PAC-Bayesian Approach to Generalization Bounds for Graph Neural Networks

Athindran Ramesh Kumar

Princeton University
arkumar@princeton.edu

September 22, 2024

Traditional generalization theory

- Targets y, inputs x from distribution $\mathcal{D}$, predictions $\hat{y}(x)$
- Loss function $\mathcal{L}(y, \hat{y}(x))$
- Typical generalization bound: w.p $1 - \delta$

$$\mathbb{E}_{x \sim \mathcal{D}}[\mathcal{L}(y, \hat{y}(x))] \leq \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(y, \hat{y}(x_i)) + \triangle(m, \delta, \chi) \tag{1}$$

$$\triangle \propto \sqrt{\frac{1}{m}} \text{ (Usually)}$$
$$\triangle \propto \chi \text{ (Some measure of model complexity)}$$

Popular complexity measures: VC-dimension, Rademacher complexity.
Another approach to generalization called PAC-Bayes

Why are these bounds vacuous?

Bias-Variance Trade-off of Machine Learning

- Applicable to regression with squared loss

$$\mathbb{E}_{\mathcal{D}}[(y-\hat{y}(x))^2] = \underbrace{\mathbb{E}_{\mathcal{D}}[(y-\theta^{*T}x)^2]}_{\text{Bias}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\hat{\theta}^T x-\theta^{*T}x)^2]}_{\text{Variance}} + \text{noise} \tag{2}$$

- Gaussian distribution over the input features - possible to characterize the variance of least norm solution.
- Variance in the bound - variance of the least norm solution.
- Bias can be minimized to global minimum.

$$\mathbb{E}_{x\sim\mathcal{D}}[\mathcal{L}(y,\hat{y}(x))] \leq \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(y,\hat{y}(x_i)) + \triangle(m,\delta,\chi) \tag{3}$$

Vacuous because **distribution-agnostic** and **algorithm-agnostic**

Generalization for graph networks

- VC-dimension ([4])
- Rademacher ([2])
- Algorithmic stability ([5])
- **PAC-Bayes bounds** ([3])) - in this talk

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)	$\mathcal{O}\left(d^{l-1}\sqrt{\log(d^{2l-3})}\right)$	$\mathcal{O}\left(h\sqrt{\log h}\right)$	$\mathcal{O}\left(\lambda\mathcal{C}\xi\sqrt{\log\left(\\|W_2\\|_2\lambda\xi^2\right)}\right)$
Ours	$\mathcal{O}\left(d^{l-1}\right)$	$\mathcal{O}\left(\sqrt{h\log h}\right)$	$\mathcal{O}\left(\lambda^{1+\frac{1}{l}}\xi^{1+\frac{1}{l}}\sqrt{\\|W_1\\|_F^2+\\|W_2\\|_F^2+\\|W_l\\|_F^2}\right)$

Table 1: Comparison of generalization bounds for GNNs. "-" means inapplicable. l is the network depth. 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 $\lambda = \|W_1\|_2\|W_l\|_2$. More details about the comparison can be found in Appendix A.5.

Figure: PAC-Bayes bounds

Benign over-fitting largely not studied

Problem Setup

- K-class graph classification
- $A \in \mathbb{R}^{n \times n}$ - Adjacency matrix
- $X \in \mathbb{R}^{n \times h_0}$ - Features in each node
- $y \in \mathbb{R}^{1 \times K}$ - Targets at each node
- Node feature of any graph is contained in a L2-ball with radius B
- Maximum hidden dimension across all layers - h

Multi-Class Margin Loss

$$L_{\mathcal{D},\gamma}(f_w) = \mathbb{P}_{z \sim \mathcal{D}} \left(f_w(X, A)[y] \leq \gamma + \max_{j \neq y} f_w(X, A)[j] \right) \tag{4}$$

$$L_{S,\gamma}(f_w) = \frac{1}{m} \sum_{z_i \in S} \mathbb{1} \left(f_w(X, A)[y] \leq \gamma + \max_{j \neq y} f_w(X, A)[j] \right) \tag{5}$$

PAC-Bayes Generalization Theory

Theorem 2.1. *(McAllester, 2003) (Two-sided) 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. size-m training set S according to $\mathcal{D}$, for all distributions Q over $\mathcal{H}$ and any $\gamma > 0$, we have*

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

Lemma 2.2. *(Neyshabur et al., 2017)[4] Let $f_w(x) : \mathcal{X} \to \mathbb{R}^K$ be any model with parameters w, and let 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:*

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

- $\mathcal{H}$ - set of hypothesis classes
- Q, P - Distribution on the parameters w

Bounds for Graph Convolutional Networks

$$H_k = \sigma_k\left(\tilde{L}H_{k-1}W_k\right) \qquad \text{(k-th Graph Convolution Layer)}$$
$$H_l = \frac{1}{n}\mathbf{1}_n H_{l-1}W_l \qquad \text{(Readout Layer)}, \qquad (1)$$

- W_k - Learnable weights, $\bar{L}$ - Laplacian

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,*

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

Final Generalization Bound for GCN's

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

Bounds for Message Passing Graph Networks

$$M_k = g(C_{\text{out}}^\top H_{k-1}) \qquad \text{(k-th step Message Computation)}$$
$$\bar{M}_k = C_{\text{in}} M_k \qquad \text{(k-th step Message Aggregation)}$$
$$H_k = \phi\left(XW_1 + \rho\left(\bar{M}_k\right) W_2\right) \qquad \text{(k-th step Node State Update)}$$
$$H_l = \frac{1}{n}\mathbf{1}_n H_{l-1} W_l \qquad \text{(Readout Layer),}$$

- $C_{in} \in \mathbb{R}^{n\times c}$, $C_{out} \in \mathbb{R}^{n\times c}$, $H_k \in \mathbb{R}^{n\times h}$
- $C_{in}[i,j] = 1$ indicates the incoming node of the j^{th} edge is the i^{th} node
- $C_{out}[i,j] = 1$ indicates the outgoing node of the j^{th} edge is the i^{th} node
- Each of the non-linearities are Lipschitz

Bounds for Message Passing Graph Networks

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,*

$$|f_{w+u}(X, A) - f_w(X, A)|_2 \leq eBl\eta\|W_1\|_2\|W_l\|_2 C_\phi \frac{(d\mathcal{C})^{l-1} - 1}{d\mathcal{C} - 1},$$

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

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,*

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

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 $\xi = C_\phi \frac{(d\mathcal{C})^{l-1} - 1}{d\mathcal{C} - 1}$.

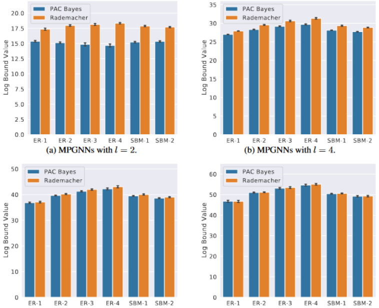
Comparison with Rademacher Bounds

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.



(a) MPGNNs with $l = 2$. (b) MPGNNs with $l = 4$.

(c) MPGNNs with $l = 6$. (d) MPGNNs with $l = 8$.

Figure 2: Bound evaluations on synthetic datasets. The maximum node degrees (*i.e.*, $d - 1$) of datasets from left to right are: 25 (ER-1), 48 (ER-2), 69 (ER-3), 87 (ER-4), 25 (SBM-1), and 36 (SBM-2). 'ER-X' and 'SBM-X' denote the Erdős–Rényi model and the stochastic block model with the 'X'-th setting respectively. Please refer to the appendix for more details.

Comments on PAC-Bayes bounds

- Bounds in this paper still vacuous - any prior P, perturbation applicable for any w
- Hope as the analysis can control P and Q
- Breaking the distribution-agnostic assumption:

$$\text{Optimal Prior} \quad P^* = \mathbb{E}_{\mathcal{A}, S \sim \mathcal{D}}[Q(S)] \tag{6}$$

- With the optimal prior, the KL divergence looks close to some concept of "variance"
- Optimal prior cannot be computed - approximation based on data - performed in [1]

Data-dependent bounds with SGD

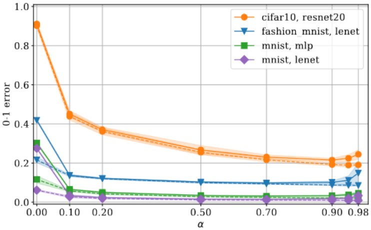


FIGURE 5. **Y-axis**: error-rate; **x-axis**: fraction α of the data used to learn the prior mean; **dashed lines**: test error; **solid lines**: bound on the error of a Gaussian Gibbs classifier whose mean and diagonal covariance are learned by optimizing the bound surrogate; **legend**: dataset and network architecture. For each scenario, under the optimal α, the bound is tight and test error is within a few percent of standard SGD-trained networks.

Looks like there is more promise here. Thank you!

References I

G. K. Dziugaite, K. Hsu, W. Gharbieh, and D. M. Roy.

On the role of data in pac-bayes bounds.

arXiv preprint arXiv:2006.10929, 2020.

V. Garg, S. Jegelka, and T. Jaakkola.

Generalization and representational limits of graph neural networks.

In *International Conference on Machine Learning*, pages 3419–3430. PMLR, 2020.

R. Liao, R. Urtasun, and R. Zemel.

A pac-bayesian approach to generalization bounds for graph neural networks.

arXiv preprint arXiv:2012.07690, 2020.

References II

F. Scarselli, A. C. Tsoi, and M. Hagenbuchner.

The vapnik–chervonenkis dimension of graph and recursive neural networks.

Neural Networks, 108:248–259, 2018.

S. Verma and Z.-L. Zhang.

Stability and generalization of graph convolutional neural networks.

In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1539–1548, 2019.