A APPENDIX

We summarize the notations used throughout the paper in Table 2. In the following sections, we provide proofs of all results in the main text and the additional details.

Symbol	Meaning
$\mathbb{R}$	the set of real numbers
$\mathbb{R}^{m\times n}$	the set of $m \times n$ real matrices
$\mathbb{N}_k^+$	the set of first k positive numbers
$\lvert\cdot\rvert_p$	the vector p-norm
$\lVert\cdot\rVert_p$	the operator norm induced by the vector p-norm
$\lVert\cdot\rVert_F$	the Frobenius norm
$X[i,j]$	the (i,j)-th element of matrix X
$X[i,:]$	the i-th row of matrix X
$X[:,i]$	the i-th column of matrix X
$\mathbf{1}_n$	a all-one vector with size n
A	an adjacency matrix
$\tilde{A}$	an adjacency matrix plus the identity matrix
B	the radius of the ℓ_2-ball where an input node feature lies
C_{in}	an incidence matrix of incoming nodes
C_{out}	an incidence matrix of outgoing nodes
ϕ, ρ, g	non-linearities in MPGNN
C_ϕ, C_ρ, C_g	Lipschitz constants of ϕ, ρ, g under the vector 2-norm
$\mathcal{C}$	the percolation complexity
D	the degree matrix
$\mathcal{D}$	the unknown data distribution
d	the maximum node degree plus one
e	the Euler's number
f_w	a model parameterized by vector w
$\mathcal{G}$	the space of graph
h	the maximum hidden dimension
H	a node representation matrix
$\mathcal{H}$	the hypothesis/model class
I	the identity matrix
l	the number of graph convolution layers / message passing steps
L	the loss function
$\tilde{L}$	the Laplacian matrix
m	the number of training samples
$\mathbb{P}$, $\mathbb{E}$	probability and expectation of a random variable
P	the prior distribution over hypothesis class
Q	the posterior distribution over hypothesis class
S	a set of training samples
W	a weight matrix
X	a node feature matrix where each row corresponds to a node
$\mathcal{X}$	the space of node feature
y	the graph class label
γ	the margin parameter
z	a data triplet (A, X, y)
$\mathcal{Z}$	the space of data triplet
log	the natural logarithm

Table 2: Summary of important notations.

A.1 PAC BAYES RESULTS

For completeness, we provide the proofs of the standard PAC-Bayes results as below.

Lemma A.1. *For non-negative continuous random variables X, we have*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq \nu) \mathrm{d}\nu.$$

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty X \mathbb{P}(X) \mathrm{d}X \\
&= \int_0^\infty \int_0^X \mathbf{1} \mathrm{d}\nu \mathbb{P}(X) \mathrm{d}X \\
&= \int_0^\infty \int_0^X \mathbb{P}(X) \mathrm{d}\nu \mathrm{d}X \\
&= \int_0^\infty \int_\nu^\infty \mathbb{P}(X) \mathrm{d}X \mathrm{d}\nu \qquad \text{(region of the integral is the same)} \\
&= \int_0^\infty \mathbb{P}(X \geq \nu) \mathrm{d}\nu
\end{aligned}$$

□

Lemma A.2. *[2-side] Let X be a random variable satisfying $\mathbb{P}(X \geq \epsilon) \leq e^{-2m\epsilon^2}$ and $\mathbb{P}(X \leq -\epsilon) \leq e^{-2m\epsilon^2}$ where $m \geq 1$ and $\epsilon > 0$, we have*

$$\mathbb{E}[e^{2(m-1)X^2}] \leq 2m.$$

Proof. If $m = 1$, the inequality holds trivially. Let us now consider $m > 1$.

$$\begin{aligned}
\mathbb{E}[e^{2(m-1)X^2}] &= \int_0^\infty \mathbb{P}\left(e^{2(m-1)X^2} \geq \nu\right) \mathrm{d}\nu \qquad \text{(Lemma A.1)} \\
&= \int_0^\infty \mathbb{P}\left(X^2 \geq \frac{\log \nu}{2(m-1)}\right) \mathrm{d}\nu \\
&= \int_0^\infty \mathbb{P}\left(X \geq \sqrt{\frac{\log \nu}{2(m-1)}}\right) \mathrm{d}\nu + \int_0^\infty \mathbb{P}\left(X \leq -\sqrt{\frac{\log \nu}{2(m-1)}}\right) \mathrm{d}\nu \qquad (5)
\end{aligned}$$

$$\begin{aligned}
\int_0^\infty \mathbb{P}\left(X \geq \sqrt{\frac{\log \nu}{2(m-1)}}\right) \mathrm{d}\nu &= \int_0^1 \mathbb{P}\left(X \geq \sqrt{\frac{\log \nu}{2(m-1)}}\right) \mathrm{d}\nu + \int_1^\infty \mathbb{P}\left(X \geq \sqrt{\frac{\log \nu}{2(m-1)}}\right) \mathrm{d}\nu \\
&\leq 1 + \int_1^\infty \mathbb{P}\left(X \geq \sqrt{\frac{\log \nu}{2(m-1)}}\right) \mathrm{d}\nu \\
&\leq 1 + \int_1^\infty e^{-2m\frac{\log \nu}{2(m-1)}} \mathrm{d}\nu \\
&= 1 + \left(-(m-1)\nu^{-\frac{1}{m-1}}\Big|_1^\infty\right) \\
&= m \qquad (6)
\end{aligned}$$

Similarly, we can show that

$$\int_0^\infty \mathbb{P}\left(X \leq -\sqrt{\frac{\log \nu}{2(m-1)}}\right) \mathrm{d}\nu \leq m \qquad (7)$$

Combining Eq. (5) and Eq. (6), we finish the proof. □

Theorem 2.1. *(Two-side) Let P be a prior distribution over $\mathcal{H}$ and let $\delta \in (0,1)$. Then, with probability $1-\delta$ over the choice of an i.i.d. training set S according to $\mathcal{D}$, for all distributions Q over $\mathcal{H}$ and any $\gamma > 0$, we have*

$$L_{\mathcal{D},\gamma}(Q) \leq L_{S,\gamma}(Q) + \sqrt{\frac{D_{\mathrm{KL}}(Q\|P) + \log\frac{2m}{\delta}}{2(m-1)}}$$

Proof. Let $\Delta(h) = L_{\mathcal{D},\gamma}(h) - L_{S,\gamma}(h)$. For any function $f(h)$, we have

$$\begin{aligned}
\mathbb{E}_{h\sim Q}[f(h)] &= \mathbb{E}_{h\sim Q}[\log e^{f(h)}] \\
&= \mathbb{E}_{h\sim Q}[\log e^{f(h)} + \log\frac{Q}{P} + \log\frac{P}{Q}] \\
&= D_{\mathrm{KL}}(Q\|P) + \mathbb{E}_{h\sim Q}\left[\log\left(\frac{P}{Q}e^{f(h)}\right)\right] \\
&\leq D_{\mathrm{KL}}(Q\|P) + \log\mathbb{E}_{h\sim Q}\left[\frac{P}{Q}e^{f(h)}\right] \qquad \text{(Jensen's inequality)} \\
&= D_{\mathrm{KL}}(Q\|P) + \log\mathbb{E}_{h\sim P}\left[e^{f(h)}\right]. \qquad (8)
\end{aligned}$$

Let $f(h) = 2(m-1)\Delta(h)^2$. We have

$$\begin{aligned}
2(m-1)\mathbb{E}_{h\sim Q}[\Delta(h)]^2 &\leq 2(m-1)\mathbb{E}_{h\sim Q}[\Delta(h)^2] \qquad \text{(Jensen's inequality)} \\
&\leq D_{\mathrm{KL}}(Q\|P) + \log\mathbb{E}_{h\sim P}\left[e^{2(m-1)\Delta(h)^2}\right]. \qquad (9)
\end{aligned}$$

Since $L_{\mathcal{D}}(h) \in [0,1]$, based on Hoeffding's inequality, for any $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{P}(\Delta(h) \geq \epsilon) &\leq e^{-2m\epsilon^2} \\
\mathbb{P}(\Delta(h) \leq -\epsilon) &\leq e^{-2m\epsilon^2}
\end{aligned}$$

Hence, based on Lemma A.2, we have

$$\begin{aligned}
\mathbb{E}_S\left[e^{2(m-1)\Delta(h)^2}\right] \leq 2m \;&\Rightarrow\; \mathbb{E}_{h\sim P}\left[\mathbb{E}_S\left[e^{2(m-1)\Delta(h)^2}\right]\right] \leq 2m \\
&\Leftrightarrow\; \mathbb{E}_S\left[\mathbb{E}_{h\sim P}\left[e^{2(m-1)\Delta(h)^2}\right]\right] \leq 2m
\end{aligned}$$

Based on Markov's inequality, we have

$$\mathbb{P}\left(\mathbb{E}_{h\sim P}\left[e^{2(m-1)\Delta(h)^2}\right] \geq \frac{2m}{\delta}\right) \leq \frac{\delta\mathbb{E}_S\left[\mathbb{E}_{h\sim P}\left[e^{2(m-1)\Delta(h)^2}\right]\right]}{2m} \leq \delta. \qquad (10)$$

Combining Eq. (9) and Eq. (10), with probability $1-\delta$, we have

$$\mathbb{E}_{h\sim Q}[\Delta(h)]^2 \leq \frac{D_{\mathrm{KL}}(Q\|P) + \log\left(\frac{2m}{\delta}\right)}{2(m-1)} \qquad (11)$$

which proves the theorem. □

Lemma 2.2. *Let $f_w(x) : \mathcal{X} \to \mathbb{R}^k$ be any model with parameters w, and P be any distribution on the parameters that is independent of the training data. For any w, we construct a posterior $Q(w+u)$ by adding any random perturbation u to w,* s.t., *$\mathbb{P}(\max_{x\in\mathcal{X}} |f_{w+u}(x) - f_w(x)|_\infty < \frac{\gamma}{4}) > \frac{1}{2}$. Then, for any $\gamma, \delta > 0$, with probability at least $1-\delta$ over an i.i.d. size-m training set S according to $\mathcal{D}$, for any w, we have:*

$$L_{\mathcal{D},0}(f_w) \leq L_{S,\gamma}(f_w) + \sqrt{\frac{2D_{\mathrm{KL}}(Q(w+u)\|P) + \log\frac{8m}{\delta}}{2(m-1)}}$$

Proof. Let $\tilde{w} = w + u$. Let $\mathcal{C}$ be the set of perturbation with the following property,

$$\mathcal{C} = \left\{ w' \middle| \max_{x \in \mathcal{X}} |f_{w'}(x) - f_w(x)|_\infty < \frac{\gamma}{4} \right\}. \tag{12}$$

$\tilde{w} = w + u$ (w is deterministic and u is stochastic) is distributed according to $Q(\tilde{w})$. We now construct a new posterior $\tilde{Q}$ as follows,

$$\tilde{Q}(\tilde{w}) = \begin{cases} \frac{1}{Z} Q(\tilde{w}) & \tilde{w} \in \mathcal{C} \\ 0 & \tilde{w} \in \bar{\mathcal{C}}. \end{cases} \tag{13}$$

Here $Z = \int_{\tilde{w} \in \mathcal{C}} \mathrm{d}Q(\tilde{w}) = \mathop{\mathbb{P}}_{\tilde{w} \sim Q}(\tilde{w} \in \mathcal{C})$ and $\bar{\mathcal{C}}$ is the complement set of $\mathcal{C}$. We know from the assumption that $Z > \frac{1}{2}$. Therefore, for any $\tilde{w} \sim \tilde{Q}$, we have

$$\begin{aligned}
&\max_{i \in \mathbb{N}_k^+, j \in \mathbb{N}_k^+, x \in \mathcal{X}} \left| |f_{\tilde{w}}(x)[i] - f_{\tilde{w}}(x)[j]| - |f_w(x)[i] - f_w(x)[j]| \right| \\
\leq &\max_{i \in \mathbb{N}_k^+, j \in \mathbb{N}_k^+, x \in \mathcal{X}} \left| f_{\tilde{w}}(x)[i] - f_{\tilde{w}}(x)[j] - f_w(x)[i] + f_w(x)[j] \right| \\
\leq &\max_{i \in \mathbb{N}_k^+, j \in \mathbb{N}_k^+, x \in \mathcal{X}} \left| f_{\tilde{w}}(x)[i] - f_w(x)[i] \right| + \left| f_{\tilde{w}}(x)[j] - f_w(x)[j] \right| \\
\leq &\max_{i \in \mathbb{N}_k^+, x \in \mathcal{X}} \left| f_{\tilde{w}}(x)[i] - f_w(x)[i] \right| + \max_{j \in \mathbb{N}_k^+, x \in \mathcal{X}} \left| f_{\tilde{w}}(x)[j] - f_w(x)[j] \right| \\
&< \frac{\gamma}{4} + \frac{\gamma}{4} = \frac{\gamma}{2}
\end{aligned} \tag{14}$$

Recall that

$$\begin{aligned}
L_{\mathcal{D}}(f_w, 0) &= \mathop{\mathbb{P}}_{z \sim \mathcal{D}} \left(f_w(x)[y] \leq \max_{j \neq y} f_w(x)[j] \right) \\
L_{\mathcal{D}}(f_{\tilde{w}}, \frac{\gamma}{2}) &= \mathop{\mathbb{P}}_{z \sim \mathcal{D}} \left(f_{\tilde{w}}(x)[y] \leq \frac{\gamma}{2} + \max_{j \neq y} f_{\tilde{w}}(x)[j] \right).
\end{aligned}$$

Denoting $j_1^* = \arg\max_{j \neq y} f_{\tilde{w}}(x)[j]$ and $j_2^* = \arg\max_{j \neq y} f_w(x)[j]$, from Eq. (14), we have

$$\begin{aligned}
&\left| f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_2^*] - f_w(x)[y] + f_w(x)[j_2^*] \right| < \frac{\gamma}{2} \\
\Rightarrow \; &f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_2^*] < f_w(x)[y] - f_w(x)[j_2^*] + \frac{\gamma}{2}
\end{aligned} \tag{15}$$

Note that since $f_{\tilde{w}}(x)[j_1^*] \geq f_{\tilde{w}}(x)[j_2^*]$, we have

$$\begin{aligned}
f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_1^*] &\leq f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_2^*] \\
&\leq f_w(x)[y] - f_w(x)[j_2^*] + \frac{\gamma}{2} \qquad \text{(Eq. (15))}
\end{aligned}$$

Therefore, we have

$$f_w(x)[y] - f_w(x)[j_2^*] \leq 0 \;\Rightarrow\; f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_1^*] \leq \frac{\gamma}{2},$$

which indicates $\mathop{\mathbb{P}}_{z \sim \mathcal{D}} (f_w(x)[y] \leq f_w(x)[j_2^*]) \leq \mathop{\mathbb{P}}_{z \sim \mathcal{D}} \left(f_{\tilde{w}}(x)[y] \leq f_{\tilde{w}}(x)[j_1^*] + \frac{\gamma}{2} \right)$, or equivalently

$$L_{\mathcal{D},0}(f_w) \leq L_{\mathcal{D},\frac{\gamma}{2}}(f_{\tilde{w}}). \tag{16}$$

Note that this holds for any perturbation $\tilde{w} \sim \tilde{Q}$.

Again, recall that

$$\begin{aligned}
L_{\mathcal{D},\frac{\gamma}{2}}(f_{\tilde{w}}) &= \mathop{\mathbb{P}}_{z \sim \mathcal{D}} \left(f_{\tilde{w}}(x)[y] \leq \frac{\gamma}{2} + \max_{j \neq y} f_{\tilde{w}}(x)[j] \right) \\
L_{\mathcal{D},\gamma}(f_w) &= \mathop{\mathbb{P}}_{z \sim \mathcal{D}} \left(f_w(x)[y] \leq \gamma + \max_{j \neq y} f_w(x)[j] \right)
\end{aligned}$$

From Eq. (14), we have

$$\begin{aligned} &\left|f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_1^*] - f_w(x)[y] + f_w(x)[j_1^*]\right| < \frac{\gamma}{2} \\ \Rightarrow\ & f_w(x)[y] - f_w(x)[j_1^*] < f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_1^*] + \frac{\gamma}{2} \end{aligned} \tag{17}$$

Note that since $f_w(x)[j_2^*] \geq f_w(x)[j_1^*]$, we have

$$\begin{aligned} f_w(x)[y] - f_w(x)[j_2^*] &\leq f_w(x)[y] - f_w(x)[j_1^*] \\ &\leq f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_1^*] + \frac{\gamma}{2} \qquad \text{(Eq. (17))} \end{aligned}$$

Therefore, we have

$$f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_1^*] \leq \frac{\gamma}{2} \ \Rightarrow\ f_w(x)[y] - f_w(x)[j_2^*] \leq \gamma,$$

which indicates $L_{\mathcal{D},\frac{\gamma}{2}}(f_{\tilde{w}}) \leq L_{\mathcal{D},\gamma}(f_w)$. Therefore, from the perspective of the empirical estimation of the probability, for any $\tilde{w} \sim \tilde{Q}$, we almost surely have

$$L_{S,\frac{\gamma}{2}}(f_{\tilde{w}}) \leq L_{S,\gamma}(f_w). \tag{18}$$

Now with probability at least $1-\delta$, we have

$$\begin{aligned} L_{\mathcal{D},0}(f_w) &\leq \mathbb{E}_{\tilde{w}\sim\tilde{Q}}\left[L_{\mathcal{D},\frac{\gamma}{2}}(f_{\tilde{w}})\right] \qquad \text{(Eq. (16))} \\ &\leq \mathbb{E}_{\tilde{w}\sim\tilde{Q}}\left[L_{S,\frac{\gamma}{2}}(f_{\tilde{w}})\right] + \sqrt{\frac{D_{\mathrm{KL}}(\tilde{Q}\|P) + \log\frac{2m}{\delta}}{2(m-1)}} \qquad \text{(Theorem 2.1)} \\ &\leq L_{S,\gamma}(f_w) + \sqrt{\frac{D_{\mathrm{KL}}(\tilde{Q}\|P) + \log\frac{2m}{\delta}}{2(m-1)}} \qquad \text{(Eq. (18))} \end{aligned} \tag{19}$$

Note that

$$\begin{aligned} D_{\mathrm{KL}}(Q\|P) &= \int_{\tilde{w}\in\mathcal{C}} Q\log\frac{Q}{P}\mathrm{d}\tilde{w} + \int_{\tilde{w}\in\bar{\mathcal{C}}} Q\log\frac{Q}{P}\mathrm{d}\tilde{w} \\ &= \int_{\tilde{w}\in\mathcal{C}} \frac{QZ}{Z}\log\frac{Q}{ZP}\mathrm{d}\tilde{w} + \int_{\tilde{w}\in\mathcal{C}} Q\log Z\mathrm{d}\tilde{w} \\ &\quad + \int_{\tilde{w}\in\bar{\mathcal{C}}} \frac{Q(1-Z)}{1-Z}\log\frac{Q}{(1-Z)P}\mathrm{d}\tilde{w} + \int_{\tilde{w}\in\bar{\mathcal{C}}} Q\log(1-Z)\mathrm{d}\tilde{w} \\ &= Z D_{\mathrm{KL}}(\tilde{Q}\|P) + (1-Z)D_{\mathrm{KL}}(\bar{Q}\|P) - H(Z), \end{aligned} \tag{20}$$

where $\bar{Q}$ denotes the normalized density of Q restricted to $\bar{\mathcal{C}}$. $H(Z)$ is the entropy of a Bernoulli random variable with parameter Z. Since we know $\frac{1}{2} \leq Z \leq 1$ from the beginning, $0 \leq H(Z) \leq \log 2$, and D_{KL} is nonnegative, from Eq. (20), we have

$$\begin{aligned} D_{\mathrm{KL}}(\tilde{Q}\|P) &= \frac{1}{Z}\left[D_{\mathrm{KL}}(Q\|P) + H(Z) - (1-Z)D_{\mathrm{KL}}(\bar{Q}\|P)\right] \\ &\leq \frac{1}{Z}\left[D_{\mathrm{KL}}(Q\|P) + H(Z)\right] \\ &\leq 2D_{\mathrm{KL}}(Q\|P) + 2\log 2. \end{aligned} \tag{21}$$

Combining Eq. (19) and Eq. (21), we have

$$L_{\mathcal{D},0}(f_w) \leq L_{S,\gamma}(f_w) + \sqrt{\frac{D_{\mathrm{KL}}(Q\|P) + \frac{1}{2}\log\frac{8m}{\delta}}{m-1}}, \tag{22}$$

which finishes the proof. □

A.2 GRAPH RESULTS

In this part, we provide a result on the graph Laplacian used by GCNs along with the proof. It is used in the perturbation analysis of GCNs.

Lemma A.3. *Let A be the binary adjacency matrix of an arbitrary simple graph $G = (V, E)$ and $\tilde{A} = A + I$. We define the graph Laplacian $L = D^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}}$ where D is the degree matrix of $\tilde{A}$. Then we have $\|L\|_1 = \|L\|_\infty \leq \sqrt{d}$, $\|L\|_2 \leq 1$, and $\|L\|_F \leq \sqrt{r}$ where r is the rank of L and $d-1$ is the maximum node degree of G.*

Proof. First, $\tilde{A}$ is symmetric and element-wise nonnegative. Denoting $n = |V|$, we have $\tilde{A} \in \mathbb{R}^{n\times n}$, $D_i = \sum_{j=1}^n \tilde{A}[i,j]$, and $1 \leq D_i \leq d, \forall i \in \mathbb{N}_n^+$. It is easy to show that $L[i,j] = \tilde{A}[i,j]/\sqrt{D_i D_j}$.

For the infinity norm and 1-norm, we have $\|L\|_1 = \|L^\top\|_\infty = \|L\|_\infty$. Moreover,

$$\begin{aligned}
\|L\|_\infty &= \max_{i\in\mathbb{N}_n^+} \sum_{j=1}^n |L[i,j]| \\
&= \max_{i\in\mathbb{N}_n^+} \sum_{j=1}^n \frac{\tilde{A}[i,j]}{\sqrt{D_i D_j}} \\
&\leq \max_{i\in\mathbb{N}_n^+} \frac{1}{\sqrt{D_i}} \sum_{j=1}^n \tilde{A}[i,j] \\
&= \max_{i\in\mathbb{N}_n^+} \sqrt{D_i} \\
&\leq \sqrt{d}
\end{aligned} \tag{23}$$

For the spectral norm, based on the definition, we have

$$\|L\|_2 = \sup_{x\neq 0} \frac{|Lx|_2}{|x|_2} = \sigma_{\max}, \tag{24}$$

where $\sigma_{\max}$ is the maximum singular value of L. Since L is symmetric, we have $\sigma_i = |\lambda_i|$ where λ_i is the i-th eigenvalue of L. Hence, $\sigma_{\max} = \max_i |\lambda_i|$. From Raylaigh quotient and Courant–Fischer minimax theorem, we have

$$\begin{aligned}
\|L\|_2 = \max_i |\lambda_i| &= \max_{x\neq 0} \left| \frac{x^\top L x}{x^\top x} \right| \\
&= \max_{x\neq 0} \left| \frac{\sum_{i=1}^n \sum_{j=1}^n L[i,j] x_i x_j}{\sum_{i=1}^n x_i^2} \right| \\
&= \max_{x\neq 0} \left| \frac{\sum_{i=1}^n \sum_{j=1}^n \tilde{A}[i,j] x_i x_j / \sqrt{D_i D_j}}{\sum_{i=1}^n x_i^2} \right| \\
&= \max_{x\neq 0} \left| \frac{\sum_{(i,j)\in\tilde{E}} x_i x_j / \sqrt{D_i D_j}}{\sum_{i=1}^n x_i^2} \right| \\
&\leq \max_{x\neq 0} \left| \frac{\frac{1}{2}\sum_{(i,j)\in\tilde{E}} \left(x_i^2/D_i + x_j^2/D_j \right)}{\sum_{i=1}^n x_i^2} \right| \\
&= \max_{x\neq 0} \left| \frac{\sum_{(i,j)\in\tilde{E}} x_i^2/D_i}{\sum_{i=1}^n x_i^2} \right| = \max_{x\neq 0} \left| \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} \right| = 1,
\end{aligned} \tag{25}$$

where $\tilde{E}$ is the union of the set of edges E in the original graph and the set of self-loops. For Frobenius norm, we have $\|L\|_F \leq \sqrt{r}\|L\|_2 \leq \sqrt{r}$ where r is the rank of L. □

A.3 GCN RESULTS

In this part, we provide the proofs of the main results regarding GCNs.

Lemma 3.1. *(GCN Perturbation Bound) For any $B > 0, l > 1$, let $f_w \in \mathcal{H} : \mathcal{X} \times \mathcal{G} \to \mathbb{R}^K$ be a l-layer GCN. Then for any w, and $x \in \mathcal{X}_{B,h_0}$, and any perturbation $u = vec(\{U_i\}_{i=1}^l)$ such that $\forall i \in \mathbb{N}_l^+$, $\|U_i\|_2 \leq \frac{1}{l}\|W_i\|_2$, the change in the output of GCN is bounded as,*

$$|f_{w+u}(X,A) - f_w(X,A)|_2 \leq eBd^{\frac{l-1}{2}} \left(\prod_{i=1}^{l} \|W_i\|_2\right) \sum_{k=1}^{l} \frac{\|U_k\|_2}{\|W_k\|_2}$$

Proof. We first perform the recursive perturbation analysis on node representations of all layers except the last one, *i.e.*, the readout layer. Then we derive the bound for the graph representation of the last readout layer.

Perturbation Analysis on Node Representations. In GCN, for any layer $j < l$ besides the last readout one, the node representations are,

$$f_w^j(X,A) = H_j = \sigma_j\left(\tilde{L}H_{j-1}W_j\right). \tag{26}$$

We add perturbation u to the weights w, *i.e.*, for the j-th layer, the perturbed weights are $W_j + U_j$. For the ease of notation, we use the superscript of prime to denote the perturbed node representations, *e.g.*, $H_j' = f_{w+u}^j(X,A)$. Let $\Delta_j = f_{w+u}^j(X,A) - f_w^j(X,A) = H_j' - H_j$. Note that $\Delta_j \in \mathbb{R}^{n \times h_j}$. Let $\Psi_j = \max_i |\Delta_j[i,:]|_2 = \max_i \left|H_j'[i,:] - H_j[i,:]\right|_2$ and $\Phi_j = \max_i |H_j[i,:]|_2$. We denote the $u_j^* = \arg\max_i |\Delta_j[i,:]|_2$ and $v_j^* = \arg\max_i |H_j[i,:]|_2$.

Upper Bound on the Max Node Representation For any layer $j < l$, we can derive an upper bound on the maximum (w.r.t. ℓ_2 norm) node representation as follows,

$$\begin{aligned}
\Phi_j = \max_i |H_j[i,:]|_2 &= \left|\left(\sigma_j\left(\tilde{L}H_{j-1}W_j\right)\right)[v_j^*,:]\right|_2 = \left|\sigma_j\left(\left(\tilde{L}H_{j-1}W_j\right)[v_j^*,:]\right)\right|_2 \\
&\leq \left|\left(\tilde{L}H_{j-1}W_j\right)[v_j^*,:]\right|_2 \qquad \text{(Lipschitz property of ReLU under vector 2-norm)} \\
&= \left|\left(\tilde{L}H_{j-1}\right)[v_j^*,:]W_j\right|_2 \\
&\leq \left|\left(\tilde{L}H_{j-1}\right)[v_j^*,:]\right|_2 \|W_j\|_2 = \left|\sum\nolimits_{k\in\mathcal{N}_{v_j^*}} \tilde{L}[v_j^*,k]H_{j-1}[k,:]\right|_2 \|W_j\|_2 \\
&\leq \sum\nolimits_{k\in\mathcal{N}_{v_j^*}} \tilde{L}[v_j^*,k]\,|H_{j-1}[k,:]|_2\,\|W_j\|_2 \\
&\leq \sum\nolimits_{k\in\mathcal{N}_{v_j^*}} \tilde{L}[v_j^*,k]\Phi_{j-1}\,\|W_j\|_2 \qquad \left(\text{since } \forall i, |H_{j-1}[i,:]|_2 \leq \Phi_{j-1}\right) \\
&\leq d^{\frac{1}{2}}\Phi_{j-1}\,\|W_j\|_2 \\
&\leq d^{\frac{j}{2}}\Phi_0 \prod_{i=1}^{j}\|W_i\|_2 \qquad \text{(unroll the recursion)} \\
&\leq d^{\frac{j}{2}}B\prod_{i=1}^{j}\|W_i\|_2,
\end{aligned} \tag{27}$$

where in the last inequality we use the fact $\Phi_0 = \max_i |X[i,:]|_2 \leq B$ based on the assumption A3. $\mathcal{N}_{v_j^*}$ is the set of neighboring nodes (including itself) of node v_j^*. In the third from the last inequality, we use the Lemma A.3 to derive the following fact that $\forall i$,

$$\sum_{k\in\mathcal{N}_i} \tilde{L}[i,k] = \sum_{k\in\mathcal{N}_i} \left|\tilde{L}[i,k]\right| \leq \left\|\tilde{L}\right\|_\infty \leq \sqrt{d}. \tag{28}$$

Upper Bound on the Max Change of Node Representation. For any layer $j < l$, we can derive an upper bound on the maximum (w.r.t. ℓ_2 norm) change between the representations with and without the weight perturbation for any node as follows,

$$
\begin{aligned}
\Psi_j &= \max_i \left| H'_j[i,:] - H_j[i,:] \right|_2 = \left| \sigma_j \left(\tilde{L} H'_{j-1} (W_j + U_j) \right) [u_j^*, :] - \sigma_j \left(\tilde{L} H_{j-1} W_j \right) [u_j^*, :] \right|_2 \\
&\leq \left| \left(\tilde{L} H'_{j-1} (W_j + U_j) \right) [u_j^*, :] - \left(\tilde{L} H_{j-1} W_j \right) [u_j^*, :] \right|_2 \qquad \text{(Lipschitz property of ReLU)} \\
&= \left| \left((\tilde{L} H'_{j-1})[u_j^*, :] \right) (W_j + U_j) - \left((\tilde{L} H_{j-1})[u_j^*, :] \right) W_j \right|_2 \\
&= \left| \left(\left((\tilde{L} H'_{j-1})[u_j^*, :] \right) - \left((\tilde{L} H_{j-1})[u_j^*, :] \right) \right) (W_j + U_j) + \left((\tilde{L} H_{j-1})[u_j^*, :] \right) U_j \right|_2 \\
&= \left| \left(\sum_{k \in \mathcal{N}_{u_j^*}} \tilde{L}[u_j^*, k] \left(H'_{j-1}[k,:] - H_{j-1}[k,:] \right) \right) (W_j + U_j) + \left(\sum_{k \in \mathcal{N}_{u_j^*}} \tilde{L}[u_j^*, k] H_{j-1}[k,:] \right) U_j \right|_2 \\
&\leq \left| \sum_{k \in \mathcal{N}_{u_j^*}} \tilde{L}[u_j^*, k] \left(H'_{j-1}[k,:] - H_{j-1}[k,:] \right) \right|_2 \|W_j + U_j\|_2 + \left| \sum_{k \in \mathcal{N}_{u_j^*}} \tilde{L}[u_j^*, k] H_{j-1}[k,:] \right|_2 \|U_j\|_2 \\
&\leq \sum_{k \in \mathcal{N}_{u_j^*}} \tilde{L}[u_j^*, k] \left| H'_{j-1}[k,:] - H_{j-1}[k,:] \right|_2 \|W_j + U_j\|_2 + \sum_{k \in \mathcal{N}_{u_j^*}} \tilde{L}[u_j^*, k] \left| H_{j-1}[k,:] \right|_2 \|U_j\|_2 \\
&\leq \sum_{k \in \mathcal{N}_{u_j^*}} \tilde{L}[u_j^*, k] \Psi_{j-1} \|W_j + U_j\|_2 + \sum_{k \in \mathcal{N}_{u_j^*}} \tilde{L}[u_j^*, k] \Phi_{j-1} \|U_j\|_2 \\
&\leq \sqrt{d} \Psi_{j-1} \|W_j + U_j\|_2 + \sqrt{d} \Phi_{j-1} \|U_j\|_2 ,
\end{aligned} \tag{29}
$$

where in the second from the last inequality we use the fact $\forall k$, $\left| H'_{j-1}[k,:] - H_{j-1}[k,:] \right|_2 \leq \Psi_{j-1}$ and $\forall k$, $\left| H_{j-1}[k,:] \right|_2 \leq \Phi_{j-1}$. In the last inequality, we again use the fact in Eq. (28). We can simplify the notations in Eq. (29) as $\Psi_j \leq a_{j-1} \Psi_{j-1} + b_{j-1}$ where $a_{j-1} = \sqrt{d} \|W_j + U_j\|_2$ and $b_{j-1} = \sqrt{d} \Phi_{j-1} \|U_j\|_2$. Since $\Delta_0 = X - X = \mathbf{0}$, we have $\Psi_0 = 0$. It is straightforward to work out the recursion as,

$$
\begin{aligned}
\Psi_j &\leq \sum_{k=0}^{j-1} b_k \left(\prod_{i=k+1}^{j-1} a_i \right) = \sum_{k=0}^{j-1} d^{\frac{1}{2}} \Phi_k \|U_{k+1}\|_2 \left(\prod_{i=k+1}^{j-1} d^{\frac{1}{2}} \|W_{i+1} + U_{i+1}\|_2 \right) \\
&= \sum_{k=0}^{j-1} d^{\frac{j-k}{2}} \Phi_k \|U_{k+1}\|_2 \left(\prod_{i=k+2}^{j} \|W_i + U_i\|_2 \right) .
\end{aligned} \tag{30}
$$

Based on Eq. (27), we can instantiate the bound in Eq. (30) as

$$
\begin{aligned}
\Psi_j &\leq \sum_{k=0}^{j-1} d^{\frac{j-k}{2}} \Phi_k \|U_{k+1}\|_2 \left(\prod_{i=k+2}^{j} \|W_i + U_i\|_2 \right) \\
&\leq \sum_{k=0}^{j-1} d^{\frac{j-k}{2}} \left(d^{\frac{k}{2}} B \prod_{i=1}^{k} \|W_i\|_2 \right) \|U_{k+1}\|_2 \left(\prod_{i=k+2}^{j} (\|W_i\|_2 + \|U_i\|_2) \right) \\
&\leq B \sum_{k=0}^{j-1} d^{\frac{j}{2}} \left(\prod_{i=1}^{k} \|W_i\|_2 \right) \|U_{k+1}\|_2 \left(\prod_{i=k+2}^{j} \left(1 + \frac{1}{l} \right) \|W_i\|_2 \right) \\
&= B \sum_{k=0}^{j-1} d^{\frac{j}{2}} \left(\prod_{i=1}^{k+1} \|W_i\|_2 \right) \frac{\|U_{k+1}\|_2}{\|W_{k+1}\|_2} \left(\prod_{i=k+2}^{j} \left(1 + \frac{1}{l} \right) \|W_i\|_2 \right) \\
&= B d^{\frac{j}{2}} \left(\prod_{i=1}^{j} \|W_i\|_2 \right) \sum_{k=0}^{j-1} \frac{\|U_{k+1}\|_2}{\|W_{k+1}\|_2} \left(1 + \frac{1}{l} \right)^{j-k-1} \\
&\leq B d^{\frac{j}{2}} \left(\prod_{i=1}^{j} \|W_i\|_2 \right) \sum_{k=1}^{j} \frac{\|U_k\|_2}{\|W_k\|_2} \left(1 + \frac{1}{l} \right)^{j-k}
\end{aligned} \tag{31}
$$

Final Bound on the Readout Layer Now let us consider the average readout function in the last layer, *i.e.*, the l-th layer. Based on Eq. (27) and Eq. (31), we can bound the change of GCN's output with and without the weight perturbation as follows,

$$
\begin{aligned}
|\Delta_l|_2 =& \left| \frac{1}{n} \mathbf{1}_n H'_{l-1}(W_l + U_l) - \frac{1}{n} \mathbf{1}_n H_{l-1} W_l \right|_2 \\
=& \left| \frac{1}{n} \mathbf{1}_n \Delta_{l-1}(W_l + U_l) + \frac{1}{n} \mathbf{1}_n H_{l-1} U_l \right|_2 \\
\leq& \frac{1}{n} \left| \mathbf{1}_n \Delta_{l-1}(W_l + U_l) \right|_2 + \frac{1}{n} \left| \mathbf{1}_n H_{l-1} U_l \right|_2 \\
\leq& \frac{1}{n} \|W_l + U_l\|_2 |\mathbf{1}_n \Delta_{l-1}|_2 + \frac{1}{n} \|U_l\|_2 |\mathbf{1}_n H_{l-1}|_2 \\
=& \frac{1}{n} \|W_l + U_l\|_2 \left| \sum_{i=1}^{n} \Delta_{l-1}[i,:] \right|_2 + \frac{1}{n} \|U_l\|_2 \left| \sum_{i=1}^{n} H_{l-1}[i,:] \right|_2 \\
\leq& \frac{1}{n} \|W_l + U_l\|_2 \left(\sum_{i=1}^{n} |\Delta_{l-1}[i,:]|_2 \right) + \frac{1}{n} \|U_l\|_2 \left(\sum_{i=1}^{n} |H_{l-1}[i,:]|_2 \right) \\
\leq& \|W_l + U_l\|_2 \Psi_{l-1} + \|U_l\|_2 \Phi_{l-1} \\
\leq& \|W_l + U_l\|_2 B d^{\frac{l-1}{2}} \left(\prod_{i=1}^{l-1} \|W_i\|_2 \right) \sum_{k=1}^{l-1} \frac{\|U_k\|_2}{\|W_k\|_2} \left(1 + \frac{1}{l} \right)^{l-1-k} + \|U_l\|_2 B d^{\frac{l-1}{2}} \prod_{i=1}^{l-1} \|W_i\|_2 \\
=& B d^{\frac{l-1}{2}} \left[\|W_l + U_l\|_2 \left(\prod_{i=1}^{l-1} \|W_i\|_2 \right) \sum_{k=1}^{l-1} \frac{\|U_k\|_2}{\|W_k\|_2} \left(1 + \frac{1}{l} \right)^{l-1-k} + \|U_l\|_2 \prod_{i=1}^{l-1} \|W_i\|_2 \right] \\
=& B d^{\frac{l-1}{2}} \left(\prod_{i=1}^{l} \|W_i\|_2 \right) \left[\frac{\|W_l + U_l\|_2}{\|W_l\|_2} \sum_{k=1}^{l-1} \frac{\|U_k\|_2}{\|W_k\|_2} \left(1 + \frac{1}{l} \right)^{l-1-k} + \frac{\|U_l\|_2}{\|W_l\|_2} \right] \\
\leq& B d^{\frac{l-1}{2}} \left(\prod_{i=1}^{l} \|W_i\|_2 \right) \left[\left(1 + \frac{1}{l} \right) \sum_{k=1}^{l-1} \frac{\|U_k\|_2}{\|W_k\|_2} \left(1 + \frac{1}{l} \right)^{l-1-k} + \frac{\|U_l\|_2}{\|W_l\|_2} \right] \\
=& B d^{\frac{l-1}{2}} \left(\prod_{i=1}^{l} \|W_i\|_2 \right) \left(1 + \frac{1}{l} \right)^{l} \left[\sum_{k=1}^{l-1} \frac{\|U_k\|_2}{\|W_k\|_2} \left(1 + \frac{1}{l} \right)^{-k} + \frac{\|U_l\|_2}{\|W_l\|_2} \left(1 + \frac{1}{l} \right)^{-l} \right] \\
\leq& e B d^{\frac{l-1}{2}} \left(\prod_{i=1}^{l} \|W_i\|_2 \right) \left[\sum_{k=1}^{l} \frac{\|U_k\|_2}{\|W_k\|_2} \right] \qquad \left(\text{Use } 1 \leq \left(1 + \frac{1}{l} \right)^{l} \leq e \right)
\end{aligned} \tag{32}
$$

which proves the lemma. □

Theorem 3.2. *(GCN Generalization Bound) For any $B > 0, l > 1$, let $f_w \in \mathcal{H} : \mathcal{X} \times \mathcal{G} \to \mathbb{R}^K$ be a l-layer GCN. Then for any $\delta, \gamma > 0$, with probability at least $1 - \delta$ over the choice of an i.i.d. size-m training set S according to $\mathcal{D}$, for any w, we have,*

$$
L_{\mathcal{D},0}(f_w) \leq L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{ \frac{B^2 d^{l-1} l^2 h \log(lh) \prod_{i=1}^{l} \|W_i\|_2^2 \sum_{i=1}^{l} \frac{\|W_i\|_F^2}{\|W_i\|_2^2} + \log \frac{ml}{\delta} }{\gamma^2 m} } \right)
$$

Proof. Let $\beta = \left(\prod_{i=1}^{l} \|W_i\|_2 \right)^{1/l}$. We normalize the weights as $\tilde{W}_i = \frac{\beta}{\|W_i\|_2} W_i$. Due to the homogeneity of ReLU, *i.e.*, $a\phi(x) = \phi(ax), \forall a \geq 0$, we have $f_w = f_{\tilde{w}}$. We can also verify that $\prod_{i=1}^{l} \|W_i\|_2 = \prod_{i=1}^{l} \|\tilde{W}_i\|_2$ and $\|W_i\|_F / \|W_i\|_2 = \|\tilde{W}_i\|_F / \|\tilde{W}_i\|_2$, *i.e.*, the terms appear in the bound stay the same after applying the normalization. Therefore, w.l.o.g., we assume that the norm is equal across layers, *i.e.*, $\forall i, \|W_i\|_2 = \beta$.

Consider the prior $P = \mathcal{N}(0, \sigma^2 I)$ and the random perturbation $u \sim \mathcal{N}(0, \sigma^2 I)$. Note that the σ of the prior and the perturbation are the same and will be set according to β. More precisely, we will set the σ based on some approximation $\tilde{\beta}$ of β since the prior P can not depend on any learned weights directly. The approximation $\tilde{\beta}$ is chosen to be a cover set which covers the meaningful range of β. For now, let us assume that we have a fix $\tilde{\beta}$ and consider β which satisfies $|\beta - \tilde{\beta}| \leq \frac{1}{l}\beta$. Note that this also implies

$$
\begin{aligned}
|\beta - \tilde{\beta}| \leq \frac{1}{l}\beta \;\Rightarrow\; & \left(1 - \frac{1}{l}\right)\beta \leq \tilde{\beta} \leq \left(1 + \frac{1}{l}\right)\beta \\
\Rightarrow\; & \left(1 - \frac{1}{l}\right)^{l-1}\beta^{l-1} \leq \tilde{\beta}^{l-1} \leq \left(1 + \frac{1}{l}\right)^{l-1}\beta^{l-1} \\
\Rightarrow\; & \left(1 - \frac{1}{l}\right)^{l}\beta^{l-1} \leq \tilde{\beta}^{l-1} \leq \left(1 + \frac{1}{l}\right)^{l}\beta^{l-1} \\
\Rightarrow\; & \frac{1}{e}\beta^{l-1} \leq \tilde{\beta}^{l-1} \leq e\beta^{l-1}
\end{aligned} \tag{33}
$$

From Tropp (2012), for $U_i \in \mathbb{R}^{h \times h}$ and $U_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$, we have,

$$
\mathbb{P}\left(\|U_i\|_2 \geq t\right) \leq 2he^{-t^2/2h\sigma^2}. \tag{34}
$$

Taking a union bound, we have

$$
\begin{aligned}
\mathbb{P}\left(\|U_1\|_2 < t \;\&\; \cdots \;\&\; \|U_l\|_2 < t\right) &= 1 - \mathbb{P}\left(\exists i, \|U_i\|_2 \geq t\right) \\
&\geq 1 - \sum_{i=1}^{l} \mathbb{P}\left(\|U_i\|_2 \geq t\right) \\
&\geq 1 - 2lhe^{-t^2/2h\sigma^2}.
\end{aligned} \tag{35}
$$

Setting $2lhe^{-t^2/2h\sigma^2} = \frac{1}{2}$, we have $t = \sigma\sqrt{2h\log(4lh)}$. This implies that the probability that the spectral norm of the perturbation of any layer is no larger than $\sigma\sqrt{2h\log(4lh)}$ holds with probability at least $\frac{1}{2}$. Plugging this bound into Lemma 3.1, we have with probability at least $\frac{1}{2}$,

$$
\begin{aligned}
|f_{w+u}(X, A) - f_w(X, A)|_2 &\leq eBd^{\frac{l-1}{2}}\left(\prod_{i=1}^{l}\|W_i\|_2\right)\sum_{k=1}^{l}\frac{\|U_k\|_2}{\|W_k\|_2} \\
&= eBd^{\frac{l-1}{2}}\beta^l\sum_{k=1}^{l}\frac{\|U_k\|_2}{\beta} \\
&\leq eBd^{\frac{l-1}{2}}\beta^{l-1}l\sigma\sqrt{2h\log(4lh)} \\
&\leq e^2Bd^{\frac{l-1}{2}}\tilde{\beta}^{l-1}l\sigma\sqrt{2h\log(4lh)} \leq \frac{\gamma}{4},
\end{aligned} \tag{36}
$$

where we can set $\sigma = \frac{\gamma}{42Bd^{\frac{l-1}{2}}\tilde{\beta}^{l-1}l\sqrt{h\log(4lh)}}$ to get the last inequality. Note that Lemma 3.1 also requires $\forall i \in \mathbb{N}_l^+$, $\|U_i\|_2 \leq \frac{1}{l}\|W_i\|_2$. The requirement is satisfied if $\sigma \leq \frac{\beta}{l\sqrt{2h\log(4lh)}}$ which in turn can be satisfied if

$$
\frac{\gamma}{4eBd^{\frac{l-1}{2}}\beta^{l-1}l\sqrt{2h\log(4lh)}} \leq \frac{\beta}{l\sqrt{2h\log(4lh)}}, \tag{37}
$$

since the chosen value of σ satisfies $\sigma \leq \frac{\gamma}{4eBd^{\frac{l-1}{2}}\beta^{l-1}l\sqrt{2h\log(4lh)}}$. Note that Eq. (37) is equivalent to $\frac{\gamma}{4eB}d^{\frac{1-l}{2}} \leq \beta^l$. We will see how to satisfy this condition later.

We now compute the KL term in the PAC-Bayes bound in Lemma 2.2.

$$\begin{aligned}\mathrm{KL}\left(Q\|P\right) = \frac{|w|_2^2}{2\sigma^2} &= \frac{42^2 B^2 d^{l-1}\tilde{\beta}^{2l-2} l^2 h \log(4lh)}{2\gamma^2}\sum_{i=1}^{l}\|W_i\|_F^2 \\ &\leq \mathcal{O}\left(\frac{B^2 d^{l-1}\beta^{2l} l^2 h\log(lh)}{\gamma^2}\sum_{i=1}^{l}\frac{\|W_i\|_F^2}{\beta^2}\right) \\ &\leq \mathcal{O}\left(B^2 d^{l-1} l^2 h\log(lh)\frac{\prod_{i=1}^{l}\|W_i\|_2^2}{\gamma^2}\sum_{i=1}^{l}\frac{\|W_i\|_F^2}{\|W_i\|_2^2}\right). \end{aligned} \tag{38}$$

From Lemma 2.2, fixing any $\tilde{\beta}$, with probability $1-\delta$ and for all w such that $|\beta-\tilde{\beta}| \leq \frac{1}{l}\beta$, we have,

$$L_{\mathcal{D},0}(f_w) \leq L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2 d^{l-1} l^2 h\log(lh)\prod_{i=1}^{l}\|W_i\|_2^2\sum_{i=1}^{l}\frac{\|W_i\|_F^2}{\|W_i\|_2^2} + \log\frac{m}{\delta}}{\gamma^2 m}}\right). \tag{39}$$

Finally, we need to consider multiple choices of $\tilde{\beta}$ so that for any β, we can bound the generalization error like Eq. (39). First, we only need to consider values of β in the following range,

$$\frac{1}{\sqrt{d}}\left(\frac{\gamma\sqrt{d}}{2B}\right)^{1/l} \leq \beta \leq \frac{1}{\sqrt{d}}\left(\frac{\gamma\sqrt{md}}{2B}\right)^{1/l}, \tag{40}$$

since otherwise the bound holds trivially as $L_{\mathcal{D},0}(f_w) \leq 1$ by definition. Note that the lower bound in Eq. (40) ensures that Eq. (37) holds which in turn justifies the applicability of Lemma 3.1. If $\beta < \frac{1}{\sqrt{d}}\left(\frac{\gamma\sqrt{d}}{2B}\right)^{1/l}$, then for any (X, A) and any $j \in \mathbb{N}_K^+$, $|f(X,A)[j]| \leq \frac{\gamma}{2}$. To see this, we have,

$$\begin{aligned}|f_w(X,A)[j]| \leq |f_w(X,A)|_2 &= |\frac{1}{n}\mathbf{1}_n H_{l-1} W_l|_2 \\ &\leq \frac{1}{n}|\mathbf{1}_n H_{l-1}|_2\|W_l\|_2 \\ &\leq \|W_l\|_2 \max_i |H_{l-1}[i,:]|_2 \\ &\leq B d^{\frac{l-1}{2}}\prod_{i=1}^{l}\|W_i\|_2 = d^{\frac{l-1}{2}}\beta^l B \qquad \text{(Use Eq. (27))} \\ &= d^{\frac{l-1}{2}} B\frac{\gamma}{2Bd^{\frac{l-1}{2}}} \leq \frac{\gamma}{2}. \end{aligned} \tag{41}$$

Therefore, by the definition in Eq. (4), we always have $L_{S,\gamma}(f_w) = 1$ when $\beta < \frac{1}{\sqrt{d}}\left(\frac{\gamma\sqrt{d}}{2B}\right)^{1/l}$. Alternatively, if $\beta > \frac{1}{\sqrt{d}}\left(\frac{\gamma\sqrt{md}}{2B}\right)^{1/l}$, the term inside the big-O notation in Eq. (39) would be,

$$\begin{aligned}\sqrt{\frac{B^2 d^{l-1} l^2 h\log(lh)\prod_{i=1}^{l}\|W_i\|_2^2\sum_{i=1}^{l}\frac{\|W_i\|_F^2}{\|W_i\|_2^2} + \log\frac{m}{\delta}}{\gamma^2 m}} &\geq \sqrt{\frac{l^2 h\log(lh)}{4}\sum_{i=1}^{l}\frac{\|W_i\|_F^2}{\|W_i\|_2^2}} \\ &\geq \sqrt{\frac{l^2 h\log(lh)}{4}} \geq 1, \end{aligned} \tag{42}$$

where we use the facts that $\|W_i\|_F \geq \|W_i\|_2$ and we typically choose $h \geq 2$ in practice and $l \geq 2$. Since we only need to consider β in the range of Eq. (40), a sufficient condition to make $|\beta-\tilde{\beta}| \leq \frac{1}{l}\beta$ hold would be $|\beta-\tilde{\beta}| \leq \frac{1}{l\sqrt{d}}\left(\frac{\gamma\sqrt{d}}{2B}\right)^{1/l}$. Therefore, if we can find a covering of the interval in Eq.

(40) with radius $\frac{1}{l\sqrt{d}}\left(\frac{\gamma\sqrt{d}}{2B}\right)^{1/l}$ and make sure bounds like Eq. (39) holds while $\tilde{\beta}$ takes all possible values from the covering, then we can get a bound which holds for all β. It is clear that we only need to consider a covering C with size $|C| = \frac{l}{2}\left(m^{\frac{1}{2l}} - 1\right)$. Therefore, denoting the event of Eq. (39) with $\tilde{\beta}$ taking the i-th value of the covering as E_i, we have

$$\mathbb{P}\left(E_1 \,\&\, \cdots \,\&\, E_{|C|}\right) = 1 - \mathbb{P}\left(\exists i, \bar{E}_i\right) \geq 1 - \sum_{i=1}^{|C|} \mathbb{P}\left(\bar{E}_i\right) \geq 1 - |C|\delta. \tag{43}$$

Note $\bar{E}_i$ denotes the complement of E_i. Hence, with probability $1-\delta$ and for all w, we have,

$$\begin{aligned} L_{\mathcal{D},0}(f_w) &\leq L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2 d^{l-1} l^2 h \log(lh) \prod_{i=1}^{l} \|W_i\|_2^2 \sum_{i=1}^{l} \frac{\|W_i\|_F^2}{\|W_i\|_2^2} + \log\frac{m|C|}{\delta}}{\gamma^2 m}}\right) \\ &= L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2 d^{l-1} l^2 h \log(lh) \prod_{i=1}^{l} \|W_i\|_2^2 \sum_{i=1}^{l} \frac{\|W_i\|_F^2}{\|W_i\|_2^2} + \log\frac{ml}{\delta}}{\gamma^2 m}}\right), \end{aligned} \tag{44}$$

which proves the theorem. □

A.4 MPGNNs Results

In this part, we provide the proofs of the main results regarding MPGNNs.

Lemma 3.3. *(MPGNN Perturbation Bound) For any $B > 0, l > 1$, let $f_w \in \mathcal{H} : \mathcal{X} \times \mathcal{G} \to \mathbb{R}^K$ be a l-step MPGNN. Then for any w, and $x \in \mathcal{X}_{B,h_0}$, and any perturbation $u = vec(\{U_1, U_2, U_l\})$ such that $\eta = \max\left(\frac{\|U_1\|_2}{\|W_1\|_2}, \frac{\|U_2\|_2}{\|W_2\|_2}, \frac{\|U_l\|_2}{\|W_l\|_2}\right) \leq \frac{1}{l}$, the change in the output of MPGNN is bounded as,*

$$|f_{w+u}(X,A) - f_w(X,A)|_2 \leq \begin{cases} eB\left(l+1\right)^2 \eta \|W_1\|_2 \|W_l\|_2 C_\phi, & \text{if } d\mathcal{C} = 1 \\ eBl\eta \|W_1\|_2 \|W_l\|_2 C_\phi \frac{(d\mathcal{C})^{l-1}-1}{d\mathcal{C}-1}, & \text{otherwise} \end{cases}$$

where $\mathcal{C} = C_\phi C_\rho C_g \|W_2\|_2$.

Proof. We first perform the recursive perturbation analysis on node representations of all steps except the last one, i.e., the readout step. Then we derive the bound for the graph representation of the last readout step.

Perturbation Analysis on Node Representations. In message passing GNNs, for any step $j < l$ besides the last readout one, the node representations are,

$$\begin{aligned} \bar{M}_j &= C_{\text{in}} g\left(C_{\text{out}}^\top H_{j-1}\right) \\ H_j &= \phi\left(XW_1 + \rho\left(\bar{M}_j\right) W_2\right), \end{aligned} \tag{45}$$

where the incidence matrices $C_{\text{in}} \in \mathbb{R}^{n\times c}$ and $C_{\text{out}} \in \mathbb{R}^{n\times c}$ (recall c is the number of edges). Moreover, since each edge only connects one incoming and one outgoing node, we have,

$$\begin{aligned} \sum_{k=1}^{c} C_{\text{in}}[i,k] &\leq \max_i \sum_{k=1}^{c} C_{\text{in}}[i,k] = \|C_{\text{in}}\|_\infty \leq d \\ \sum_{t=1}^{n} C_{\text{out}}[t,k] &\leq \max_k \sum_{t=1}^{n} C_{\text{out}}[t,k] \leq \|C_{\text{out}}\|_1 \leq 1 \end{aligned} \tag{46}$$

where $d-1$ is the maximum node degree. Note that one actually has $\|C_{\text{in}}\|_\infty \leq d-1$ for simple graphs. Since some models in the literature pre-process the graphs by adding self-loops, we thus relax it to $\|C_{\text{in}}\|_\infty \leq d$ which holds in both cases.

We add perturbation u to the weights w, *i.e.*, the perturbed weights are $W_1 + U_1$, $W_2 + U_2$ and $W_l + U_l$. For the ease of notation, we use the superscript of prime to denote the perturbed node representations, *e.g.*, $H'_j = f^j_{w+u}(X, A)$. Let $\Delta_j = f^j_{w+u}(X, A) - f^j_w(X, A) = H'_j - H_j$. Note that $\Delta_j \in \mathbb{R}^{n \times h_j}$. Let $\Psi_j = \max_i |\Delta_j[i,:]|_2 = \max_i \left|H'_j[i,:] - H_j[i,:]\right|_2$ and $\Phi_j = \max_i |H_j[i,:]|_2$. We denote the $u^*_j = \arg\max_i |\Delta_j[i,:]|_2$ and $v^*_j = \arg\max_i |H_j[i,:]|_2$. To simplify the derivation, we abbreviate the following statistics $\kappa = C_\phi B \|W_1\|_2$ and $\tau = d\mathcal{C}$ throughout the proof where $\mathcal{C} = C_\phi C_\rho C_g \|W_2\|_2$ is the *percolation complexity*.

Upper Bound on the Max Node Representation. For any step $j < l$, we can derive an upper bound on the ℓ_2 norm of the aggregated message of any node i as follows,

$$
\begin{aligned}
\left|\bar{M}_j[i,:]\right|_2 &= \left|\sum_{k=1}^{c} C_{\text{in}}[i,k] \left(g\left(C^\top_{\text{out}} H_{j-1}\right)\right)[k,:]\right|_2 = \left|\sum_{k=1}^{c} C_{\text{in}}[i,k] g\left(C^\top_{\text{out}}[k,:] H_{j-1}\right)\right|_2 \\
&\le \sum_{k=1}^{c} C_{\text{in}}[i,k] \left|g\left(C^\top_{\text{out}}[k,:] H_{j-1}\right)\right|_2 \\
&\le \sum_{k=1}^{c} C_{\text{in}}[i,k] C_g \left|C^\top_{\text{out}}[k,:] H_{j-1}\right|_2 = \sum_{k=1}^{c} C_{\text{in}}[i,k] C_g \left|\sum_{t=1}^{n} C^\top_{\text{out}}[k,t] H_{j-1}[t,:]\right|_2 \\
&\le \sum_{k=1}^{c} C_{\text{in}}[i,k] C_g \left(\sum_{t=1}^{n} C_{\text{out}}[t,k] \left|H_{j-1}[t,:]\right|_2\right) \\
&\le \sum_{k=1}^{c} C_{\text{in}}[i,k] C_g \left(\sum_{t=1}^{n} C_{\text{out}}[t,k] \Phi_{j-1}\right) \\
&\le d C_g \Phi_{j-1}.
\end{aligned} \tag{47}
$$

Then we can derive an upper bound on the maximum (w.r.t. ℓ_2 norm) node representation as follows,

$$
\begin{aligned}
\Phi_j &= \max_i |H_j[i,:]|_2 = \left|\phi\left(XW_1 + \rho\left(\bar{M}_j\right) W_2\right)[v^*_j,:]\right|_2 \\
&= \left|\phi\left(\left(XW_1 + \rho\left(\bar{M}_j\right) W_2\right)[v^*_j,:]\right)\right|_2 \\
&\le C_\phi \left|\left(XW_1 + \rho\left(\bar{M}_j\right) W_2\right)[v^*_j,:]\right|_2 \\
&= C_\phi \left|\left(XW_1\right)[v^*_j,:] + \left(\rho\left(\bar{M}_j\right) W_2\right)[v^*_j,:]\right|_2 \\
&\le C_\phi \left|\left(XW_1\right)[v^*_j,:]\right|_2 + C_\phi \left|\left(\rho\left(\bar{M}_j\right) W_2\right)[v^*_j,:]\right|_2 \\
&= C_\phi \left|X[v^*_j,:] W_1\right|_2 + C_\phi \left|\rho\left(\bar{M}_j\right)[v^*_j,:] W_2\right|_2 \\
&\le C_\phi \left|X[v^*_j,:]\right|_2 \|W_1\|_2 + C_\phi \left|\rho\left(\bar{M}_j\right)[v^*_j,:]\right|_2 \|W_2\|_2 \\
&\le C_\phi B \|W_1\|_2 + C_\phi \left|\rho\left(\bar{M}_j[v^*_j,:]\right)\right|_2 \|W_2\|_2 \\
&\le C_\phi B \|W_1\|_2 + C_\phi C_\rho \left|\bar{M}_j[v^*_j,:]\right|_2 \|W_2\|_2 \\
&\le C_\phi B \|W_1\|_2 + d C_\phi C_\rho C_g \Phi_{j-1} \|W_2\|_2 = \kappa + \tau \Phi_{j-1} \\
&\le \tau^j \Phi_0 + \sum_{i=0}^{j-1} \tau^{j-1-i} \kappa \qquad \text{(Unroll recursion)} \\
&= \sum_{i=0}^{j-1} \tau^{j-1-i} \kappa \qquad \text{(Use } \Phi_0 = 0\text{)} \\
&= \begin{cases} j\kappa, & \text{if } \tau = 1 \\ \kappa \frac{\tau^j - 1}{\tau - 1}, & \text{otherwise} \end{cases}
\end{aligned} \tag{48}
$$

Upper Bound on the Max Change of Node Representation. For any step $j < l$, we can derive an upper bound on the maximum (w.r.t. ℓ_2 norm) change between the aggregated message with and without the weight perturbation for any node i as follows,

$$
\begin{aligned}
\left|\bar{M}'_j[i,:] - \bar{M}_j[i,:]\right|_2 &\leq \left|\left(C_{\text{in}} g\left(C_{\text{out}}^\top H'_{j-1}\right)\right)[i,:] - \left(C_{\text{in}} g\left(C_{\text{out}}^\top H_{j-1}\right)\right)[i,:]\right|_2 \\
&= \left|\sum_{k=1}^{c} C_{\text{in}}[i,k]\left(g\left(C_{\text{out}}^\top H'_{j-1}\right) - g\left(C_{\text{out}}^\top H_{j-1}\right)\right)[k,:]\right|_2 \\
&\leq \sum_{k=1}^{c} C_{\text{in}}[i,k]\left|\left(g\left(C_{\text{out}}^\top H'_{j-1}\right) - g\left(C_{\text{out}}^\top H_{j-1}\right)\right)[k,:]\right|_2 \\
&= \sum_{k=1}^{c} C_{\text{in}}[i,k]\left|g\left(\left(C_{\text{out}}^\top H'_{j-1}\right)[k,:]\right) - g\left(\left(C_{\text{out}}^\top H_{j-1}\right)[k,:]\right)\right|_2 \\
&\leq \sum_{k=1}^{c} C_{\text{in}}[i,k] C_g \left|\left(C_{\text{out}}^\top H'_{j-1}\right)[k,:] - \left(C_{\text{out}}^\top H_{j-1}\right)[k,:]\right|_2 \\
&= \sum_{k=1}^{c} C_{\text{in}}[i,k] C_g \left|\sum_{t=1}^{n} C_{\text{out}}[t,k] H'_{j-1}[t,:] - \sum_{t=1}^{n} C_{\text{out}}[t,k] H_{j-1}[t,:]\right|_2 \\
&= \sum_{k=1}^{c} C_{\text{in}}[i,k] C_g \left|\sum_{t=1}^{n} C_{\text{out}}[t,k]\left(H'_{j-1}[t,:] - H_{j-1}[t,:]\right)\right|_2 \\
&\leq \sum_{k=1}^{c} C_{\text{in}}[i,k] C_g \left(\sum_{t=1}^{n} C_{\text{out}}[t,k]\left|H'_{j-1}[t,:] - H_{j-1}[t,:]\right|_2\right) \\
&\leq d C_g \Psi_{j-1}
\end{aligned} \tag{49}
$$

Based on Eq. (49), we can derive an upper bound on the maximum (w.r.t. ℓ_2 norm) change between the representations with and without the weight perturbation for any node as follows,

$$
\begin{aligned}
\Psi_j =& \max_i \left|H'_j[i,:] - H_j[i,:]\right|_2 \\
=& \left|\left(\phi\left(X(W_1+U_1) + \rho\left(\bar{M}'_j\right)(W_2+U_2)\right)\right)[u_j^*,:] - \left(\phi\left(XW_1 + \rho\left(\bar{M}_j\right)W_2\right)\right)[u_j^*,:]\right|_2 \\
=& \left|\phi\left(\left(X(W_1+U_1) + \rho\left(\bar{M}'_j\right)(W_2+U_2)\right)[u_j^*,:]\right) - \phi\left(\left(XW_1 + \rho\left(\bar{M}_j\right)W_2\right)[u_j^*,:]\right)\right|_2 \\
\leq& C_\phi \left|\left(X(W_1+U_1) + \rho\left(\bar{M}'_j\right)(W_2+U_2)\right)[u_j^*,:] - \left(XW_1 + \rho\left(\bar{M}_j\right)W_2\right)[u_j^*,:]\right|_2 \\
\leq& C_\phi \left|X[u_j^*,:]U_1 + \left(\rho\left(\bar{M}'_j\right)\right)[u_j^*,:](W_2+U_2) - \left(\rho\left(\bar{M}_j\right)\right)[u_j^*,:]W_2\right|_2 \\
=& C_\phi \left|X[u_j^*,:]U_1 + \left(\rho\left(\bar{M}'_j\right) - \rho\left(\bar{M}_j\right)\right)[u_j^*,:](W_2+U_2) + \rho\left(\bar{M}_j\right)[u_j^*,:]U_2\right|_2 \\
\leq& C_\phi B \left\|U_1\right\|_2 + C_\phi C_\rho \left|\bar{M}'_j[u_j^*,:] - \bar{M}_j[u_j^*,:]\right|_2 \left\|W_2+U_2\right\|_2 + C_\phi C_\rho \left|\bar{M}_j[u_j^*,:]\right|_2 \left\|U_2\right\|_2 \\
\leq& \kappa \frac{\|U_1\|_2}{\|W_1\|_2} + d\mathcal{C}\Psi_{j-1}\frac{\|W_2+U_2\|_2}{\|W_2\|_2} + d\mathcal{C}\Phi_{j-1}\frac{\|U_2\|_2}{\|W_2\|_2} \qquad \text{(Use Eq. (47) and (49))} \\
\leq& \tau\left(1 + \frac{\|U_2\|_2}{\|W_2\|_2}\right)\Psi_{j-1} + \kappa\frac{\|U_1\|_2}{\|W_1\|_2} + \tau\Phi_{j-1}\frac{\|U_2\|_2}{\|W_2\|_2}
\end{aligned} \tag{50}
$$

If $\tau = 1$, then we have,

$$
\begin{aligned}
\Psi_j \leq& \tau\left(1 + \frac{\|U_2\|_2}{\|W_2\|_2}\right)\Psi_{j-1} + \kappa\frac{\|U_1\|_2}{\|W_1\|_2} + \tau\Phi_{j-1}\frac{\|U_2\|_2}{\|W_2\|_2} \\
\leq& \left(1 + \frac{\|U_2\|_2}{\|W_2\|_2}\right)\Psi_{j-1} + \kappa\left(\frac{\|U_1\|_2}{\|W_1\|_2} + \frac{\|U_2\|_2}{\|W_2\|_2}(j-1)\right) \qquad \text{(Use Eq. (48))} \\
\leq& (1+\eta)\Psi_{j-1} + \kappa\eta(1 + (j-1)) \qquad \left(\text{Use } \eta = \max\left(\frac{\|U_1\|_2}{\|W_1\|_2}, \frac{\|U_2\|_2}{\|W_2\|_2}, \frac{\|U_l\|_2}{\|W_l\|_2}\right)\right) \\
=& (1+\eta)\Psi_{j-1} + \kappa\eta j.
\end{aligned} \tag{51}
$$

If $\tau \neq 1$, then we have,

$$\begin{aligned}
\Psi_j \leq & \tau\left(1+\frac{\|U_2\|_2}{\|W_2\|_2}\right)\Psi_{j-1}+\kappa\frac{\|U_1\|_2}{\|W_1\|_2}+\tau\Phi_{j-1}\frac{\|U_2\|_2}{\|W_2\|_2} \\
\leq & \tau\left(1+\frac{\|U_2\|_2}{\|W_2\|_2}\right)\Psi_{j-1}+\kappa\left(\frac{\|U_1\|_2}{\|W_1\|_2}+\tau\frac{\|U_2\|_2}{\|W_2\|_2}\frac{\tau^{j-1}-1}{\tau-1}\right) \qquad \text{(Use Eq. (48))} \\
\leq & \tau\left(1+\eta\right)\Psi_{j-1}+\kappa\eta\left(1+\frac{\tau^j-\tau}{\tau-1}\right) \qquad \left(\text{Use } \eta=\max\left(\frac{\|U_1\|_2}{\|W_1\|_2},\frac{\|U_2\|_2}{\|W_2\|_2},\frac{\|U_l\|_2}{\|W_l\|_2}\right)\right) \\
= & \tau\left(1+\eta\right)\Psi_{j-1}+\kappa\eta\left(\frac{\tau^j-1}{\tau-1}\right). \qquad (52)
\end{aligned}$$

Recall from Eq. (30), if $\Psi_j \leq a_{j-1}\Psi_{j-1}+b_{j-1}$ and $\Psi_0=0$, then $\Psi_j \leq \sum_{k=0}^{j-1} b_k\left(\prod_{i=k+1}^{j-1} a_i\right)$. If $\tau=1$, then we have $a_{j-1}=1+\eta$, $b_{j-1}=\kappa\eta j$ in our case and,

$$\begin{aligned}
\Psi_j \leq & \sum_{k=0}^{j-1} b_k\left(\prod_{i=k+1}^{j-1} a_i\right)=\sum_{k=0}^{j-1}\kappa\eta(k+1)\left(1+\eta\right)^{j-k-1} \\
\leq & \kappa\eta\left(1+\frac{1}{l}\right)^j\sum_{k=0}^{j-1}(k+1)\left(1+\frac{1}{l}\right)^{-k-1} \qquad \left(\text{Use } \eta\leq\frac{1}{l}\right) \\
= & \kappa\eta\left(1+\frac{1}{l}\right)^j\sum_{k=1}^{j}k\left(1+\frac{1}{l}\right)^{-k} \\
= & \kappa\eta\left(1+\frac{1}{l}\right)^j\left(1+\frac{1}{l}\right)^{-1}\frac{1-(j+1)\left(1+\frac{1}{l}\right)^{-j}+j\left(1+\frac{1}{l}\right)^{-j-1}}{\left(1-\left(1+\frac{1}{l}\right)^{-1}\right)^2} \\
= & \kappa\eta\frac{\left(1+\frac{1}{l}\right)^{j+1}-(j+1)\left(1+\frac{1}{l}\right)+j}{\left(\left(1+\frac{1}{l}\right)^1-1\right)^2} \\
= & \kappa\eta l^2\left(\left(1+\frac{1}{l}\right)^{j+1}-(j+1)\left(1+\frac{1}{l}\right)+j\right) \\
\leq & \kappa\eta l^2\left(\left(1+\frac{1}{l}\right)^{j+1}-1\right) \\
\leq & \kappa\eta l\left(l+1\right)\left(1+\frac{1}{l}\right)^j \qquad (53)
\end{aligned}$$

If $\tau \neq 1$, then we have $a_{j-1}=\tau\left(1+\eta\right)$, $b_{j-1}=\kappa\eta\left(\frac{\tau^j-1}{\tau-1}\right)$ in our case and,

$$\begin{aligned}
\Psi_j \leq & \sum_{k=0}^{j-1} b_k\left(\prod_{i=k+1}^{j-1} a_i\right)=\sum_{k=0}^{j-1}\kappa\eta\left(\frac{\tau^{k+1}-1}{\tau-1}\right)\tau^{j-k-1}\left(1+\eta\right)^{j-k-1} \\
\leq & \kappa\eta\tau^j\left(1+\frac{1}{l}\right)^j\sum_{k=0}^{j-1}\left(\frac{\tau^{k+1}-1}{\tau-1}\right)\tau^{-k-1}\left(1+\frac{1}{l}\right)^{-k-1} \qquad \left(\text{Use } \eta\leq\frac{1}{l}\right) \\
\leq & \kappa\eta\tau^j\left(1+\frac{1}{l}\right)^j\sum_{k=0}^{j-1}\left(\frac{1-\tau^{-k-1}}{\tau-1}\right)\left(1+\frac{1}{l}\right)^{-k-1} \\
\leq & \frac{\kappa\eta\tau^j}{\tau-1}\left(1+\frac{1}{l}\right)^j\sum_{k=0}^{j-1}\left(1-\tau^{-k-1}\right)\left(1+\frac{1}{l}\right)^{-k-1} \\
\leq & \frac{\kappa\eta\tau^j}{\tau-1}\left(1+\frac{1}{l}\right)^j\sum_{k=1}^{j}\left(1-\tau^{-k}\right) \qquad (54)
\end{aligned}$$

Final Bound with Readout Function Now let us consider the readout function. Since the last readout layer produces a vector in $\mathbb{R}^{1\times C}$, we have,

$$\begin{aligned}
|\Delta_l|_2 =& \left|\frac{1}{n}\mathbf{1}_n H'_{l-1}(W_l+U_l) - \frac{1}{n}\mathbf{1}_n H_{l-1}W_l\right|_2 \\
=& \left|\frac{1}{n}\mathbf{1}_n \Delta_{l-1}(W_l+U_l) + \frac{1}{n}\mathbf{1}_n H_{l-1}U_l\right|_2 \\
\leq& \frac{1}{n}\left|\mathbf{1}_n\Delta_{l-1}(W_l+U_l)\right|_2 + \frac{1}{n}\left|\mathbf{1}_n H_{l-1}U_l\right|_2 \\
\leq& \frac{1}{n}\|W_l+U_l\|_2|\mathbf{1}_n\Delta_{l-1}|_2 + \frac{1}{n}\|U_l\|_2|\mathbf{1}_n H_{l-1}|_2 \\
\leq& \|W_l+U_l\|_2\Psi_{l-1} + \|U_l\|_2\Phi_{l-1}
\end{aligned} \tag{55}$$

If $\tau = 1$, we have,

$$\begin{aligned}
|\Delta_l|_2 \leq& \|W_l\|_2\left(1+\frac{1}{l}\right)\kappa\eta l\,(l+1)\left(1+\frac{1}{l}\right)^{l-1} + (l-1)\kappa\|U_l\|_2 \qquad \text{(Use Eq. (48), (53))} \\
\leq& \|W_l\|_2\left(1+\frac{1}{l}\right)^{l}\kappa\left(\eta l\,(l+1) + (l-1)\frac{\|U_l\|_2}{\|W_l\|_2}\left(1+\frac{1}{l}\right)^{-l}\right) \\
\leq& \|W_l\|_2 e\kappa\eta\left(l(l+1)+(l-1)\right) \\
=& \|W_l\|_2 e\kappa\eta\left(l^2+2l-1\right) \\
\leq& \|W_l\|_2 e\kappa\eta\,(l+1)^2
\end{aligned} \tag{56}$$

Otherwise, we have,

$$\begin{aligned}
|\Delta_l|_2 \leq& \|W_l\|_2\frac{\kappa\eta\tau^{l-1}}{\tau-1}\left(1+\frac{1}{l}\right)^{l}\sum_{k=1}^{l-1}\left(1-\tau^{-k}\right) + \kappa\|U_l\|_2\frac{\tau^{l-1}-1}{\tau-1} \qquad \text{(Use Eq. (48), (54))} \\
\leq& \|W_l\|_2\frac{\kappa\tau^{l-1}}{\tau-1}\left(1+\frac{1}{l}\right)^{l}\left(\eta\sum_{k=1}^{l-1}\left(1-\tau^{-k}\right) + \frac{\|U_l\|_2}{\|W_l\|_2}(1-\tau^{1-l})\right)
\end{aligned} \tag{57}$$

If $\tau > 1$, then $\frac{1-\tau^{-k}}{\tau-1} \leq \frac{1-\tau^{1-l}}{\tau-1}$ when $1 \leq k \leq l-1$. If $\tau < 1$, we also have $\frac{1-\tau^{-k}}{\tau-1} \leq \frac{1-\tau^{1-l}}{\tau-1}$ when $1 \leq k \leq l-1$. Therefore, Eq. (57) can be further relaxed as,

$$\begin{aligned}
|\Delta_l|_2 \leq& \|W_l\|_2\frac{\kappa\tau^{l-1}}{\tau-1}\left(1+\frac{1}{l}\right)^{l}\left(\eta\sum_{k=1}^{l-1}\left(1-\tau^{-k}\right) + \frac{\|U_l\|_2}{\|W_l\|_2}(1-\tau^{1-l})\right) \\
=& \|W_l\|_2\kappa\tau^{l-1}\left(1+\frac{1}{l}\right)^{l}\left(\eta\sum_{k=1}^{l-1}\frac{1-\tau^{-k}}{\tau-1} + \frac{\|U_l\|_2}{\|W_l\|_2}\frac{1-\tau^{1-l}}{\tau-1}\right) \\
\leq& \|W_l\|_2\kappa\tau^{l-1}e\left(\eta(l-1)\frac{(1-\tau^{1-l})}{\tau-1} + \frac{\|U_l\|_2}{\|W_l\|_2}\frac{(1-\tau^{1-l})}{\tau-1}\right) \\
\leq& \|W_l\|_2\kappa\tau^{l-1}e\eta l\frac{(1-\tau^{1-l})}{\tau-1} \qquad \left(\text{Use } \frac{\|U_l\|_2}{\|W_l\|_2} \leq \eta\right) \\
=& e\eta l\kappa\|W_l\|_2\frac{\tau^{l-1}-1}{\tau-1},
\end{aligned} \tag{58}$$

Therefore, combining Eq. (56) and Eq. (58), we have,

$$|\Delta_l|_2 \leq \begin{cases} e\eta\kappa\,(l+1)^2\,\|W_l\|_2, & \text{if } d\mathcal{C}=1 \\ e\eta\kappa l\|W_l\|_2\frac{\tau^{l-1}-1}{\tau-1}, & \text{otherwise.}\end{cases} \tag{59}$$

which proves the lemma. □

Theorem 3.4. *(MPGNN Generalization Bound) For any $B > 0, l > 1$, let $f_w \in \mathcal{H} : \mathcal{X} \times \mathcal{G} \to \mathbb{R}^K$ be a l-step MPGNN. Then for any $\delta, \gamma > 0$, with probability at least $1 - \delta$ over the choice of an i.i.d. size-m training set S according to $\mathcal{D}$, for any w, we have,*

1. *If $d\mathcal{C} \neq 1$, then*

$$L_{\mathcal{D},0}(f_w) \leq L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2 \left(\max\left(\zeta^{-(l+1)}, (\lambda\xi)^{(l+1)/l}\right)\right)^2 l^2 h \log(lh) |w|_2^2 + \log\frac{m(l+1)}{\delta}}{\gamma^2 m}}\right).$$

2. *If $d\mathcal{C} = 1$, then*

$$L_{\mathcal{D},0}(f_w) \leq L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2 \max\left(\zeta^{-6}, \lambda^3 C_\phi^3\right) (l+1)^4 h \log(lh) |w|_2^2 + \log\frac{m}{\delta}}{\gamma^2 m}}\right).$$

where $\zeta = \min\left(\|W_1\|_2, \|W_2\|_2, \|W_l\|_2\right)$, $|w|_2^2 = \|W_1\|_F^2 + \|W_2\|_F^2 + \|W_l\|_F^2$, $\mathcal{C} = C_\phi C_\rho C_g \|W_2\|_2$, $\lambda = \|W_1\|_2 \|W_l\|_2$, and $\xi = C_\phi \frac{(d\mathcal{C})^{l-1} - 1}{d\mathcal{C} - 1}$.

Proof. We will derive the results conditioning on the value of $d\mathcal{C}$.

General Case $d\mathcal{C} \neq 1$ We first consider the general case $d\mathcal{C} \neq 1$. To derive the generalization bound, we construct a special statistic of the learned weights $\beta = \max\left(\frac{1}{\zeta}, (\lambda\xi)^{\frac{1}{l}}\right)$. It is clear that $\frac{1}{\zeta} \leq \beta$, $\lambda\xi \leq \beta^l$, and $\lambda\xi/\zeta \leq \beta^{l+1}$. Note that $\frac{1}{\zeta} = \max\left(\frac{1}{\|W_1\|_2}, \frac{1}{\|W_2\|_2}, \frac{1}{\|W_l\|_2}\right)$.

Consider the prior $P = \mathcal{N}(0, \sigma^2 I)$ and the random perturbation $u \sim \mathcal{N}(0, \sigma^2 I)$. Note that the σ of the prior and the perturbation are the same and will be set according to β. More precisely, we will set the σ based on some approximation $\tilde{\beta}$ of β since the prior P can not depend on any learned weights directly. The approximation $\tilde{\beta}$ is chosen to be a cover set which covers the meaningful range of β. For now, let us fix any $\tilde{\beta}$ and consider β which satisfies $|\beta - \tilde{\beta}| \leq \frac{1}{l+1}\beta$. This also implies,

$$\begin{aligned}
|\beta - \tilde{\beta}| \leq \frac{1}{l+1}\beta \;&\Rightarrow\; \left(1 - \frac{1}{l+1}\right)\beta \leq \tilde{\beta} \leq \left(1 + \frac{1}{l+1}\right)\beta \\
&\Rightarrow\; \left(1 - \frac{1}{l+1}\right)^{l+1} \beta^{l+1} \leq \tilde{\beta}^{l+1} \leq \left(1 + \frac{1}{l+1}\right)^{l+1} \beta^{l+1} \\
&\Rightarrow\; \frac{1}{e}\beta^{l+1} \leq \tilde{\beta}^{l+1} \leq e\beta^{l+1}
\end{aligned} \tag{60}$$

From Tropp (2012), for $U_i \in \mathbb{R}^{h \times h}$ and $U_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$, we have,

$$\mathbb{P}\left(\|U_i\|_2 \geq t\right) \leq 2he^{-t^2/2h\sigma^2}. \tag{61}$$

Taking a union bound, we have

$$\begin{aligned}
\mathbb{P}\left(\|U_1\|_2 < t \;\&\; \cdots \;\&\; \|U_l\|_2 < t\right) &= 1 - \mathbb{P}\left(\exists i, \|U_i\|_2 \geq t\right) \\
&\geq 1 - \sum_{i=1}^{l} \mathbb{P}\left(\|U_i\|_2 \geq t\right) \\
&\geq 1 - 2lhe^{-t^2/2h\sigma^2}.
\end{aligned} \tag{62}$$

Setting $2lhe^{-t^2/2h\sigma^2} = \frac{1}{2}$, we have $t = \sigma\sqrt{2h\log(4lh)}$. This implies that the probability that the spectral norm of the perturbation of any layer is no larger than $\sigma\sqrt{2h\log(4lh)}$ holds with probability

at least $\frac{1}{2}$. Plugging this bound into Lemma 3.3, we have with probability at least $\frac{1}{2}$,

$$\begin{aligned}|f_{w+u}(X,A) - f_w(X,A)|_2 &\leq e\frac{t}{\zeta} lC_\phi B\|W_1\|_2\|W_l\|_2\frac{(d\mathcal{C})^{l-1}-1}{d\mathcal{C}-1} \\ &= etlB\frac{\lambda\xi}{\zeta} \\ &= eBl\beta^{l+1}t \leq e^2Bl\tilde{\beta}^{l+1}\sigma\sqrt{2h\log(4lh)} \leq \frac{\gamma}{4}, \end{aligned} \tag{63}$$

where we can set $\sigma = \frac{\gamma}{42Bl\tilde{\beta}^{l+1}\sqrt{h\log(4lh)}}$ to get the last inequality. Note that Lemma 3.3 also requires $\max\left(\frac{\|U_1\|_2}{\|W_1\|_2}, \frac{\|U_2\|_2}{\|W_2\|_2}, \frac{\|U_l\|_2}{\|W_l\|_2}\right) \leq \frac{1}{l}$. The requirement is satisfied if $\sigma \leq \frac{\zeta}{l\sqrt{2h\log(4lh)}}$ which in turn can be satisfied if

$$\frac{\gamma}{4eBl\beta^{l+1}\sqrt{2h\log(4lh)}} \leq \frac{1}{\beta l\sqrt{2h\log(4lh)}}, \tag{64}$$

since the chosen value of σ satisfies $\sigma \leq \frac{\gamma}{4eBl\beta^{l+1}\sqrt{2h\log(4lh)}}$ and $\frac{1}{\beta} \leq \zeta$. Therefore, one sufficient condition to make Eq. (64) hold is $\frac{\gamma}{4eB} \leq \beta^l$. We will see how to satisfy this condition later.

We now compute the KL term in the PAC-Bayes bound in Lemma 2.2.

$$\begin{aligned}\mathrm{KL}\left(Q\|P\right) &= \frac{|w|_2^2}{2\sigma^2} \\ &= \frac{42^2B^2\tilde{\beta}^{2l+2}l^2h\log(4lh)}{2\gamma^2}\left(\|W_1\|_F^2 + \|W_2\|_F^2 + \|W_l\|_F^2\right) \\ &\leq \mathcal{O}\left(\frac{B^2\beta^{2l+2}l^2h\log(lh)}{\gamma^2}\left(\|W_1\|_F^2 + \|W_2\|_F^2 + \|W_l\|_F^2\right)\right) \end{aligned} \tag{65}$$

From Lemma 2.2, fixing any $\tilde{\beta}$, with probability $1-\delta$ and for all w such that $|\beta - \tilde{\beta}| \leq \frac{1}{l+1}\beta$, we have,

$$L_{\mathcal{D},0}(f_w) \leq L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2\beta^{2l+2}l^2h\log(lh)|w|_2^2 + \log\frac{m}{\delta}}{\gamma^2 m}}\right). \tag{66}$$

Finally, we need to consider multiple choices of $\tilde{\beta}$ so that for any β, we can bound the generalization error like Eq. (66). In particular, we only need to consider values of β in the following range,

$$\left(\frac{\gamma}{2B}\right)^{\frac{1}{l}} \leq \beta \leq \left(\frac{\gamma\sqrt{m}}{2B}\right)^{\frac{1}{l}}, \tag{67}$$

since otherwise the bound holds trivially as $L_{\mathcal{D},0}(f_w) \leq 1$ by definition. To see this, if $\beta^l < \frac{\gamma}{2B}$, then for any (X,A) and any $j \in \mathbb{N}_K^+$, we have,

$$\begin{aligned}|f_w(X,A)[j]| \leq |f_w(X,A)|_2 &= |\frac{1}{n}\mathbf{1}_n H_{l-1}W_l|_2 \\ &\leq \frac{1}{n}|\mathbf{1}_n H_{l-1}|_2\|W_l\|_2 \\ &\leq \|W_l\|_2 \max_i |H_{l-1}[i,:]|_2 \\ &\leq BC_\phi\|W_1\|_2\|W_l\|_2\frac{(d\mathcal{C})^{l-1}-1}{d\mathcal{C}-1} \qquad \text{(Use Eq. (48))} \\ &\leq B\lambda\xi \qquad \text{(Use definition of } \lambda \text{ and } \xi) \\ &\leq B\beta^l \qquad \text{(Use definition of } \beta) \\ &< \frac{\gamma}{2}. \end{aligned} \tag{68}$$

Therefore, based on the definition in Eq. (4), we always have $L_{S,\gamma}(f_w) = 1$ when $\beta^l < \frac{\gamma}{2B}$. It is hence sufficient to consider $\beta^l \geq \frac{\gamma}{2B} > \frac{\gamma}{4eB}$ which also makes Eq. (64) hold. Alternatively, if $\beta^l > \frac{\gamma\sqrt{m}}{2B}$, the term inside the big-O notation in Eq. (66) would be,

$$\sqrt{\frac{B^2\beta^{2l}l^2h\log(lh)(\beta^2|w|_2^2) + \log\frac{m}{\delta}}{\gamma^2 m}} \geq \sqrt{\frac{l^2h\log(lh)(|w|_2^2/\zeta^2)}{4}} \geq 1, \tag{69}$$

The last inequality uses the fact that we typically choose $h \geq 2$ in practice, $l \geq 2$ and $|w|_2^2 \geq \min\left(\|W_1\|_F^2, \|W_2\|_F^2, \|W_l\|_F^2\right) \geq \zeta^2$. Since we only need to consider β in the range of Eq. (67), one sufficient condition to ensure $|\beta - \tilde{\beta}| \leq \frac{1}{l+1}\beta$ holds would be $|\beta - \tilde{\beta}| \leq \frac{1}{l+1}\left(\frac{\gamma}{2B}\right)^{\frac{1}{l}}$. Therefore, if we can find a covering of the interval in Eq. (67) with radius $\frac{1}{l+1}\left(\frac{\gamma}{2B}\right)^{\frac{1}{l}}$ and make sure bounds like Eq. (66) holds while $\tilde{\beta}$ takes all possible values from the covering, then we can get a bound which holds for all β. It is clear that we only need to consider a covering C with size $|C| = \frac{(l+1)}{2}\left(m^{1/2l} - 1\right)$. Therefore, denoting the event of Eq. (66) with $\tilde{\beta}$ taking the i-th value of the covering as E_i, we have

$$\mathbb{P}\left(E_1 \,\&\, \cdots \,\&\, E_{|C|}\right) = 1 - \mathbb{P}\left(\exists i, \bar{E}_i\right) \geq 1 - \sum_{i=1}^{|C|}\mathbb{P}\left(\bar{E}_i\right) \geq 1 - |C|\delta, \tag{70}$$

where $\bar{E}_i$ denotes the complement of E_i. Hence, with probability $1-\delta$ and for all w, we have,

$$\begin{aligned}
L_{\mathcal{D},0}(f_w) &\leq L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2\beta^{2l+2}l^2h\log(lh)|w|_2^2 + \log\frac{m|C|}{\delta}}{\gamma^2 m}}\right) \\
&= L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2\beta^{2l+2}l^2h\log(lh)|w|_2^2 + \log\frac{m(l+1)}{\delta} + \frac{1}{2l}\log m}{\gamma^2 m}}\right) \\
&= L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2\max\left(\zeta^{-1}, (\lambda\xi)^{\frac{1}{l}}\right)^{2(l+1)}l^2h\log(lh)|w|_2^2 + \log\frac{m(l+1)}{\delta}}{\gamma^2 m}}\right)
\end{aligned} \tag{71}$$

which proves the theorem for the case of $d\mathcal{C} \neq 1$.

Special Case $d\mathcal{C} = 1$ Now we consider $d\mathcal{C} = 1$ of which the proof follows the logic of the one for $d\mathcal{C} \neq 1$. Note that this case happens rarely in practice. We only include it for the completeness of the analysis. We again construct a statistic $\beta = \max\left(\frac{1}{\zeta}, \sqrt{\lambda C_\phi}\right)$. For now, let us fix any $\tilde{\beta}$ and consider β which satisfies $|\beta - \tilde{\beta}| \leq \frac{1}{3}\beta$. This also implies $\frac{1}{e}\beta^3 \leq \tilde{\beta}^3 \leq e\beta^3$. Based on Lemma 3.3, we have,

$$\begin{aligned}
|f_{w+u}(X,A) - f_w(X,A)|_2 &\leq e\frac{t}{\zeta}(l+1)^2 C_\phi B\|W_1\|_2\|W_l\|_2 \\
&= et(l+1)^2 B\frac{\lambda C_\phi}{\zeta} \leq eB(l+1)^2\beta^3 t \\
&\leq e^2 B(l+1)^2\tilde{\beta}^3\sigma\sqrt{2h\log(4lh)} \leq \frac{\gamma}{4},
\end{aligned} \tag{72}$$

where we can set $\sigma = \frac{\gamma}{42B(l+1)^2\tilde{\beta}^3\sqrt{h\log(4lh)}}$ to get the last inequality. Note that Lemma 3.3 also requires $\max\left(\frac{\|U_1\|_2}{\|W_1\|_2}, \frac{\|U_2\|_2}{\|W_2\|_2}, \frac{\|U_l\|_2}{\|W_l\|_2}\right) \leq \frac{1}{l}$. The requirement is satisfied if $\sigma \leq \frac{\zeta}{l\sqrt{2h\log(4lh)}}$ which in turn can be satisfied if

$$\frac{\gamma}{4eB(l+1)^2\beta^3\sqrt{2h\log(4lh)}} \leq \frac{1}{\beta l\sqrt{2h\log(4lh)}}, \tag{73}$$

since the chosen value of σ satisfies $\sigma \leq \frac{\gamma}{4eB(l+1)^2\beta^3\sqrt{2h\log(4lh)}}$ and $\frac{1}{\beta} \leq \zeta$. As shown later, we only need to consider a certain range of values of β which naturally satisfy the condition $\frac{\gamma l}{4eB(l+1)^2} \leq \beta^2$, *i.e.*, the equivalent form of Eq. (73). This assures the applicability of Lemma 3.3. Now we compute the KL divergence,

$$\begin{aligned}
\mathrm{KL}\left(Q\|P\right) &= \frac{|w|_2^2}{2\sigma^2} \\
&= \frac{42^2B^2\tilde{\beta}^6(l+1)^4h\log(4lh)}{2\gamma^2}\left(\|W_1\|_F^2+\|W_2\|_F^2+\|W_l\|_F^2\right) \\
&\leq \mathcal{O}\left(\frac{B^2\beta^6(l+1)^4h\log(lh)}{\gamma^2}\left(\|W_1\|_F^2+\|W_2\|_F^2+\|W_l\|_F^2\right)\right)
\end{aligned} \tag{74}$$

In particular, we only need to consider values of β in the following range,

$$\sqrt{\frac{\gamma}{2Bl}} \leq \beta \leq \sqrt{\frac{\gamma\sqrt{m}}{2Bl}}, \tag{75}$$

since otherwise the bound holds trivially as $L_{\mathcal{D},0}(f_w) \leq 1$ by definition. To see this, if $\beta < \frac{\gamma}{2Bl}$, then for any (X, A) and any $j \in \mathbb{N}_K^+$, we have,

$$\begin{aligned}
|f_w(X,A)[j]| \leq |f_w(X,A)|_2 &= |\frac{1}{n}\mathbf{1}_n H_{l-1}W_l|_2 \\
&\leq \frac{1}{n}|\mathbf{1}_n H_{l-1}|_2\|W_l\|_2 \\
&\leq \|W_l\|_2 \max_i |H_{l-1}[i,:]|_2 \\
&\leq B(l-1)C_\phi\|W_1\|_2\|W_l\|_2 \qquad \text{(Use Eq. (48))} \\
&\leq B(l-1)\lambda C_\phi \qquad \text{(Use definition of } \lambda\text{)} \\
&\leq Bl\beta^2 \qquad \text{(Use definition of } \beta\text{)} \\
&< \frac{\gamma}{2}.
\end{aligned} \tag{76}$$

Therefore, based on the definition in Eq. (4), we always have $L_{S,\gamma}(f_w) = 1$ when $\beta < \frac{\gamma}{2Bl}$. It it hence sufficient to consider $\beta^2 \geq \frac{\gamma}{2Bl} \geq \frac{\gamma l}{4eBl^2} \geq \frac{\gamma l}{4eB(l+1)^2}$ which means the condition in Eq. (73) is indeed satisfied. Alternatively, if $\beta > \sqrt{\frac{\gamma\sqrt{m}}{2Bl}}$, the term inside the big-O notation in Eq. (74) would be,

$$\sqrt{\frac{B^2\beta^4(l+1)^4h\log(lh)\beta^2|w|_2^2+\log\frac{m}{\delta}}{\gamma^2 m}} \geq \sqrt{\frac{(l+1)^4h\log(lh)\frac{|w|_2^2}{\zeta^2}}{4l^2}} \geq 1, \tag{77}$$

where the first inequality hold since $\beta \geq \frac{1}{\zeta}$. The last inequality uses the fact that we typically choose $h \geq 2$ in practice, $l \geq 2$, and $|w|_2^2 \geq \min\left(\|W_1\|_F^2, \|W_2\|_F^2, \|W_l\|_F^2\right) \geq \zeta^2$. Since we only need to consider β in the range of Eq. (75), one sufficient condition to ensure $|\beta - \tilde{\beta}| \leq \frac{1}{3}\beta$ always holds would be $|\beta - \tilde{\beta}| \leq \frac{1}{3}\sqrt{\frac{\gamma}{2Bl}}$. Therefore, if we can find a covering of the interval in Eq. (75) with radius $\frac{1}{3}\sqrt{\frac{\gamma}{2Bl}}$ and make sure bounds like Eq. (66) holds while $\tilde{\beta}$ takes all possible values from the covering, then we can get a bound which holds for all β. It is clear that we only need to consider a covering C with size $|C| = \frac{3}{2}\left(m^{\frac{1}{4}} - 1\right)$.

Statistics	Max Node Degree $d-1$	Max Hidden Dim h	Spectral Norm of Learned Weights
VC-Dimension (Scarselli et al., 2018)	-	$\mathcal{O}\left(h^4\right)$	-
Rademacher Complexity (Garg et al., 2020) Case A	$\mathcal{O}\left(d^{l-1}\sqrt{\log(d^{l-1})}\right)$	$\mathcal{O}\left(h\right)$	$\mathcal{O}\left(\lambda\mathcal{C}\xi\sqrt{\log\left(\lambda\mathcal{C}\xi\right)}\right)$
Rademacher Complexity (Garg et al., 2020) Case B	$\mathcal{O}\left(d^{l-1}\sqrt{\log(d^{l-2})}\right)$	$\mathcal{O}\left(h\sqrt{\log\sqrt{h}}\right)$	$\mathcal{O}\left(\lambda\mathcal{C}\xi\sqrt{\log\left(\lambda\xi\right)}\right)$
Rademacher Complexity (Garg et al., 2020) Case C	$\mathcal{O}\left(d^{l-1}\sqrt{\log(d^{2l-3})}\right)$	$\mathcal{O}\left(h\sqrt{\log\sqrt{h}}\right)$	$\mathcal{O}\left(\lambda\mathcal{C}\xi\sqrt{\log\left(\|W_2\|_2\lambda\xi^2\right)}\right)$
Ours Case A	-	$\mathcal{O}\left(\sqrt{h\log h}\right)$	$\mathcal{O}\left(\zeta^{-(l+1)}\sqrt{\|W_1\|_F^2+\|W_2\|_F^2+\|W_l\|_F^2}\right)$
Ours Case B	$\mathcal{O}\left(d^{\frac{(l+1)(l-2)}{l}}\right)$	$\mathcal{O}\left(\sqrt{h\log h}\right)$	$\mathcal{O}\left(\lambda^{1+\frac{1}{l}}\xi^{1+\frac{1}{l}}\sqrt{\|W_1\|_F^2+\|W_2\|_F^2+\|W_l\|_F^2}\right)$

Table 3: Detailed comparison of Generalization Bounds for GNNs. "-" means inapplicable. We only consider the general case $d\mathcal{C}\|W_2\|_2 \neq 1$ and simplify the Rademacher complexity based bounds (w.r.t. spectral norm of weights) based on the assumption that $C_\phi \ll d\mathcal{C}\xi$ which generally holds in practice. Here $\mathcal{C} = C_\phi C_\rho C_g \|W_2\|_2$, $\xi = C_\phi \frac{(d\mathcal{C})^{l-1}-1}{d\mathcal{C}-1}$, $\zeta = \min\left(\|W_1\|_2, \|W_2\|_2, \|W_l\|_2\right)$, and $\lambda = \|W_1\|_2\|W_l\|_2$. Note that $d^{\frac{(l+1)(l-2)}{l}} \leq d^{\frac{l^2-l}{l}} = d^{l-1}$.

Hence, with probability $1-\delta$ and for all w, we have,

$$
\begin{aligned}
L_{\mathcal{D},0}(f_w) &\leq L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2\beta^6(l+1)^4 h\log(lh)|w|_2^2 + \log\frac{m|C|}{\delta}}{\gamma^2 m}}\right) \\
&= L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2\beta^6(l+1)^4 h\log(lh)|w|_2^2 + \log\frac{m}{\delta} + \frac{1}{4}\log m}{\gamma^2 m}}\right) \\
&= L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2\max\left(\zeta^{-6}, \lambda^3 C_\phi^3\right)(l+1)^4 h\log(lh)|w|_2^2 + \log\frac{m}{\delta}}{\gamma^2 m}}\right)
\end{aligned} \tag{78}
$$

which proves the theorem for the case of $d\mathcal{C} = 1$.

□

Remark. Note that our proof applies to both homogeneous and non-homogeneous GNNs.

A.5 Bound Comparison

In this section, we explain the details of the comparison with Rademacher complexity based generalization bounds of GNNs.

A.5.1 Rademacher Complexity based Bound

We first restate the Rademacher complexity bound from (Garg et al., 2020) as below:

$$
\begin{aligned}
&L_{D,0}(f_w) \leq L_{S,\gamma}(f_w) \\
&\quad + \mathcal{O}\left(\frac{1}{\gamma m} + hB_l Z\sqrt{\frac{\log\left(B_l\sqrt{m}\max\left(Z, M\sqrt{h}\max\left(BB_1, \bar{R}B_2\right)\right)\right)}{\gamma^2 m}} + \sqrt{\frac{\log\frac{1}{\delta}}{m}}\right)
\end{aligned} \tag{79}
$$

where $M = C_\phi\frac{(C_\phi C_\rho C_g d B_2)^{l-1}-1}{C_\phi C_\rho C_g d B_2 - 1}$, $Z = C_\phi\left(BB_1 + \bar{R}B_2\right)$, $\bar{R} \leq C_\rho C_g d\min\left(b\sqrt{h}, BB_1 M\right)$, b is the uniform upper bound of ϕ (*i.e.*, $\forall x \in \mathbb{R}^h$, $\phi(x) \leq b$), and B_1, B_2, B_l are the spectral norms

of the weight matrices W_1, W_2, W_l. Note that the numerator of M has the exponent $l-1$ since we count the readout function in the number of layers/steps, *i.e.*, there are $l-1$ message passing steps in total.

A.5.2 Comparison in Our Context

For typical message passing GNNs presented in the literature, node state update function ϕ could be a neural network like MLP or GRU, a ReLU unit, etc. This makes the assumption of the uniform upper bound on ϕ impractical, *e.g.*, $b = \infty$ when ϕ is ReLU. Therefore, we dot not adopt this assumption in our analysis[8].

Rademacher Complexity Based Bound Based on the above consideration, we have $\bar{R} \leq C_\rho C_g d B B_1 M$. We further convert some notations in the original bound to the ones in our context.

$$M = C_\phi \frac{(C_\phi C_\rho C_g d B_2)^{l-1} - 1}{C_\phi C_\rho C_g d B_2 - 1} = \xi \tag{80}$$

$$\bar{R} \leq C_\rho C_g d B B_1 M = C_\rho C_g d B \|W_1\|_2 \xi \tag{81}$$

$$Z = C_\phi \left(B B_1 + \bar{R} B_2 \right) = B \|W_1\|_2 \left(C_\phi + d\mathcal{C}\xi \right), \tag{82}$$

where we use the same abbreviations as in Theorem 3.4, $\xi = C_\phi \frac{(d\mathcal{C})^{l-1}-1}{d\mathcal{C}-1}$, $\lambda = \|W_1\|_2 \|W_l\|_2$, $\mathcal{C} = C_\phi C_\rho C_g \|W_2\|_2$.

We need to consider three cases for the big-O term of the original bound in Eq. (79) depending on the outcomes of the two point-wise maximum functions.

Case A If $\max\left(Z, M\sqrt{h}\max\left(BB_1, \bar{R}B_2\right)\right) = Z$, then the generalization bound is,

$$\begin{aligned}
&\mathcal{O}\left(h B_l Z \sqrt{\frac{\log\left(B_l \sqrt{m} Z\right)}{m}} \right) \\
&= \mathcal{O}\left(h \|W_l\| B \|W_1\|_2 \left(C_\phi + d\mathcal{C}\xi\right) \sqrt{\frac{\log\left(\|W_l\|_2 \sqrt{m} B \|W_1\|_2 \left(C_\phi + d\mathcal{C}\xi\right)\right)}{m}} \right) \\
&= \mathcal{O}\left(h B \lambda \left(C_\phi + d\mathcal{C}\xi\right) \sqrt{\frac{\log\left(\sqrt{m} B \lambda \left(C_\phi + d\mathcal{C}\xi\right)\right)}{m}} \right).
\end{aligned} \tag{83}$$

Case B If $\max\left(Z, M\sqrt{h}\max\left(BB_1, \bar{R}B_2\right)\right) = M\sqrt{h}\max\left(BB_1, \bar{R}B_2\right)$ and $BB_1 = \max\left(BB_1, \bar{R}B_2\right)$, then the generalization bound is,

$$\begin{aligned}
&\mathcal{O}\left(h B_l Z \sqrt{\frac{\log\left(B_l \sqrt{m} M \sqrt{h} B B_1\right)}{m}} \right) \\
&= \mathcal{O}\left(h \|W_l\| B \|W_1\|_2 \left(C_\phi + d\mathcal{C}\xi\right) \sqrt{\frac{\log\left(\|W_l\|_2 \sqrt{m} \xi \sqrt{h} B \|W_1\|_2\right)}{m}} \right) \\
&= \mathcal{O}\left(h B \lambda \left(C_\phi + d\mathcal{C}\xi\right) \sqrt{\frac{\log\left(\sqrt{m} \lambda \xi \sqrt{h} B\right)}{m}} \right)
\end{aligned} \tag{84}$$

[8] If we introduce the uniform upper bound on ϕ in our analysis, we can also obtain a similar functional dependency in our bound like $\min(b\sqrt{h}, \cdot)$. But as aforementioned, it is somewhat impractical and leads to a more cumbersome bound.

Case C If $\max\left(Z, M\sqrt{h}\max\left(BB_1, \bar{R}B_2\right)\right) = M\sqrt{h}\max\left(BB_1, \bar{R}B_2\right)$ and $\bar{R}B_2 = \max\left(BB_1, \bar{R}B_2\right)$, then the generalization bound is,

$$\begin{aligned}
&\mathcal{O}\left(hB_lZ\sqrt{\frac{\log\left(B_l\sqrt{m}M\sqrt{h}\bar{R}B_2\right)}{m}}\right) \\
&= \mathcal{O}\left(h\|W_l\|B\|W_1\|_2\left(C_\phi + d\mathcal{C}\xi\right)\sqrt{\frac{\log\left(\|W_l\|_2\sqrt{m}\xi\sqrt{h}C_\rho C_g dB\|W_1\|_2\xi\|W_2\|_2\right)}{m}}\right) \\
&= \mathcal{O}\left(hB\lambda\left(C_\phi + d\mathcal{C}\xi\right)\sqrt{\frac{\log\left(\lambda\sqrt{m}\sqrt{h}C_\rho C_g dB\xi^2\|W_2\|_2\right)}{m}}\right)
\end{aligned} \tag{85}$$

We show the detailed dependencies of the Rademarcher complexity based bound under three cases in Table 3. In practice, we found message passing GNNs typically do not behave like a contraction mapping. In other words, we have $d\mathcal{C} > 1$ and $\xi \gg 1$ hold for many datasets. Therefore, the case C happens more often in practice, *i.e.*, $\max\left(Z, M\sqrt{h}\max\left(BB_1, \bar{R}B_2\right)\right) = M\sqrt{h}\bar{R}B_2$.

PAC Bayes Bound For our PAC-Bayes bound in Theorem 3.4, we also need to consider two cases which correspond to $\max\left(\zeta^{-1}, (\lambda\xi)^{\frac{1}{l}}\right) = \zeta^{-1}$ (case A) and $\max\left(\zeta^{-1}, (\lambda\xi)^{\frac{1}{l}}\right) = (\lambda\xi)^{\frac{1}{l}}$ (case B) respectively. Here $\zeta = \min\left(\|W_1\|_2, \|W_2\|_2, \|W_l\|_2\right)$. We show the detailed dependencies of our bound under three cases in Table 3. Again, in practice, we found $d\mathcal{C} > 1$, $\xi \gg 1$ and $\zeta \leq 1$. Therefore, case B occurs more often.

VC-dim Bound (Scarselli et al., 2018) show that the upper bound of the VC-dimension of general GNNs with Sigmoid or Tanh activations is $\mathcal{O}(p^4N^2)$ where p is the total number of parameters and N is the maximum number of nodes. Since $p = \mathcal{O}(h^2)$ in our case, the VC-dim bound is $\mathcal{O}(h^8N^2)$. Therefore, the corresponding generalization bound scales as $\mathcal{O}(\frac{h^4N}{\sqrt{m}})$. Note that N is at least d and could be much larger than d for some datasets.

A.6 Connections with Existing Bounds of MLPs/CNNs

ReLU Networks are Special GCNs Since regular feedforward neural networks could be viewed as a special case of GNNs by treating each sample as the node feature of a single-node graph, it is natural to investigate the connections between these two classes of models. In particular, we consider the class of ReLU networks studied in Neyshabur et al. (2017),

$$\begin{aligned}
H_0 &= X && \text{(Input Node Feature)} \\
H_k &= \sigma_k\left(H_{k-1}W_k\right) && (k\text{-th Layer}) \\
H_l &= H_{l-1}W_l && \text{(Readout Layer)},
\end{aligned} \tag{86}$$

where σ_k = ReLU. It includes two commonly-seen types of deep neural networks, *i.e.*, fully connected networks (or MLPs) and convolutional neural networks (CNNs), as special cases. Comparing Eq. (86) against Eq. (1), it is clear that these ReLU networks can be further viewed as special cases of GCNs which operate on single-node graphs, *i.e.*, $\tilde{L} = I$.

Statistics	COLLAB	IMDB-BINARY	IMDB-MULTI	PROTEINS
max # nodes	492	136	89	620
max # edges	80727	2634	3023	2718
# classes	3	2	3	2
# graphs	5000	1000	1500	1113
train/test	4500/500	900/100	1350/150	1002/111
feature dimension	367	65	59	3
max node degree	491	135	88	25

Table 4: Statistics of real-world datasets.

Statistics	ER-1	ER-2	ER-3	ER-4	SBM-1	SBM-2
max # nodes	100	100	100	100	100	100
max # edges	1228	3266	5272	7172	2562	1870
# classes	2	2	2	2	2	2
# graphs	200	200	200	200	200	200
train/test	180/20	180/20	180/20	180/20	180/20	180/20
feature dimension	16	16	16	16	16	16
max node degree	25	48	69	87	25	36

Table 5: Statistics of synthetic datasets.

Connections of Generalization Bounds Let us restate the PAC-Bayes bound of ReLU networks in Neyshabur et al. (2017) as below,

$$L_{D,0}(f_w) \leq L_{S,\gamma}(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2 l^2 h \log(lh) \prod_{i=1}^{l} \|W_i\|_2^2 \sum_{i=1}^{l} \frac{\|W_i\|_F^2}{\|W_i\|_2^2} + \log \frac{ml}{\delta}}{\gamma^2 m}}\right). \tag{87}$$

Comparing it with the bound in Theorem 3.2, we can find that our bound only adds a factor d^{l-1} to the first term inside the square root of the big-O notation which is brought by the underlying graph structure of the data. If we consider GCNs operating on single-node graphs, *i.e.*, the case where GCNs degenerate to ReLU networks, two bounds coincide since $d = 1$. Therefore, our Theorem 3.2 directly generalizes the result in Neyshabur et al. (2017) to GCNs which is a strictly larger class of models than ReLU networks.

A.7 Experimental Details

Datasets We create 6 synthetic datasets by generating random graphs from different random graph models. In particular, the first 4 synthetic datasets correspond to the Erdős–Rényi models with different edge probabilities: 1) Erdős–Rényi-1 (ER-1), edge probability = 0.1; 2) Erdős–Rényi-2 (ER-2), edge probability = 0.3; 3) Erdős–Rényi-3 (ER-3), edge probability = 0.5; 4) Erdős–Rényi-4 (ER-4), edge probability = 0.7. The remaining 2 synthetic datasets correspond to the stochastic block model with the following settings: 1) Stochastic-Block-Model-1 (SBM-1), two blocks, sizes = [40, 60], edge probability = [[0.25, 0.13], [0.13, 0.37]]; 2) Stochastic-Block-Model-2 (SBM-2), three blocks, sizes = [25, 25, 50], edge probability = [[0.25, 0.05, 0.02], [0.05, 0.35, 0.07], [0.02, 0.07, 0.40]]. Each synthetic dataset has 200 graphs where the number of nodes of individual graph is 100, the number of classes is 2, and the random train-test split ratio is $90\%/10\%$. For each random graph of individual synthetic dataset, we generate the 16-dimension random Gaussian node feature (normalized to have unit ℓ_2 norm) and a binary class label following a uniform distribution. We summarize the statistics of the real-world and synthetic datasets in Table 4 and Table 5 respectively.

$l=2$	PROTEINS	IMDB-MULTI	IMDB-BINARY	COLLAB
Rademacher	11.80 ± 0.18	16.66 ± 0.04	17.37 ± 0.02	21.26 ± 0.07
PAC-Bayes	$\mathbf{8.45} \pm 0.28$	$\mathbf{15.26} \pm 0.07$	$\mathbf{15.44} \pm 0.03$	$\mathbf{19.37} \pm 0.17$
$l=4$				
Rademacher	24.04 ± 0.23	29.94 ± 0.10	31.38 ± 0.09	41.03 ± 0.33
PAC-Bayes	$\mathbf{22.10} \pm 0.23$	$\mathbf{28.35} \pm 0.11$	$\mathbf{29.53} \pm 0.08$	$\mathbf{40.31} \pm 0.36$

Table 6: Bound (log value) comparisons on real-world datasets.

$l=2$	ER-1	ER-2	ER-3	ER-4	SBM-1	SBM-2
Rademacher	17.37 ± 0.16	17.98 ± 0.13	18.15 ± 0.15	18.35 ± 0.10	17.88 ± 0.11	17.71 ± 0.09
PAC-Bayes	$\mathbf{15.38} \pm 0.12$	$\mathbf{15.13} \pm 0.13$	$\mathbf{14.86} \pm 0.25$	$\mathbf{14.69} \pm 0.24$	$\mathbf{15.23} \pm 0.12$	$\mathbf{15.35} \pm 0.10$
$l=4$						
Rademacher	27.92 ± 0.02	29.57 ± 0.12	30.64 ± 0.18	31.34 ± 0.20	29.35 ± 0.14	28.87 ± 0.07
PAC-Bayes	$\mathbf{27.00} \pm 0.04$	$\mathbf{28.32} \pm 0.07$	$\mathbf{29.18} \pm 0.12$	$\mathbf{29.70} \pm 0.14$	$\mathbf{28.14} \pm 0.05$	$\mathbf{27.74} \pm 0.04$
$l=6$						
Rademacher	37.10 ± 0.29	40.22 ± 0.19	42.00 ± 0.26	43.08 ± 0.39	40.04 ± 0.25	39.02 ± 0.19
PAC-Bayes	$\mathbf{36.85} \pm 0.25$	$\mathbf{39.65} \pm 0.14$	$\mathbf{41.30} \pm 0.22$	$\mathbf{42.24} \pm 0.34$	$\mathbf{39.50} \pm 0.17$	$\mathbf{38.63} \pm 0.17$
$l=8$						
Rademacher	$\mathbf{46.72} \pm 0.51$	51.16 ± 0.21	53.44 ± 0.39	55.06 ± 0.38	50.60 ± 0.17	49.29 ± 0.34
PAC-Bayes	46.79 ± 0.48	$\mathbf{51.02} \pm 0.21$	$\mathbf{53.10} \pm 0.36$	$\mathbf{54.67} \pm 0.38$	$\mathbf{50.44} \pm 0.16$	$\mathbf{49.22} \pm 0.36$

Table 7: Bound (log value) comparisons on synthetic datasets.

Experimental Setup For all MPGNNs used in the experiments, we specify $\phi = \text{ReLU}$, $\rho = \text{Tanh}$, and $g = \text{Tanh}$ which imply $C_\phi = C_\rho = C_g = 1$. For experiments on real-world datasets, we set $h = 128$, the number of training epochs to 50, and try 2 values of network depth, *i.e.*, $l = 2$ and $l = 4$. The batch size is set to 20 (due to the GPU memory constraint) on COLLAB and 128 for others. For experiments on synthetic datasets, we set $h = 128$ and try 4 values of network depth, *i.e.*, $l = 2$, $l = 4$, $l = 6$ and $l = 8$. Since these generated datasets essentially require GNNs to fit to random labels which is arguably hard, we extend the number of training epochs to 200. For all above experiments, we use Adam as the optimizer with learning rate set to $1.0e^{-2}$. The batch size is 128 for all synthetic datasets.

Bound Computations For all datasets, we compute the bound values for the learned model saved in the end of the training. We also consider the constants of both bounds in the computation. In particular, for our bound, we compute the following quantity

$$\sqrt{\frac{42^2 B^2 \left(\max\left(\zeta^{-(l+1)}, (\lambda\xi)^{\frac{l+1}{l}}\right)\right)^2 l^2 h \log(4lh)|w|_2^2}{\gamma^2 m}}. \tag{88}$$

For the Rademacher complexity based bound, we compute the following quantity

$$2 \times 24 h B_l Z \sqrt{\frac{3\log\left(24 B_l \sqrt{m} \max\left(Z, M\sqrt{h}\max\left(BB_1, \bar{R}B_2\right)\right)\right)}{\gamma^2 m}}, \tag{89}$$

where the variables are the same as Eq. (79).

Experimental Results In addition to the figures shown in the main paper, we also provide the numerical values of the bound evaluations in Table 6 (real-world datasets) and Table 7 (synthetic

datasets). As you can see, our bound is tighter than the Rademacher complexity based one under all settings except for one synthetic setting which falls in the scenario "small d (max-node-degree) and large l (number-of-steps)". This makes sense since we have a square term on the number of steps l and it will play a role when the term involved with d is comparable (*i.e.*, when d is small). Again, all quantities are in the log domain.