Graph Convolution for Semi-Supervised Classification: Improved Linear Separability and Out-of-Distribution Generalization (Supplementary Material)

Aseem Baranwal [1] **Kimon Fountoulakis** [1] **Aukosh S. Jagannath** [2]

1. Outline

This article provides the supplementary material to support the content from the main paper. In Section 2, we briefly describe our model and state the main results for the readers. In Section 3, we show a concentration bound for the class sizes and the degree of a node. Our results rely heavily on these concentration bounds. We then prove our first main result about separability thresholds in Section 4. Finally, we prove our second main result about generalization in Section 5.

2. Summary of Model and Results

Let $(\varepsilon_k)_{k\in[n]}$ be i.i.d. $\mathrm{Ber}(\frac{1}{2})$ random variables. Corresponding to these, consider a stochastic block model consisting of two classes $C_0 = \{i \in [n] : \varepsilon_i = 0\}$ and $C_1 = C_0^{\mathsf{c}}$ with inter-class edge probability q and intra-class edge probability p with no self-loops. In particular, conditionally on (ε_k) the adjacency matrix $A = (a_{ij})$ is Bernoulli with $a_{ij} \sim \mathrm{Ber}(p)$ if i, j are in the same class and $a_{ij} \sim \mathrm{Ber}(q)$ if they are in distinct classes. Along with this, consider $X \in \mathbb{R}^{n\times d}$ to be the feature matrix such that each row X_i is an independent d-dimensional Gaussian random vector with $X_i \sim N(\boldsymbol{\mu}, \frac{1}{d}I)$ if $i \in C_0$ and $X_i \sim N(\boldsymbol{\nu}, \frac{1}{d}I)$ if $i \in C_1$. Here $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^d$ are fixed vectors with $\|\boldsymbol{\mu}\|_2, \|\boldsymbol{\nu}\|_2 \leq 1$ and I is the identity matrix. Denote by $\mathrm{CSBM}(n, p, q, \boldsymbol{\mu}, \boldsymbol{\nu})$ the coupling of a stochastic block model with a two component Gaussian mixture model with means $\boldsymbol{\mu}, \boldsymbol{\nu}$ and covariance $\frac{1}{d}I$ as described above and we denote a sample by $(A, X) \sim \mathrm{CSBM}(n, p, q, \boldsymbol{\mu}, \boldsymbol{\nu})$.[1] Observe that the marginal distribution for A is a stochastic block model and that the marginal distribution for X is a two-component Gaussian mixture model. Finally, define $\tilde{A} = (\tilde{a}_{ij}) = A + I$ and D, the diagonal degree matrix for $\tilde{A}$ where $D_{ii} = \sum_{j\in[n]} \tilde{a}_{ij}$ for all $i \in [n]$. Then the graph convolution of some data X is given by $\tilde{X} = D^{-1}\tilde{A}X$.

For parameters $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, the label predictions are given by $\hat{\mathbf{y}} = \sigma(D^{-1}\tilde{A}X\mathbf{w} + b\mathbf{1})$, where $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function applied element-wise in the usual sense. Note that we work in the semi-supervised setting where only a fraction of the labels are available. In particular, we will assume that for some fixed $0 < \beta_0, \beta_1 \leq \frac{1}{2}$, the number of labels available for class C_0 is $\beta_0 n$ and for class C_1 is $\beta_1 n$. Let $S = \{i : y_i \text{ is available}\}$ so that $|S| = (\beta_0 + \beta_1)n$. The loss function we use is the binary cross entropy,

$$L(A, X, \mathbf{w}, b) = -\frac{1}{|S|}\sum_{i\in S} y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i), \tag{1}$$

where y_i is the given label of node i, and $\hat{y}_i$ is the predicted label of node i (also, the i-th component of vector $\hat{\mathbf{y}}$). Observe that the binary cross-entropy loss used in Logistic regression can be written as $L(I, X, \mathbf{w}, b)$.

[1]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada [2]Department of Statistics and Actuarial Science, Department of Applied Mathematics, University of Waterloo, Waterloo, Canada. Correspondence to: Aseem Baranwal <aseem.baranwal@uwaterloo.ca>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]We note here that, we could also have considered $\sigma^2 I$ instead of I/d, in which case all of our results still hold after rescaling the thresholds appropriately. For example, if we took $\sigma^2 = 1$, then the relevant critical thresholds for linear separability become $\|\boldsymbol{\mu} - \boldsymbol{\nu}\| \sim 1$ and $\|\boldsymbol{\mu} - \boldsymbol{\nu}\| \sim 1/\sqrt{D}$ for the mixture model and the CSBM respectively.

The optimization problem we consider is the following.

$$\mathrm{OPT}_d(A, X, R) = \min_{\substack{\|\mathbf{w}\|_2 \le R, \\ b \in \mathbb{R}}} L(A, X, \mathbf{w}, b) \tag{2}$$

Here R is a real number which can vary in d. We usually think of $R \sim d$. We also define the following quantities that are used throughout the paper.

$$\gamma = \frac{\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2}{2}, \quad \Gamma(p, q) = \frac{p - q}{p + q}. \tag{3}$$

We work under two scaling assumptions that are defined below.

Assumption 1. *We say that n satisfies* Assumption 1 *if*

$$\omega(d \log d) \le n \le O(\mathrm{poly}(d)).$$

Assumption 2. *We say that (p, q) satisfies* Assumption 2 *if*

$$p, q = \omega(\log^2(n)/n) \quad \textit{and} \quad \Gamma(p, q) = \Omega(1).$$

The first several sections of this supplement are intended to provided the proofs of our main results, which we briefly recall here for the reader.

Theorem 1. *Suppose that n satisfies Assumption 1 and that (p, q) satisfies Assumption 2. Fix $0 < \beta_0, \beta_1 \le 1/2$ and let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{B}^d$. For any $(A, X) \sim \mathrm{CSBM}(n, p, q, \boldsymbol{\mu}, \boldsymbol{\nu})$, we have the following:*

1. *For any $K \ge 0$ if $\|\boldsymbol{\mu} - \boldsymbol{\nu}\| \le K/\sqrt{d}$, then there are some $C, c > 0$ such that for $d \ge 1$*

   $$\mathbb{P}((X_i)_{i \in S} \textit{ is linearly separable}) \le C \exp(-cd).$$

   Furthermore, for any $t > 0$ there is a $c > 0$ such that for every $R > 0$,

   $$\mathrm{OPT}_d(I, X, R) \ge 2(\beta_0 \wedge \beta_1)\Phi\left(-\frac{K}{2}(1 + t)\right)\log(2)$$

   with probability $1 - \exp(-cd)$.

2. *If $\|\boldsymbol{\mu} - \boldsymbol{\nu}\| = \omega(\frac{\log n}{\sqrt{dn(p+q)/2}})$, then*

   $$\mathbb{P}((\tilde{X}_i)_{i \in S} \textit{ is linearly separable}) = 1 - o_d(1),$$

   where $o_d(1)$ denotes a quantity that converges to 0 as $d \to \infty$. Furthermore, with probability $1 - o_d(1)$, we have for all $R > 0$

   $$\mathrm{OPT}_d(A, X, R) \le \exp\left(-\frac{R}{2}\Gamma(p, q)\|\boldsymbol{\mu} - \boldsymbol{\nu}\|(1 - o_d(1))\right).$$

3. *For any $K \ge 0$ if $\|\boldsymbol{\mu} - \boldsymbol{\nu}\| \le K/\sqrt{dn(p+q)/2}$, then*

   $$\mathbb{P}((\tilde{X}_i)_{i \in S} \textit{ is linearly separable}) = o_d(1).$$

   Furthermore, for any $t > 0$ with probability $1 - o_d(1)$, for all $R > 0$

   $$\mathrm{OPT}_d(A, X, R) \ge 2(\beta_0 \wedge \beta_1)\Phi\Big(-\frac{K}{2}(1 + t)\Big)\log(2).$$

Note that Assumption 1 states $\omega(d \log d) = n = \mathcal{O}(\mathrm{poly}(d))$. Thus, one could obtain the same results by interchanging the dependence of n and d. We do not really require $d \to \infty$. Theorem 1 part 1 also holds for fixed $d \ge 1$ and large n by taking $L = \sqrt{n}$ in the proof of Lemma 1. In this case, the probability of separability is $\exp(-\Omega(n))$.

In Theorem 2, we state the generalization result and show that the graph convolution obtains diminishing loss on out-of-distribution test data as well.

Theorem 2. *Suppose that n and n' satisfy Assumption 1. Suppose furthermore that the pairs (p,q) and (p',q') satisfy Assumption 2. Fix $0 < \beta_1, \beta_2 \leq 1/2$ and $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{B}^d$. Let $(A,X) \sim \mathrm{CSBM}(n,p,q,\boldsymbol{\mu},\boldsymbol{\nu})$. Let $(\mathbf{w}^*(R), b^*(R))$ be the optimizer of (2). Then for any sample $(A',X') \sim \mathrm{CSBM}(n',p',q',\boldsymbol{\mu},\boldsymbol{\nu})$ independent of (A,X), there is a $C > 0$ such that with probability $1 - o_d(1)$ we have that for all $R > 0$*

$$L(A',X',\mathbf{w}^*(R),b^*(R)) \leq C\exp\Big(-\frac{R}{2}\|\boldsymbol{\mu}-\boldsymbol{\nu}\|\Gamma(p',q')(1-o(1))\Big)$$

where the loss (1) is with respect to the full test set $S = [n']$.

3. Degree Concentration

We note here the following elementary concentration results for the class size and degrees, which are all straightforward consequences of the Chernoff bound for sums of independent Bernoulli random variables, see, e.g., (Vershynin, 2018).

Since $(\varepsilon_i)_{i\in[n]} \sim \mathrm{Ber}(\frac{1}{2})$, by the Chernoff bound we have for any $\delta > 0$ that

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i - \frac{1}{2}\Big| \geq \delta/2\Big) \leq 2\exp(-n\delta^2/6).$$

In particular, we have that for any $\delta > 0$ the number of nodes in each class satisfies

$$\mathbb{P}\left(\frac{|C_0|}{n}, \frac{|C_1|}{n} \in \left[\frac{1}{2}-\delta, \frac{1}{2}+\delta\right]\right) \geq 1 - C\exp(-cn\delta^2), \tag{4}$$

for some $C, c > 0$.

The degrees are sums of Bernoulli random variables. Hence, by the Chernoff bound, for each i, we have for $\delta \in (0,1)$ that

$$\mathbb{P}\left(\Big|D_{ii} - \mathbb{E}[D_{ii}]\Big| \geq \delta\mathbb{E}[D_{ii}]\right) \leq 2\exp(-\mathbb{E}[D_{ii}]\delta^2/3),$$

where for any i,

$$\mathbb{E}[D_{ii}] = \frac{1}{2}(\mathbb{E}[D_{ii} \mid \varepsilon_i = 0] + \mathbb{E}[D_{ii} \mid \varepsilon_i = 1]) = 1 + \frac{n-1}{2}(p+q).$$

In particular, it follows that for any $\delta \in (0,1)$ we have

$$\mathbb{P}\left(\frac{D_{ii}}{n} \in \left[\frac{1}{2}(p+q)(1-\delta), \frac{1}{2}(p+q)(1+\delta)\right]^{\mathsf{c}}\right) \leq C\exp(-cn(p+q)\delta^2), \tag{5}$$

for some $C, c > 0$. As we will frequently work on the event that the degrees and class sizes concentrate, for fixed $\delta, \delta' > 0$ we define the event

$$B(\delta,\delta') = \left\{\frac{n}{2}(1-\delta) \leq |C_0|, |C_1| \leq \frac{n}{2}(1+\delta)\right\} \bigcap_{i\in[n]} \left\{\frac{n}{2}(p+q)(1-\delta') \leq D_{ii} \leq \frac{n}{2}(p+q)(1+\delta')\right\}. \tag{6}$$

Since $p, q = \omega(\frac{\log^2 n}{n})$, by the union bound, if we choose $\delta = n^{-1/2+\epsilon}$ and $\delta' = (\log n)^{-1/2+\epsilon}$ then for $\epsilon > 0$ small enough, for any $c > 1$ there is $C > 0$ such that

$$\mathbb{P}(B(n^{-1/2+\epsilon}, (\log n)^{-1/2+\epsilon})) \geq 1 - \frac{C}{n^c}. \tag{7}$$

Let N_i denote those vertices connected to i (including i), and let

$$\tilde{B}(\delta,\delta') = B(\delta,\delta') \bigcap_{i\in[n]} \left\{\frac{(1-\varepsilon_i)p + \varepsilon_i q}{p+q}(1-\delta') \leq \frac{|C_0 \cap N_i|}{D_{ii}} \leq \frac{(1-\varepsilon_i)p + \varepsilon_i q}{p+q}(1+\delta')\right\}$$
$$\bigcap_{i\in[n]} \left\{\frac{\varepsilon_i p + (1-\varepsilon_i)q}{p+q}(1-\delta') \leq \frac{|C_1 \cap N_i|}{D_{ii}} \leq \frac{\varepsilon_i p + (1-\varepsilon_i)q}{p+q}(1+\delta')\right\}$$

by similar reasoning, a Chernoff bound and union bound yields, for $\epsilon > 0$ small enough, we have that for any $c > 0$, and some $C > 0$,

$$\mathbb{P}\left(\tilde{B}(n^{-1/2+\epsilon}, (\log n)^{-1/2+\epsilon})\right) \geq 1 - \frac{C}{n^c}. \tag{8}$$

4. Separability Thresholds

In this section, we prove Theorem 1. We begin by first proving a bound on a certain Gaussian process. We then develop concentration bounds for the convolved data. We end the section by proving the three parts of the theorem in turn.

4.1. Bounds for the isonormal process

Consider the Gaussian process, $g(\mathbf{v}) = \langle Z, \mathbf{v}\rangle$ for all $\mathbf{v} \in \mathbb{R}^d$, for some standard Gaussian vector $Z \sim N(\mathbf{0}, I)$. The process g is sometimes called the *isonormal process* or *canonical Gaussian process*. Controlling its behaviour will be an essential step in showing that the mixture model or CSBM data is not linearly separable below a certain threshold. Let $g_i(\mathbf{v})$ denote i.i.d. copies of this process and Define the events

$$A_{k,\gamma,n}(\mathbf{v}) = \{\exists J \subseteq [n] : g_i(\mathbf{v}) > \gamma \text{ for } i \in J, \quad |J| = k\},$$
$$\tilde{A}_{k,\gamma,n}(\mathbf{v}) = \{\exists J \subseteq [n] : g_i(\mathbf{v}) > \gamma \text{ for } i \in J, g_i(\mathbf{v}) < \gamma \text{ for } i \in J^c \quad |J| = k\}$$

Observe that for each k, $A_{k,\gamma,n}(\mathbf{v})$ is the event that at $\mathbf{v}$ the k-th largest of the $g_i(\mathbf{v})$ is large and the tilded version is the event that this occurs and the remaining $g_i(\mathbf{v})$ are all small. Let

$$H(x) = -x\log x - (1-x)\log(1-x).$$

Finally for $\epsilon > 0$, let $\Sigma_{\epsilon,d}$ denote an ϵ-net of the unit sphere, $\mathbb{S}^{d-1}$. We begin by showing the following result about the isonormal process.

Lemma 1. *Suppose that n satisfies Assumption 1. Then for any $\gamma > 0$ and $0 < t < 1 - \Phi(\gamma)$, there is a $C > 0$ such that for $d \geq 1$,*

$$\frac{1}{d}\log \mathbb{P}\Big(\bigcup_{\mathbf{v}\in\mathbb{S}^{d-1}} A_{\lfloor tn\rfloor,\gamma,n}(\mathbf{v})^c\Big) \leq -C, \tag{9}$$

and for any $0 < t \leq 1$, there is a $C > 0$ such that for $d \geq 1$

$$\frac{1}{d}\log \mathbb{P}\Big(\bigcup_{\mathbf{v}\in\mathbb{S}^{d-1}} A_{1,\gamma,\lfloor tn\rfloor}(\mathbf{v})^c\Big) \leq -C. \tag{10}$$

Proof. It will suffice to consider only the first case as the second case clearly follows by the same argument. For L sufficiently large and fixed $\epsilon < \gamma$, let $\epsilon' = \frac{\epsilon}{L\sqrt{d}}$. Then we have that

$$\begin{aligned}
\mathbb{P}\Big(\bigcup_{\mathbf{v}\in\mathbb{S}^{d-1}} A_{\lfloor tn\rfloor,\gamma,n}(\mathbf{v})^c\Big) &\leq n\mathbb{P}(\|Z_1\| > L\sqrt{d}) + \mathbb{P}\Big(\bigcup_{\mathbf{v}\in\mathbb{S}^{d-1}} A_{\lfloor tn\rfloor,\gamma,n}(\mathbf{v})^c \cap \{\|Z_i\| \leq L\sqrt{d} \ \ \forall i \in [n]\}\Big) \\
&\leq n\mathbb{P}(\|Z_1\| > L\sqrt{d}) + \mathbb{P}\Big(\bigcup_{\mathbf{v}\in\Sigma_{\epsilon',d}} A_{\lfloor tn\rfloor,\gamma+\epsilon,n}(\mathbf{v})^c \cap \{\|Z_i\| \leq L\sqrt{d} \ \ \forall i \in [n]\}\Big) \\
&\leq n\mathbb{P}(\|Z_1\| > L\sqrt{d}) + \mathbb{P}\Big(\bigcup_{\mathbf{v}\in\Sigma_{\epsilon',d}} A_{\lfloor tn\rfloor,\gamma+\epsilon,n}(\mathbf{v})^c\Big) = A + B,
\end{aligned}$$

The first inequality above follows from the law of total probability and then a union bound over all $i \in [n]$. For the second inequality, observe that since $\Sigma_{\epsilon',d}$ is an ϵ'-net, we have that for a fixed $\mathbf{v} \in \mathbb{S}^{d-1}$ if $\mathbf{u} \in \Sigma_{\epsilon',d}$ is the vector in the ϵ'-net nearest to $\mathbf{v}$ then if we let $E = \{\|Z_i\| \leq L\sqrt{d}\, i \in [n]\}$ then $A_{\lfloor tn\rfloor,\gamma,n}(\mathbf{v})^c \cap E \subseteq A_{\lfloor tn\rfloor,\gamma+\epsilon,n}(\mathbf{u})^c \cap E$.

We bound these terms in turn. Let us begin with A. Recall that by the norm concentration of a standard Gaussian vector (Vershynin, 2018), there exist $C, c > 0$ such that for any $L > 1$ and $d \geq 1$,

$$\mathbb{P}(\|Z_1\| > L\sqrt{d}) \leq C\exp(-cdL^2).$$

On the other hand, for B, we have that for some $C' > 0$ and any fixed $\mathbf{v} \in \mathbb{S}^{d-1}$

$$\begin{aligned}
B &\le |\Sigma_{\epsilon',d}|\mathbb{P}(A_{\lfloor nt\rfloor,\gamma+\epsilon,d,n}(\mathbf{v})^c) \le \exp(C'd\log(d/\epsilon))\mathbb{P}(A_{\lfloor nt\rfloor,\gamma+\epsilon,n}(\mathbf{v})^c) \\
&\le \exp(C'd\log(d/\epsilon)) \sum_{s<nt} \mathbb{P}(\tilde{A}_{s,\gamma+\epsilon,n}(\mathbf{v})) \\
&\le \exp(C'd\log(d/\epsilon)) \sum_{s<nt} \binom{n}{s} \mathbb{P}(g_1(\mathbf{v}) > \gamma+\epsilon)^s \mathbb{P}(g_1(\mathbf{v}) < \gamma+\epsilon)^{n-s} \\
&\le \exp(C'd\log(d/\epsilon)) \sum_{s<nt} \exp\left(nH\left(\frac{s}{n}\right)\right)(1-\Phi(\gamma+\epsilon))^s \Phi(\gamma+\epsilon)^{n-s} \\
&= \exp(C'd\log(d/\epsilon)) \sum_{s<nt} \exp\left(n\left[H\left(\frac{s}{n}\right) + \frac{s}{n}\log(1-\Phi(\gamma+\epsilon)) + \left(1-\frac{s}{n}\right)\log\Phi(\gamma+\epsilon)\right]\right) \\
&\le nt\exp\{n[H(t) + t\log(1-\Phi(\gamma+\epsilon)) + (1-t)\log\Phi(\gamma+\epsilon)] + O(d\log(d/\epsilon))\}.
\end{aligned}$$

The first inequality follows by a union bound. The second inequality follows from $|\Sigma_{\epsilon,d}| \le (2/\epsilon+1)^d$ for any $\epsilon \in (0, \frac{1}{2})$ (Vershynin, 2018), the third follows by union bound since $A^c_{\lfloor nt\rfloor,\gamma+\epsilon,n} \subseteq \cup_{s<nt}\tilde{A}_{s,\gamma+\epsilon,n}$, the fourth follows since $g_i(\mathbf{v})$ are i.i.d., and the fifth by the Stirling bound $\binom{n}{nt} \le \exp(nH(t))$. For the final inequality, note that since $\gamma+\epsilon > 0$, the function

$$f(x) = H(x) + x\log(1-\Phi(\gamma+\epsilon)) + (1-x)\log\Phi(\gamma+\epsilon)$$

is negative and increasing for $0 < x < 1-\Phi(\gamma+\epsilon)$ so that each summand is bounded above by the value at $s = nt$ since $t < 1-\Phi(\gamma+\epsilon)$. Since by Assumption 1, $n = \omega(d\log d)$, we have that there is some $C > 0$ such that

$$B \le C\exp(-cd\log d)$$

Consequently, $0 \le B/A \le C$ for all d for some $C > 0$ Combining the bounds on A and B we obtain

$$\frac{1}{d}\log(A+B) \le \frac{1}{d}\log A + \frac{1}{d}\log(1+C) = -cL^2 + O(\frac{1}{d}),$$

from which the result follows. □

4.2. Proof of part 1 of Theorem 1

We are now ready to prove part 1 of Theorem 1, which shows the threshold for data to be linearly separable, along with a corresponding lower bound for the loss.

Proof of Theorem 1 part 1. Observe that X_i can be written as

$$X_i = (1-\varepsilon_i)\boldsymbol{\mu} + \varepsilon_i\boldsymbol{\nu} + \frac{Z_i}{\sqrt{d}}$$

where Z_i are i.i.d. standard Gaussian vectors.

By (4), it suffices to bound these terms on the event from (4). If (X_i) are linearly separable, then there is a unit vector $\mathbf{v}$ and $b \in \mathbb{R}$ such that

$$\langle\boldsymbol{\mu},\mathbf{v}\rangle + \frac{\langle Z_i,\mathbf{v}\rangle}{\sqrt{d}} + b < 0, \quad i \in S_0 \quad \text{and} \quad \langle\boldsymbol{\nu},\mathbf{v}\rangle + \frac{\langle Z_i,\mathbf{v}\rangle}{\sqrt{d}} + b > 0, \quad i \in S_1. \tag{11}$$

Recall that $\|\boldsymbol{\mu}-\boldsymbol{\nu}\| \le K/\sqrt{d}$. Hence, writing $b = b' - \frac{\langle\boldsymbol{\mu}+\boldsymbol{\nu},\mathbf{v}\rangle}{2}$, we see that if the above holds then there is a pair $(\mathbf{v}, b')$ such that

$$\max_{i\in S_0}\frac{\langle Z_i,\mathbf{v}\rangle}{\sqrt{d}} + b' < \frac{K}{2\sqrt{d}} \quad \text{and} \quad \min_{i\in S_1}\frac{\langle Z_i,\mathbf{v}\rangle}{\sqrt{d}} + b' > -\frac{K}{2\sqrt{d}}. \tag{12}$$

Such a pair $(\mathbf{v}, b')$ exists only if at least one of the above two holds with $b' = 0$. Conditionally on the event $|S_0| = k$, the probability of this occurring is at most sum of the probability of these two events:

$$\begin{aligned}
&\mathbb{P}\Big(\exists\mathbf{v}\in\mathbb{S}^{d-1} : \max_{i\le k} g_i(\mathbf{v}) < K/2\Big) + \mathbb{P}\Big(\exists\mathbf{v}\in\mathbb{S}^{d-1} : \min_{i\in[|S|-k,|S|]} g_i(\mathbf{v}) > -K/2\Big) \\
&\le 2\mathbb{P}\Big(\exists\mathbf{v}\in\mathbb{S}^{d-1} : \max_{i\le k\wedge|S|-k} g_i(\mathbf{v}) < K/2\Big) = I_k
\end{aligned}$$

As this function is decreasing in k, it suffices to bound it in the case that $k = (\frac{1}{2} - \delta)\beta_0 n$, by (4). Note that

$$I_{tn} = 2\mathbb{P}\Big(\bigcup_{\mathbf{v}\in\mathbb{S}^{d-1}} A_{1,\gamma,\lfloor tn \rfloor}(\mathbf{v})^c \Big)$$

with $\gamma = K/2$ and $t = (\frac{1}{2} - \delta)\beta_0$. Thus, using (10) we have that

$$\frac{1}{d}\log(I_{tn}) \leq -C.$$

The first result then follows by combining this with (4).

Let us now turn to the lower bound on the loss. Take $t < 1 - \Phi((1+\epsilon)K/2)$ for some $\epsilon > 0$. Since $K > 0$ and $\beta_0, \beta_1 \leq 1/2$, using (9) with $\gamma = \frac{K}{2}(1+\epsilon)$ we have that with probability at least $1 - C\exp(-cd)$, for all $\mathbf{v}$ with $\|\mathbf{v}\| = 1$ there are $t\beta_0 n$ choices of $i \in S_0$ and $t\beta_1 n$ choices of $i \in S_1$ with

$$\langle Z_i, \mathbf{v}\rangle > (1+\epsilon)\frac{K}{2} \qquad \text{and} \qquad \langle Z_i, \mathbf{v}\rangle < -(1+\epsilon)\frac{K}{2} \tag{13}$$

respectively. Let these sets of indices be denoted by $J(\mathbf{v}), J'(\mathbf{v})$, and let $l(X_i, \varepsilon_i, \mathbf{v}, b)$ denote the loss, given by

$$l(X_i, \varepsilon_i, \mathbf{v}, b) = -\varepsilon_i \log(\sigma(\langle X_i, \mathbf{v}\rangle + b)) - (1-\varepsilon_i)\log(1 - \sigma(\langle X_i, \mathbf{v}\rangle + b)).$$

Then using (13) we have that for each $\mathbf{v} \in \mathbb{R}^d$

$$\begin{aligned}
\min_{i\in J(\mathbf{v}/\|\mathbf{v}\|)} l(X_i, \varepsilon_i, \mathbf{v}, b) &= -\log(1 - \sigma(\langle X_i, \mathbf{v}\rangle + b)) \\
&= -\log\left(1 - \sigma\left(\langle \boldsymbol{\mu}, \mathbf{v}\rangle + b + \frac{\langle Z_i, \mathbf{v}\rangle}{\sqrt{d}}\right)\right) \\
&\geq -\log\left(1 - \sigma\left(\epsilon\frac{K\|\mathbf{v}\|}{2\sqrt{d}} + b'\right)\right).
\end{aligned}$$

Similarly,

$$\min_{i\in J'(\mathbf{v}/\|\mathbf{v}\|)} l(X_i, \varepsilon_i, \mathbf{v}, b) \geq -\log\sigma\left(-\epsilon\frac{K\|\mathbf{v}\|}{2\sqrt{d}} + b'\right).$$

Thus, using (1) we have that

$$L(I, X, \mathbf{v}, b) \geq t f\left(\epsilon\frac{K\|\mathbf{v}\|}{2\sqrt{d}}, b'\right),$$

where

$$f(x,y) = -\beta_0\log(1 - \sigma(x+y)) - \beta_1\log\sigma(-x+y) = \beta_0\log(1+e^{x+y}) + \beta_1\log(1+e^{x-y}).$$

Note that by optimizing in x, y, we see that for $\beta_0 = \beta_1$ and $x \geq 0$, we have

$$f(x,y) \geq f(0,0) = 2\beta_0\log(2),$$

so that for any $0 < \beta_0, \beta_1 \leq \frac{1}{2}$ and $x \geq 0$ we have $f(x,y) \geq (\beta_0 \wedge \beta_1)2\log 2$, so that

$$L(I, X, \mathbf{v}, b) \geq 2t \cdot \beta_0 \wedge \beta_1 \cdot \log 2.$$

Combining the above and minimizing in $\mathbf{v}, b$, we see that for every $0 < t < 1 - \Phi((1+\epsilon)K/2)$, there is some $c > 0$ such that

$$\min_{\mathbf{v}\in\mathbb{R}^d, b\in\mathbb{R}} L(I, X, \mathbf{v}, b) \geq 2t\beta_0 \wedge \beta_1 \cdot \log 2$$

with probability $1 - \exp(-cd)$ as desired. □

4.3. Decomposition of the convolved data

In this subsection we provide a decomposition of $\tilde{X}$ which we will use frequently throughout the rest of this paper. Note that conditionally on (ε_i), we have that $X_j \sim \mathcal{N}(\boldsymbol{\mu}_j, \frac{1}{d}I)$ where $\boldsymbol{\mu}_j = \boldsymbol{\mu}$ if $j \in C_0$ and $\boldsymbol{\mu}_j = \boldsymbol{\nu}$ if $j \in C_1$. Thus, we can write

$$X_j = (1-\varepsilon_j)\boldsymbol{\mu} + \varepsilon_j \boldsymbol{\nu} + \frac{g_j}{\sqrt{d}}, \tag{14}$$

where $g_j \sim \mathcal{N}(\mathbf{0}, I)$ are i.i.d. copies of a standard normal vector.

Lemma 2. *Conditionally on A and (ε_k), for any $\eta > 0$ small enough and $c > 0$ large enough we have that with probability at least $1 - 1/n^c$, for every $i \in [n]$,*

$$\left\| \tilde{X}_i - \frac{p\boldsymbol{\mu} + q\boldsymbol{\nu}}{p+q}(1+o(1)) \right\|_2 = O\left(\frac{d^\eta}{\sqrt{dn(p+q)}}\right) \text{ if } \varepsilon_i = 0,$$
$$\left\| \tilde{X}_i - \frac{q\boldsymbol{\mu} + p\boldsymbol{\nu}}{p+q}(1+o(1)) \right\|_2 = O\left(\frac{d^\eta}{\sqrt{dn(p+q)}}\right) \text{ if } \varepsilon_i = 1.$$

Proof. Consider the random variables $\tilde{X}_i = [D^{-1}\tilde{A}X]_i$. For any fixed i, we define $m(i)$ to be conditional mean of $\tilde{X}_i$ on the adjacency matrix A and class memberships (ε_j),

$$m(i) = \mathbb{E}[\tilde{X}_i \mid A, \varepsilon] = \frac{1}{D_{ii}} \sum_{j\in[n]} \tilde{a}_{ij}\boldsymbol{\mu}_j.$$

From (14) we can rewrite

$$\tilde{X}_i = \frac{1}{D_{ii}} \sum_{j\in[n]} \tilde{a}_{ij} X_j = m(i) + \frac{1}{D_{ii}\sqrt{d}} \sum_{j\in[n]} \tilde{a}_{ij} g_j, \tag{15}$$

When $\varepsilon_i = 0$, we have that

$$m(i) = \frac{1}{D_{ii}} \left(\sum_{j\in C_0} \tilde{a}_{ij}\boldsymbol{\mu} + \sum_{j\in C_1} \tilde{a}_{ij}\boldsymbol{\nu} \right) = \frac{1}{D_{ii}}(|C_0 \cap N_i|\boldsymbol{\mu} + |C_1 \cap N_i|\boldsymbol{\nu}), \tag{16}$$

and similarly when $\varepsilon_i = 1$.

Note that by (8), we have that with probability $1 - 1/n^c$ for $c > 0$ large enough,

$$\frac{|C_0 \cap N_i|}{D_{ii}} = \left[(1-\varepsilon_i)\frac{p}{p+q} + \varepsilon_i \frac{q}{p+q}\right](1+o(1)),$$
$$\frac{|C_1 \cap N_i|}{D_{ii}} = \left[\varepsilon_i \frac{p}{p+q} + (1-\varepsilon_i)\frac{q}{p+q}\right](1+o(1))$$
$$\frac{1}{D_{ii}} = \frac{2}{n(p+q)}(1+o(1))$$

for all $i \in [n]$, and so we have that

$$m(i) = \frac{p\boldsymbol{\mu} + q\boldsymbol{\nu}}{p+q}(1+o(1)) \qquad \text{for } \varepsilon_i = 0, \tag{17}$$
$$m(i) = \frac{q\boldsymbol{\mu} + p\boldsymbol{\nu}}{p+q}(1+o(1)) \qquad \text{for } \varepsilon_i = 1. \tag{18}$$

Next, we consider $F = \frac{1}{D_{ii}\sqrt{d}} \sum_{j\in[n]} \tilde{a}_{ij} g_j$. Note that for a given adjacency matrix A, we have that $F \sim \mathcal{N}(\mathbf{0}, \frac{1}{dD_{ii}}I)$. As above, by concentration of the norm of Gaussian vectors, we have

$$\mathbb{P}(\|F\|_2 > t \mid A) \le 2\exp(-t^2 d D_{ii}/2). \tag{19}$$

Define the event $Q = Q(K) = \{\|F\|_2 \leq K\}$ and note that if we let $\tilde{B} = \tilde{B}(n^{-1/2+\epsilon}, (\log n)^{\epsilon-1/2})$,

$$\mathbb{P}(Q^{\mathsf{c}}) \leq \mathbb{P}(\tilde{B} \cap Q^{\mathsf{c}}) + \mathbb{P}(\tilde{B}^{\mathsf{c}}) \leq 2\exp(-c'K^2 dn(p+q)) + \frac{1}{n^c},$$

for some $c' > 0$ and any $c > 0$ large enough. Subsequently, we have

$$\mathbb{P}(\tilde{B} \cap Q) \geq 1 - \mathbb{P}(\tilde{B}^{\mathsf{c}}) - \mathbb{P}(Q^{\mathsf{c}}) \geq 1 - \frac{2}{n^c} - 2\exp(-c'K^2 dn(p+q)).$$

We now choose $K = \frac{d^\eta}{\sqrt{dn(p+q)}}$ for some $\eta > 0$. Then using (8) and (19) with a union bound, it follows that

$$\mathbb{P}(\tilde{B} \cap Q) \geq 1 - \frac{2}{n^c} - 2\exp(-c'd^{2\eta}).$$

Note that $\|\boldsymbol{\mu}\|_2, \|\boldsymbol{\nu}\|_2 \leq 1$. Thus, on the event $\tilde{B} \cap Q$, we have that

$$\|\tilde{X}_i - m(i)\|_2 = O\left(\frac{d^\eta}{\sqrt{dn(p+q)}}\right),$$

from which the result follows upon recalling (17) and (18). □

4.4. Rate of decay of the loss for chosen parameters

Here we show that there exists a choice of parameters $(\tilde{\mathbf{w}}, \tilde{b})$ such that the loss incurred on any sample $(A, X) \sim \mathrm{CSBM}(n, p, q, \boldsymbol{\mu}, \boldsymbol{\nu})$ is exponentially small with a high probability.

Lemma 3. *Consider the following parameters that satisfy the constraints of the problem in* (2)*.*

$$\tilde{\mathbf{w}}(R) = \frac{R}{2\gamma}(\boldsymbol{\nu} - \boldsymbol{\mu}), \qquad \tilde{b}(R) = -\frac{\langle \boldsymbol{\mu} + \boldsymbol{\nu}, \tilde{\mathbf{w}}(R)\rangle}{2},$$

where $\gamma = \frac{1}{2}\|\boldsymbol{\mu} - \boldsymbol{\nu}\|$*. Consider a sample* $(A, X) \sim \mathrm{CSBM}(n, p, q, \boldsymbol{\mu}, \boldsymbol{\nu})$ *such that* $p > q$*. Then for any* $0 < \beta_1, \beta_2 \leq 1/2$ *and any* $c > 0$ *large enough, with probability at least* $1 - n^{-c}$ *, we have that for all* $R > 0$*,*

$$L(A, X, \tilde{\mathbf{w}}, \tilde{b}) = \exp\left(-R\gamma\Gamma(p,q)(1 + o(1))\right).$$

Proof. For readability, we suppress the dependence of R in $(\tilde{\mathbf{w}}, \tilde{b})$ when it is clear from context. Consider the loss for a single node i for which we know the label y_i,

$$\begin{aligned} L_i(A, X, \tilde{\mathbf{w}}, \tilde{b}) &= -y_i \log(\sigma(\langle \tilde{X}_i, \tilde{\mathbf{w}}\rangle + \tilde{b})) - (1 - y_i)\log(1 - \sigma(\langle \tilde{X}_i, \tilde{\mathbf{w}}\rangle + \tilde{b})) \\ &= \log\left(1 + \exp\left((1 - 2\varepsilon_i)(\langle \tilde{X}_i, \tilde{\mathbf{w}}\rangle + \tilde{b})\right)\right). \end{aligned}$$

We will work on the case where $\varepsilon_i = 0$ as the analysis for $\varepsilon_i = 1$ is symmetric. Using Lemma 2 and the fact that $\|\boldsymbol{\mu}\|, \|\boldsymbol{\nu}\| \leq 1$, it follows that with probability at least $1 - n^{-c}$ we have that for all $i \in [n]$,

$$\begin{aligned} \langle \tilde{X}_i, \tilde{\mathbf{w}}\rangle + \tilde{b} &= \frac{\langle p\boldsymbol{\mu} + q\boldsymbol{\nu}, \tilde{\mathbf{w}}\rangle}{p+q}(1 + o(1)) + O\left(\frac{\|\tilde{\mathbf{w}}\| d^\eta}{\sqrt{dn(p+q)}}\right) + \tilde{b} \\ &= \frac{p-q}{2(p+q)}\langle \boldsymbol{\mu} - \boldsymbol{\nu}, \tilde{\mathbf{w}}\rangle(1 + o(1)) + o(\|\tilde{\mathbf{w}}\|) \\ &= -\|\tilde{\mathbf{w}}\|\gamma\Gamma(p,q)(1 + o(1)), \end{aligned}$$

where the error terms here are uniform in i. In the second equation, we have used the definition of $\tilde{b}$ and the fact that the error term in the first equation is $o(\|\tilde{\mathbf{w}}\|)$ by combining Assumption 1 and Assumption 2. The third equality follows from the definitions of γ and $\Gamma(p, q)$.

The expression is symmetric for $\varepsilon_i = 1$. Hence, with probability at least $1 - n^{-c}$, we have that for all i, and all $R > 0$,

$$\langle \tilde{X}_i, \tilde{\mathbf{w}}(R) \rangle + \tilde{b}(R) = (2\varepsilon_i - 1) R\gamma\Gamma(p,q)(1 + o_d(1)), \tag{20}$$

where the error term is uniform in i. On this event, we have for each i that

$$L_i(A, X, \tilde{\mathbf{w}}(R), \tilde{b}(R)) = \log\Big(1 + \exp\Big(- R\gamma\Gamma(p,q)(1 + o(1))\Big)\Big).$$

Thus the total loss is given by

$$L(A, X, \tilde{\mathbf{w}}, \tilde{b}) = \frac{1}{|S|}\sum_{i \in S} L_i(A, X, \tilde{\mathbf{w}}, \tilde{b}) = \log\Big(1 + \exp\Big(- R\gamma\Gamma(p,q)(1 + o(1))\Big)\Big).$$

Observe that for $x < 0$, we have that

$$e^{x-1} \le \log(1 + e^x) \le e^x. \tag{21}$$

Hence, we conclude that

$$L(A, X, \tilde{\mathbf{w}}(R), \tilde{b}(R)) = \exp\left(-R\gamma\Gamma(p,q)(1 + o(1))\right)$$

as desired. □

4.5. Proof of part 2 of Theorem 1

We now turn to show the improvement achieved through the graph convolution.

Proof of Theorem 1 part 2. We begin by observing that conditionally on A and $(\varepsilon_k)_{k\in[n]}$, $\tilde{X}_i$ are Gaussian vectors with independent entries and have mean and covariance

$$\mathbb{E}\Big(\tilde{X}_i \mid A, \epsilon\Big) = m(i) = \frac{1}{D_{ii}}\Big(\sum_{j\in[n]} \tilde{a}_{ij}\mathbb{E}[X_j \mid \varepsilon]\Big),$$
$$\mathrm{Cov}(\tilde{X}_i \mid A, \varepsilon) = \frac{1}{dD_{ii}} I.$$

To have linear separability, we need that there is some unit vector $\mathbf{v} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that

$$\begin{cases} \langle m(i), \mathbf{v}\rangle + \frac{1}{\sqrt{dD_{ii}}} g_i(\mathbf{v}) + b < 0 & i \in S_0, \\ \langle m(i), \mathbf{v}\rangle + \frac{1}{\sqrt{dD_{ii}}} g_i(\mathbf{v}) + b > 0 & i \in S_1. \end{cases} \tag{22}$$

We now turn to the event $\tilde{B}(\delta, \delta)$ from (8). On this event we have that $m(i)$ is given by (17) and (18). Recall that for fixed $\mathbf{v} \in \mathbb{S}^{n-1}$, $g_i(\mathbf{v})$ are n i.i.d. standard Gaussians. Therefore for some $C, c > 0$ we have

$$\mathbb{P}(\max |g_i(\mathbf{v})| > K\sqrt{\log n}) \le C\exp(-cK^2),$$

where we have used Borell's inequality (Vershynin, 2018) and the elementary fact that for n i.i.d. standard Gaussians, (z_i),

$$\mathbb{E}[\max_{i\in[n]} |g_i(\mathbf{v})|] = \mathbb{E}\max_i |z_i| \le c\sqrt{\log n}$$

for some $c > 0$. We take $K = C'\sqrt{\log n}$ for some large constant $C' > 0$ so that this probability is $O(1/n^\alpha)$ for some $\alpha > 0$.

Fix $\mathbf{v} = \frac{\boldsymbol{\nu} - \boldsymbol{\mu}}{\|\boldsymbol{\mu} - \boldsymbol{\nu}\|}$ and $b = -\frac{1}{2}\langle \boldsymbol{\mu} + \boldsymbol{\nu}, \mathbf{v}\rangle$. Then using the degree concentration from (8), we have

$$\begin{cases} -\gamma\Gamma(p,q)(1 + o(1)) + O(\frac{\log n}{\sqrt{dn(p+q)}}) < 0 & i \in S_0, \\ \gamma\Gamma(p,q)(1 + o(1)) + O(\frac{\log n}{\sqrt{dn(p+q)}}) > 0 & i \in S_1. \end{cases}$$

The above holds since we have $\gamma = \omega(\frac{\log n}{\sqrt{dn(p+q)}})$.

Now to bound the loss, we take a multiple of the unit vector above, $\mathbf{v} = \frac{R}{2\gamma}(\boldsymbol{\nu} - \boldsymbol{\mu})$, where R is the norm constraint in (2). Then the bound on the loss follows directly from Lemma 3. □

4.6. Proof of part 3 of Theorem 1

We now provide the non-separability threshold for the convolved data, $\tilde{X}$.

Proof of Theorem 1 part 3. The proof of this result is identical to that of Theorem 1 part 1, so we only explain here what changes.

As before it suffices to find $(\mathbf{v}, b)$ with $\mathbf{v}$ a unit vector and $b \in \mathbb{R}$ such that (22) holds. If there is such a $(\mathbf{v}, b)$ then if we write $b = b' - \langle \boldsymbol{\mu} + \boldsymbol{\nu}, \mathbf{v}\rangle/2$, we see that there is some $(\mathbf{v}, b')$ such that

$$\begin{cases} \langle m(i) - \frac{1}{2}(\boldsymbol{\mu}+\boldsymbol{\nu}), \mathbf{v}\rangle + \frac{1}{\sqrt{dD_{ii}}} g_i(\mathbf{v}) + b' < 0 & i \in S_0, \\ \langle m(i) - \frac{1}{2}(\boldsymbol{\mu}+\boldsymbol{\nu}), \mathbf{v}\rangle + \frac{1}{\sqrt{dD_{ii}}} g_i(\mathbf{v}) + b' > 0 & i \in S_1. \end{cases} \tag{23}$$

By the definition of $m(i)$, we see as in (16) that for $i \in S_0$,

$$\begin{aligned} m(i) - \frac{\boldsymbol{\mu}+\boldsymbol{\nu}}{2} &= \frac{2|C_0 \cap N_i|\boldsymbol{\mu} + 2|C_1 \cap N_i|\boldsymbol{\nu} - D_{ii}(\boldsymbol{\mu}+\boldsymbol{\nu})}{2D_{ii}}, \\ &= \frac{|C_0 \cap N_i| - |C_1 \cap N_i|}{2D_{ii}}(\boldsymbol{\mu} - \boldsymbol{\nu}). \end{aligned}$$

and similarly for $i \in S_1$, so that

$$|\langle m(i) - \frac{\boldsymbol{\mu}+\boldsymbol{\nu}}{2}, \mathbf{v}\rangle| \le \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2.$$

Furthermore, since $\|\boldsymbol{\mu} - \boldsymbol{\nu}\| \le K/\sqrt{dn(p+q)/2}$, we see that such a pair $(\mathbf{v}, b')$ exists only if

$$\begin{aligned} \max_{i \in S_0} \frac{\langle Z_i, \mathbf{v}\rangle}{\sqrt{dn(p+q)/2(1+o(1))}} + b' &< \frac{K}{2\sqrt{dn(p+q)/2}} \\ \min_{i \in S_1} \frac{\langle Z_i, \mathbf{v}\rangle}{\sqrt{dn(p+q)/2(1+o(1))}} + b' &> -\frac{K}{2\sqrt{dn(p+q)/2}}. \end{aligned} \tag{24}$$

with probability $1 - n^{-c}$, where we have used here the degree concentration from (5) for $\delta = (\log n)^{-1/2+\epsilon}$ with $\epsilon > 0$ small enough. We see that such a pair exists if and only if one of the above holds with $b' = 0$. Comparing this with (12), we see that the remainder of the proof follows *mutatis mutandis*. □

5. Generalization

In this section we provide the proof for Theorem 2.

5.1. Characterizing the optimizer

We begin by characterizing $\mathbf{w}^*$, the optimizer of (2). Define the following quantities.

$$m_0 = \frac{p\boldsymbol{\mu} + q\boldsymbol{\nu}}{p+q}, \qquad m_1 = \frac{q\boldsymbol{\mu} + p\boldsymbol{\nu}}{p+q}. \tag{25}$$

Lemma 4. *For any $R > 0$, let $(\mathbf{w}^*(R), b^*(R))$ be the optimizer to the problem in (2) for a given training sample $(A, X) \sim \mathrm{CSBM}(n, p, q, \boldsymbol{\mu}, \boldsymbol{\nu})$ with $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^d$ and with norm constraint R. Then for any $c > 0$ fixed but large enough, with probability at least $1 - n^{-c}$ we have that for any $R > 0$,*

$$\mathbf{w}^*(R) = \frac{R}{2\gamma}(\boldsymbol{\nu} - \boldsymbol{\mu})(1 - o(1)), \tag{26}$$

where $\gamma = \frac{1}{2}\|\boldsymbol{\mu} - \boldsymbol{\nu}\|$ and that

$$\langle m_0, \mathbf{w}^*(R)\rangle + b^*(R) \le -R\gamma\Gamma(p,q)(1 - o(1)), \tag{27}$$

$$\langle m_1, \mathbf{w}^*(R)\rangle + b^*(R) \ge R\gamma\Gamma(p,q)(1 - o(1)). \tag{28}$$

Proof. Fix $R > 0$ and let $(\mathbf{w}^*(R), b^*(R))$ be the solutions to the problem in (2) with norm constraint R. Let the training sample be $(A, X) \sim \text{CSBM}(n, p, q, \boldsymbol{\mu}, \boldsymbol{\nu})$. Then we have that

$$\text{OPT}_d(A, X, R) = L(A, X, \mathbf{w}^*, b^*) \leq L(A, X, \tilde{\mathbf{w}}, \tilde{b}),$$

where $(\tilde{\mathbf{w}}, \tilde{b})$ are defined in Lemma 3. Let $\tilde{X}_i = (D^{-1}\tilde{A}X)_i$. Now we focus our scope to the event that for every $i \in [n]$, and $R > 0$

$$\langle \tilde{X}_i, \tilde{\mathbf{w}} \rangle + \tilde{b} = (2\varepsilon_i - 1)R\gamma\Gamma(p, q)(1 + o(1)).$$

Note that from (20) this event occurs with probability at least $1 - n^{-c}$ for c large but $O(1)$. Since $(\mathbf{w}^*, b^*)$ are solutions to (2), on this event we have for all i that

$$\langle \tilde{X}_i, \mathbf{w}^* \rangle + b^* \leq -R\gamma\Gamma(p, q)(1 - o(1)) \text{ for } \varepsilon_i = 0,$$
$$\langle \tilde{X}_i, \mathbf{w}^* \rangle + b^* \geq R\gamma\Gamma(p, q)(1 - o(1)) \text{ for } \varepsilon_i = 1.$$

Note that Lemma 2 implies that with probability at least $1 - n^{-c}$, for all i we also have

$$|\langle \tilde{X}_i - m_{\varepsilon_i}(1 + o(1)), \mathbf{w}^* \rangle| \leq O\left(\frac{\|\tilde{\mathbf{w}}\| d^\eta}{\sqrt{dn(p+q)}}\right)$$

Since $\|\mathbf{w}^*\| \leq R$, we conclude by triangle inequality that

$$\langle m_0, \mathbf{w}^* \rangle + b^* \leq -R\gamma\Gamma(p, q)(1 - o(1)),$$
$$\langle m_1, \mathbf{w}^* \rangle + b^* \geq R\gamma\Gamma(p, q)(1 - o(1)).$$

that is, (27) and (28) hold as desired. It remains to show (26).

Subtracting (27) from (28) we obtain

$$\langle m_1 - m_0, \mathbf{w}^* \rangle = \frac{p - q}{p + q}\langle \boldsymbol{\nu} - \boldsymbol{\mu}, \mathbf{w}^* \rangle \geq 2R\gamma\Gamma(p, q)(1 - o(1)). \tag{29}$$

This implies that $\|\mathbf{w}^*\| \geq R(1 - o(1))$. Since $\|\mathbf{w}^*\| \leq R$ due to the optimization constraint, we have that

$$1 - o(1) \leq \frac{\langle \boldsymbol{\nu} - \boldsymbol{\mu}, \mathbf{w}^* \rangle}{\|\boldsymbol{\mu} - \boldsymbol{\nu}\| \|\mathbf{w}^*\|} \leq 1$$

as desired. □

5.2. Proof of Theorem 2

Now we turn to the proof of Theorem 2.

Proof of Theorem 2. Consider a test sample $(A', X') \sim \text{CSBM}(n', p', q', \boldsymbol{\mu}, \boldsymbol{\nu})$. Let $\tilde{X}'$ be the corresponding convolution $D'^{-1}\tilde{A}'X'$. Similar to (17), (18) and (25) we also define $m'(i)$, m'_0 and m'_1 corresponding to the sample (A', X'). We restrict our calculations to the case where $\varepsilon_i = 0$. Note that

$$m'_0 - m_0 = \frac{qp' - pq'}{(p + q)(p' + q')}(\boldsymbol{\mu} - \boldsymbol{\nu}).$$

From Lemmas 2 and 4, we see that for $c' > 0$ large but $O(1)$, with probability at least $1 - (n')^{-c'}$ we have that for any $R > 0$

$$\langle \tilde{X}'_i, \mathbf{w}^* \rangle = \langle m'_{\varepsilon_i}, \mathbf{w}^* \rangle(1 + o(1)) + o(\frac{d^\eta}{\sqrt{d}})$$

for some $\eta > 0$ and for all $i \in [n']$.

Let $\gamma = \frac{\|\boldsymbol{\mu}-\boldsymbol{\nu}\|}{2}$. Therefore, by the same lemmas, we have that for any $c, c' > 0$ large enough, with probability $1 - O(1/(n')^{c'} + 1/n^c)$, when $\varepsilon_i = 0$

$$\begin{aligned}
\langle \tilde{X}_i', \mathbf{w}^* \rangle + b^* &= \langle m_0', \mathbf{w}^* \rangle (1 + o(1)) + b^* + o(\frac{d^\eta}{\sqrt{d}}) \\
&= \langle m_0' - m_0, \mathbf{w}^* \rangle (1 + o(1)) + \langle m_0, \mathbf{w}^* \rangle (1 + o(1)) + b^* + o(\frac{d^\eta}{\sqrt{d}}) \\
&\leq \frac{qp' - pq'}{(p+q)(p'+q')} \langle \boldsymbol{\mu} - \boldsymbol{\nu}, \mathbf{w}^* \rangle - R\gamma\Gamma(p,q)(1 - o(1)) \\
&\leq -\frac{4d\gamma^2(qp' - pq')}{2\gamma(p+q)(p'+q')}(1 - o(1)) - R\gamma\Gamma(p,q)(1 - o(1)) \\
&= R\gamma \left(\frac{2(pq' - qp')}{(p+q)(p'+q')} - \frac{p-q}{p+q} \right) (1 - o(1)) \\
&= -R\gamma\Gamma(p',q')(1 - o(1)).
\end{aligned}$$

The first inequality above uses (27), while the second inequality follows from Lemma 4. Similarly, for $\varepsilon_i = 1$ we obtain

$$\langle \tilde{X}_i', \mathbf{w}^* \rangle + b^* \geq R\gamma\Gamma(p',q')(1 - o(1)).$$

The loss is then given by

$$\begin{aligned}
L(A', X', \mathbf{w}^*, b^*) &= \frac{1}{n'} \sum_{i \in [n']} \log \left(1 + \exp \left((1 - 2\varepsilon_i)(\langle \tilde{X}', \mathbf{w}^* \rangle + b^*) \right) \right) \\
&\leq \frac{1}{n'} \sum_{i \in [n']} \log \left(1 + \exp \left(- R\gamma\Gamma(p',q')(1 - o(1)) \right) \right)
\end{aligned}$$

Now it follows from (21) that on this event

$$L(A', X', \mathbf{w}^*, b^*) \leq C \exp\left(-R\gamma\Gamma(p',q')(1 - o(1)) \right). \qquad \square$$

6. Additional Experiments

We now present additional experiments to support our results. Our conclusions are similar to those in the main paper. For the separability thresholds, we show an additional plot in Fig. 1 in the log scale which describes the decrease in the loss once the distance between the means crosses the corresponding threshold.

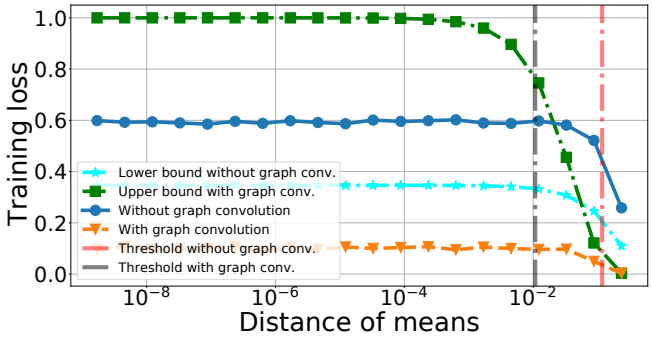
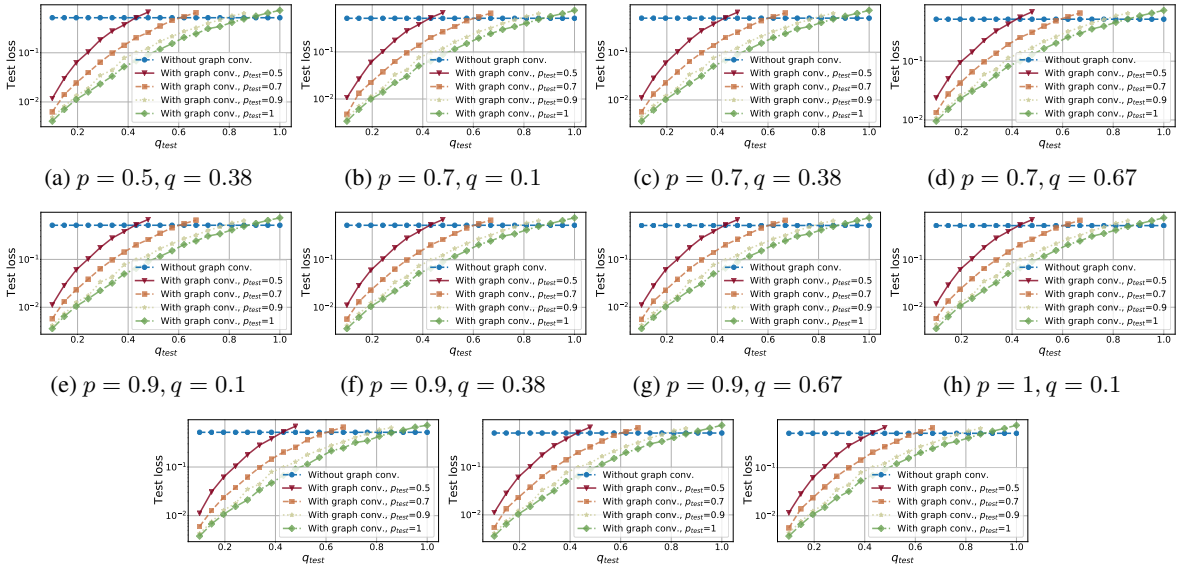


Figure 1. Training loss with/without graph convolution for increasing distance between the means. The vertical dashed red and black lines correspond to the separability thresholds from Parts 1 and 3 of Theorem 1, respectively. The green dashed line with square markers illustrates the theoretical rate from Theorem 2. The cyan dashed line with star markers corresponds to the lower bound from Part 1 of Theorem 1. We train and test on a CSBM with $p = 0.5$, $q = 0.1$, $n = 400$ and $d = 60$. The y-axis is in log-scale.

6.1. Out-of-distribution generalization

In this experiment we test the performance of the trained classifier on out-of-distribution datasets. We perform this experiment for two different distances between the means, $16/\sqrt{d}$ and $2/\sqrt{d}$. We train on a CSBM using various combinations of p_{train} and q_{train}, while $p_{train} > q_{train}$. In all experiments we set $n = 400$ and $d = 60$. We test on CSBMs with $n = 400$, $d = 60$ and varying p_{test} and q_{test} while $p_{test} > q_{test}$. The results for distance of means equal to $2/\sqrt{d}$ are presented in Figure 2, and the results for distance between the means equal to $16/\sqrt{d}$ are presented in Figure 3.



(a) $p = 0.5, q = 0.38$ (b) $p = 0.7, q = 0.1$ (c) $p = 0.7, q = 0.38$ (d) $p = 0.7, q = 0.67$

(e) $p = 0.9, q = 0.1$ (f) $p = 0.9, q = 0.38$ (g) $p = 0.9, q = 0.67$ (h) $p = 1, q = 0.1$

(i) $p = 1, q = 0.38$ (j) $p = 1, q = 0.67$ (k) $p = 1, q = 0.95$

Figure 2. Out-of-distribution generalization for distance between the means equal to $2/\sqrt{d}$. The subcaption of each figure is the p_{train} and q_{train} pair. Note that we omit the sub-index $train$ from p and q in the subcaption due to space limitation. We test on CSBMs with $n = 400$, $d = 60$ and varying p_{test} and q_{test} while $p_{test} > q_{test}$ and fixed means. The y-axis is in log-scale.

6.2. Out-of-distribution generalization on real data

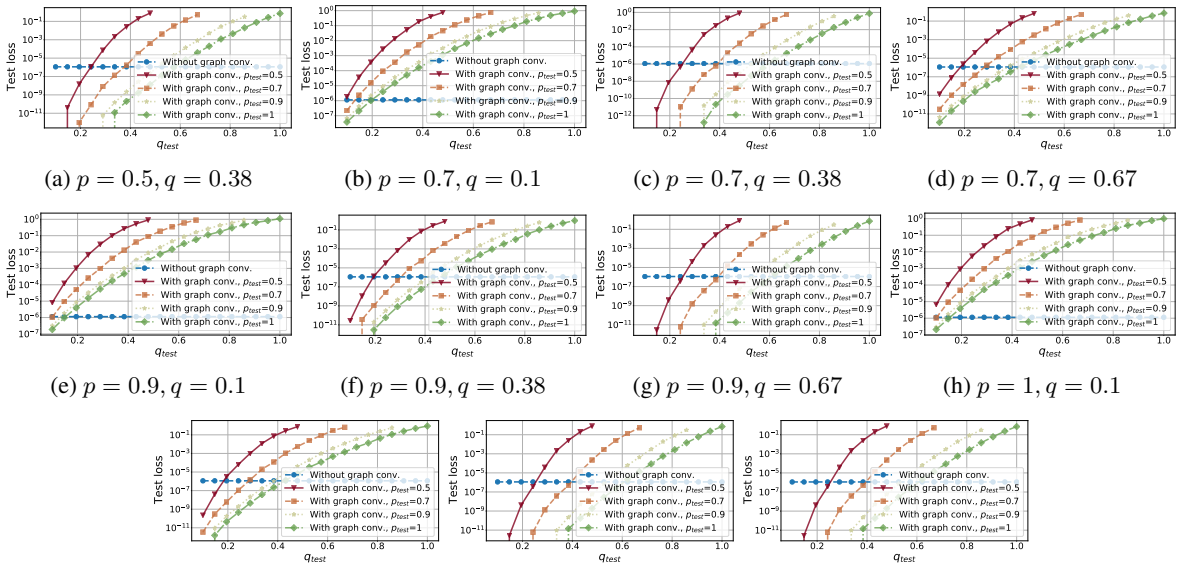
In this experiment we illustrate the generalization performance on real data for the linear classifier obtained by minimizing cross-entropy; see details about the optimization problem in the main paper. In particular, we use the partially labelled real data to train two linear classifiers, with and without graph convolution. We generate new graphs by adding inter-class edges uniformly at random. Then we test the performance of the trained classifiers on the noisy graphs with the original attributes. Therefore, the only thing that changes in the new unseen data are the graphs, the attributes remain the same.

We use the popular real data Cora, PubMed and WikipediaNetwork. These data are publicly available and can be downloaded from (Fey & Lenssen, 2019). The datasets come with multiple classes, however, for each of our experiments we do a one-v.s.-all classification for a single class. WikipediaNetwork comes with multiple masks for the labels, in our experiments we use the first mask. Moreover, this is a semi-supervised problem, meaning that only a fraction of the training nodes have labels. Details about the classes of the datasets that were omitted from the main paper are given in Table 1. The results of the experiments are shown in Figures 4 to 6.

References

Fey, M. and Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.



(a) $p = 0.5, q = 0.38$ (b) $p = 0.7, q = 0.1$ (c) $p = 0.7, q = 0.38$ (d) $p = 0.7, q = 0.67$

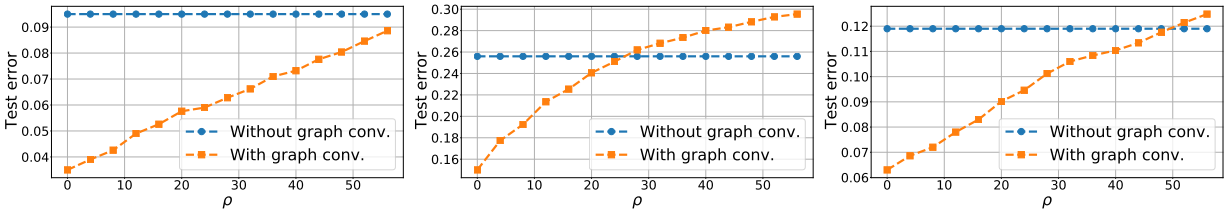
(e) $p = 0.9, q = 0.1$ (f) $p = 0.9, q = 0.38$ (g) $p = 0.9, q = 0.67$ (h) $p = 1, q = 0.1$

(i) $p = 1, q = 0.38$ (j) $p = 1, q = 0.67$ (k) $p = 1, q = 0.95$

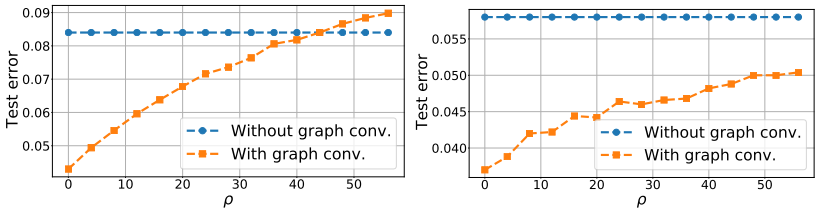
Figure 3. Out-of-distribution generalization for distance between the means equal to $16/\sqrt{d}$. The subcaption of each figure is the p_{train} and q_{train} pair. Note that we omit the sub-index $train$ from p and q in the subcaption due to space limitation. We test on CSBMs with $n = 400$, $d = 60$ and varying p_{test} and q_{test} while $p_{test} > q_{test}$ and fixed means. The y-axis is in log-scale.

Table 1. Information about the classes of the datasets. Here, the letter of the class refers to the original class of the dataset. Then number of nodes and attributes for Cora dataset is 2708 and 1433, respectively. The number of nodes and attributes for PubMed dataset is 19717 and 500, respectively. The number of nodes and attributes for Wiki.Net dataset is 2277 and 2325, respectively.

Dataset	Class	β_0	β_1	$\|\boldsymbol{\mu} - \boldsymbol{\nu}\|$
Cora	C	5.2e-02	4.8e-02	9.2e-01
	D	6.3e-02	2.4e-02	6.8e-01
	E	5.3e-02	4.7e-02	7.7e-01
	F	5.0e-02	6.7e-02	8.5e-01
	G	4.7e-02	1.1e-01	8.6e-01
PubMed	C	3.4e-03	2.5e-03	7.0e-02
Wiki.Net	C	4.8e-01	4.8e-01	2.1e-01
	D	4.8e-01	4.6e-01	4.3e-01
	E	4.8e-01	4.9e-01	5.3e-01

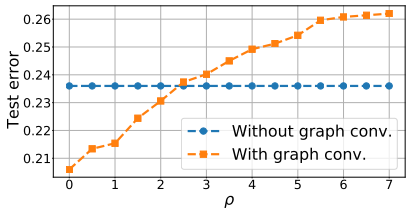


(a) Cora, class C (b) Cora, class D (c) Cora, class E



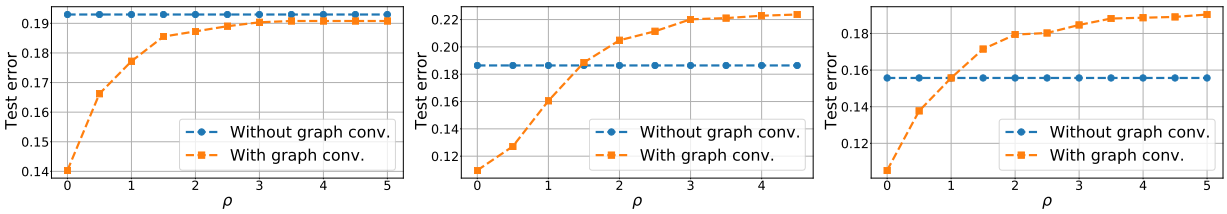
(d) Cora, class F (e) Cora, class G

Figure 4. Test loss as the number of nodes increases for Cora. The test error measures the number of misclassified nodes over the number of nodes in the graph. Here, ρ denotes the ratio of added inter-class edges over the number of inter-class edges of the original graph. The y-axis is in log-scale.



(a) PubMed, class C

Figure 5. Test loss as the number of nodes increases for PubMed.



(a) Wiki.Net, class C (b) Wiki.Net, class D (c) Wiki.Net, class E

Figure 6. Test loss as the number of nodes increases for Wiki.Net.