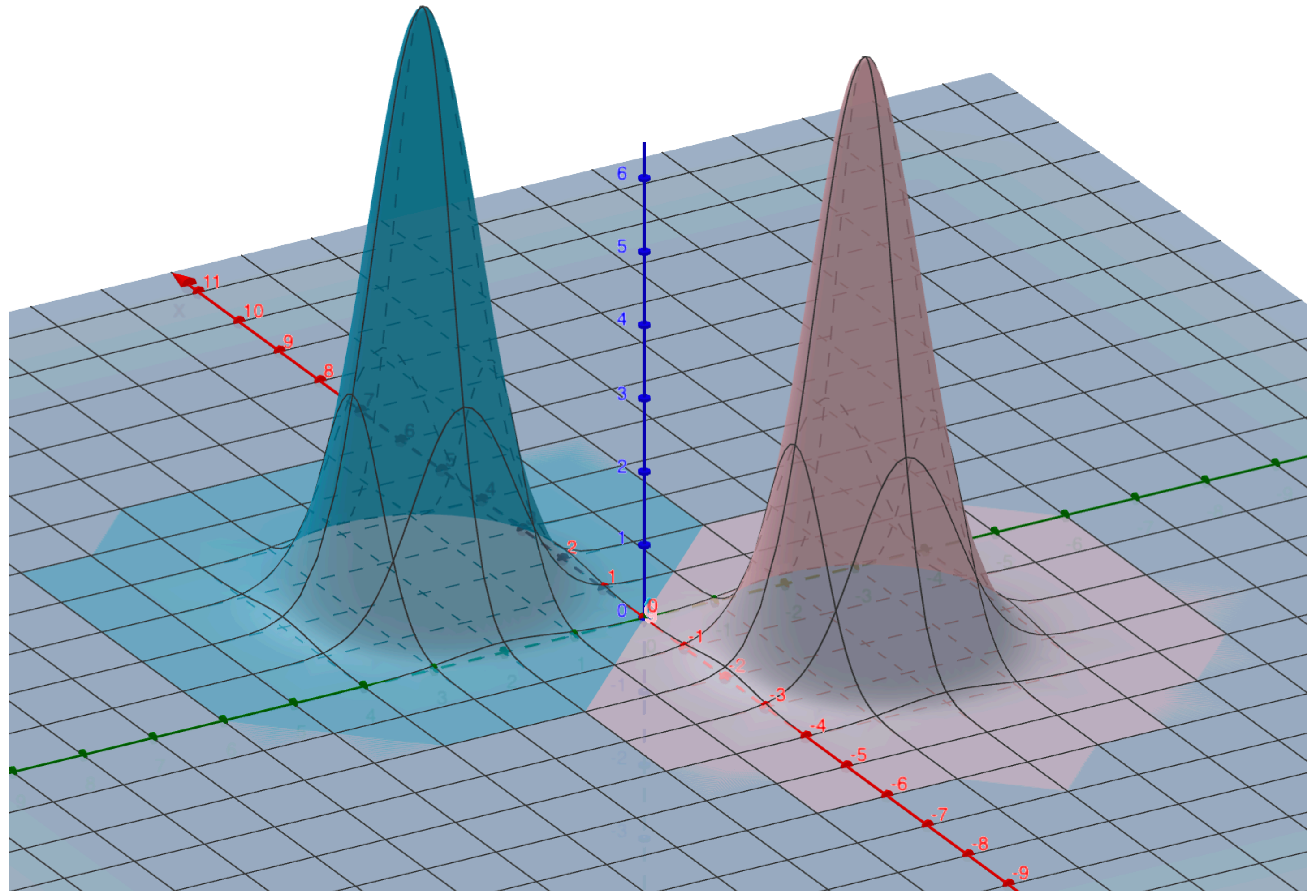
Example

$$\mathbb{P} = \{\mathcal{N}(\pm\mu, \sigma^2 I)\}$$

Graph signal $\Gamma = \dfrac{a-b}{a+b}$

$$Q = \frac{1}{n}\begin{pmatrix} a & b \\ b & a \end{pmatrix}$$

Feature signal $\gamma = \dfrac{2\|\mu\|}{\sigma}$



$$h^*_\ell(u, \{X_v\}_{v\in\eta_\ell(u)}) = \mathrm{sgn}\Big(\langle X_u, \mu\rangle + \sum_{v\in\eta_\ell(u)\setminus\{u\}} M_{d(u,v)}(X_v)\Big)$$

$$M_k(x) = \mathrm{sgn}(a-b)\cdot \mathrm{CLIP}(\langle x, \mu\rangle, \pm c_k), \qquad c_k = \log\left(\frac{1+\Gamma^k}{1-\Gamma^k}\right)$$

Results

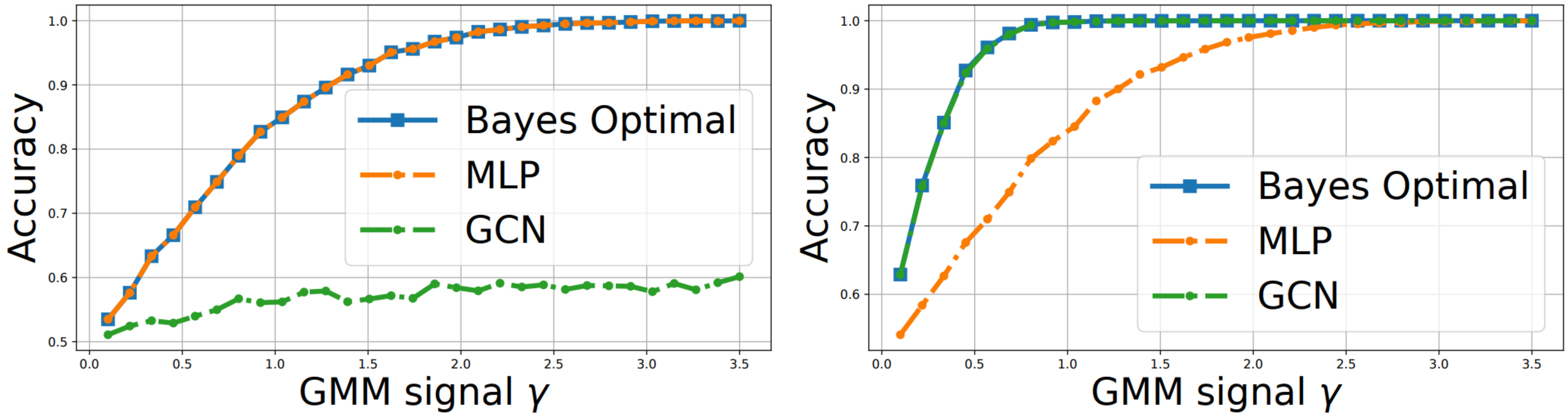
$$h_\ell^*(u, \{X_v\}_{v \in \eta_\ell(u)}) = \mathrm{sgn}\Big(\langle X_u, \mu \rangle + \sum_{v \in \eta_\ell(u) \setminus \{u\}} M_{d(u,v)}(X_v) \Big)$$

$$M_k(x) = \mathrm{sgn}(a - b) \cdot \mathrm{CLIP}(\langle x, \mu \rangle, \pm c_k), \qquad c_k = \log\left(\frac{1 + \Gamma^k}{1 - \Gamma^k}\right)$$

Theorem

- When $\Gamma \to 0$, h_ℓ^* ignores all messages and collapses to a simple MLP.
- When $\Gamma \to 1$, h_ℓ^* collapses to a typical GCN.
- When $\Gamma \in (0,1)$, h_ℓ^* interpolates and is superior to MLP and GCN.

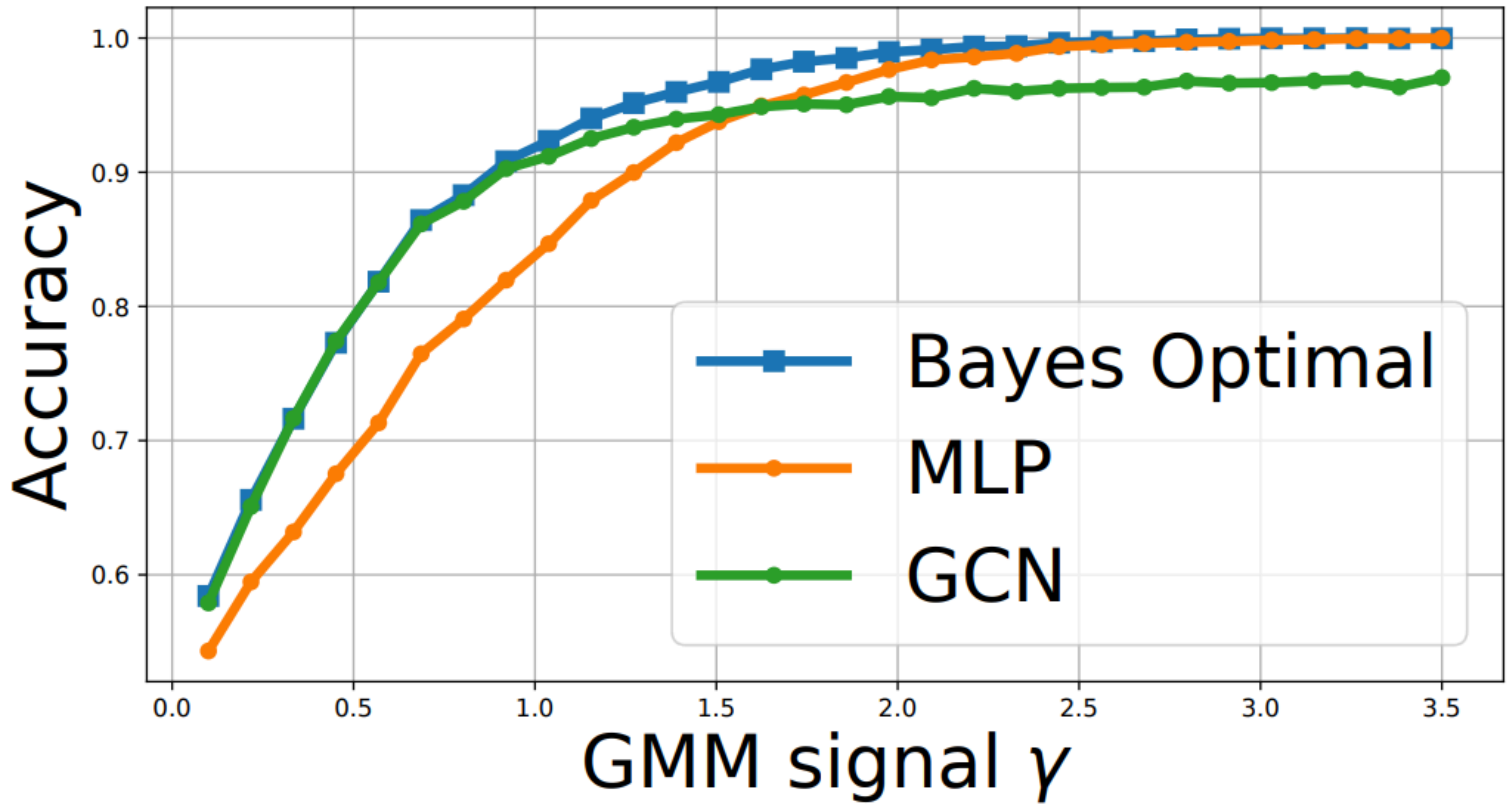
Results



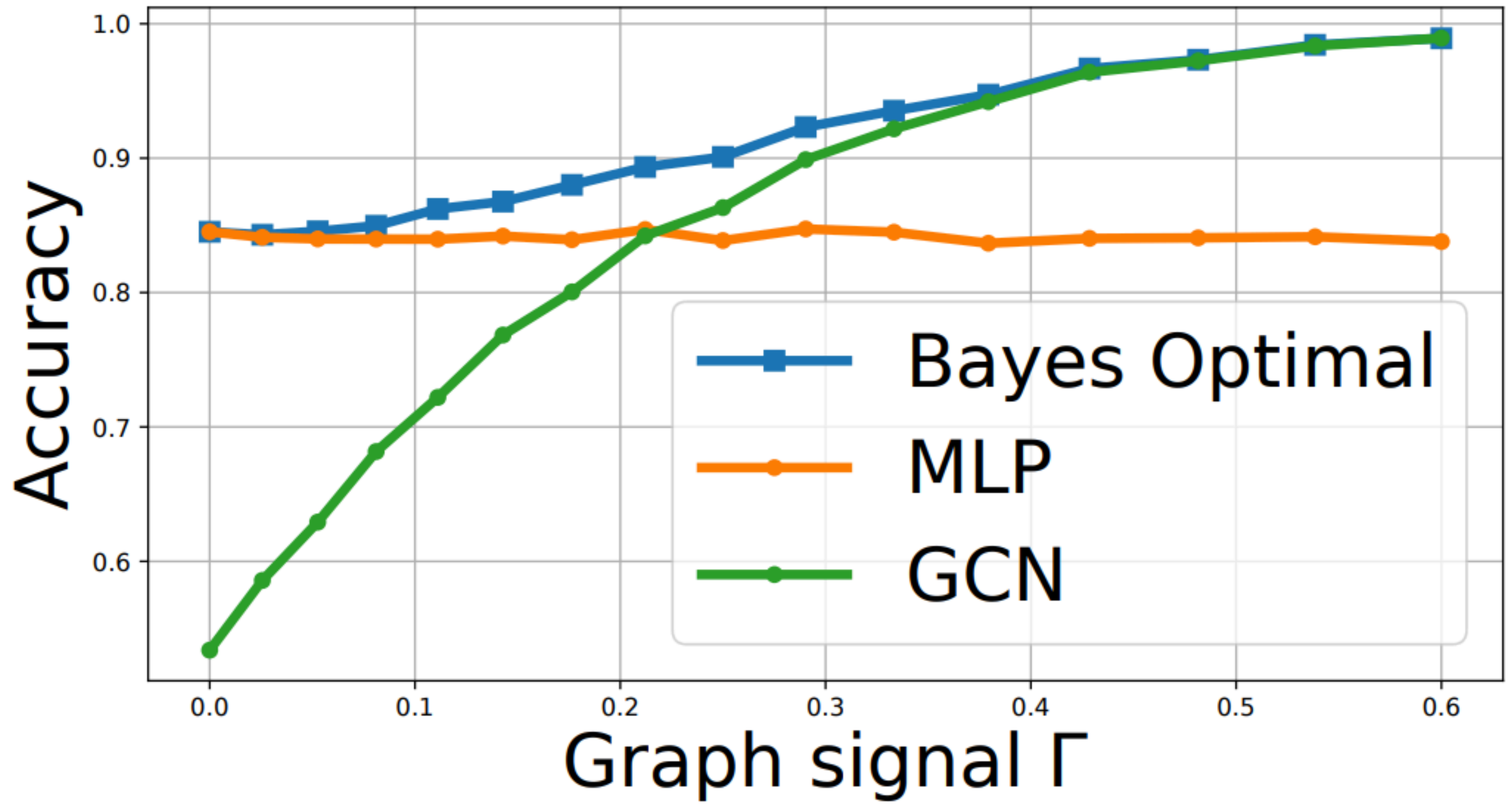
(a) Fixed graph signal $\Gamma = 0$.

(b) Fixed graph signal $\Gamma = 1$.

Results



(a) Varying γ with fixed $\Gamma = 0.42$.



(b) Varying Γ with fixed $\gamma = 1$.

Non-Asymptotic Result

Theorem

For fixed number of nodes n and $4\ell \leq \log_{\mathbb{E}\deg}(n)$, the classifier h_ℓ^* is $o_n(1)$ away from the true optimal in terms misclassification probability.

- h_ℓ^* minimizes probability of misclassification in the local weak limit of the model
- $h_{\ell,n}^*$ minimizes probability of misclassification in the finite n model
- We show that $\text{Error}(h_\ell^*) - \text{Error}(h_{\ell,n}^*) = o_n(1)$
- Proof technique utilizes Stein's method