Towards Generalization Bounds of GCNs for Adversarially Robust Node Classification

Wen Wen [1] Han Li [1,2] Tieliang Gong [3] Hong Chen [1,2]

[1]Huazhong Agricultural University [2]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education [3]Xi'an Jiaotong University

Abstract

Adversarially robust generalization of Graph Convolutional Networks (GCNs) has garnered significant attention in various security-sensitive application areas, driven by intrinsic adversarial vulnerability. Albeit remarkable empirical advancement, theoretical understanding of the generalization behavior of GCNs subjected to adversarial attacks remains elusive. To make progress on the mystery, we establish unified high-probability generalization bounds for GCNs in the context of node classification, by leveraging adversarial Transductive Rademacher Complexity (TRC) and developing a novel contraction technique on graph convolution. Our bounds capture the interaction between generalization error and adversarial perturbations, revealing the importance of key quantities in mitigating the negative effects of perturbations, such as low-dimensional feature projection, perturbation-dependent norm regularization, normalized graph matrix, proper number of network layers, etc. Furthermore, we provide TRC-based bounds of popular GCNs with ℓ_r-norm-additive perturbations for arbitrary $r \geq 1$. A comparison of theoretical results demonstrates that specific network architectures (e.g., residual connection) can help alleviate the cumulative effect of perturbations during the forward propagation of deep GCNs. Experimental results on benchmark datasets validate our theoretical findings.

Preliminaries

- **Attributed graph $\mathcal{G} = (\boldsymbol{A}, \boldsymbol{X})$: $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \in \mathbb{R}^{n\times d}$ is the node feature matrix, and $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is the adjacency matrix.**
- **Adversarial nodes $\widetilde{\boldsymbol{X}}_* = [\widetilde{\boldsymbol{x}}_{1*}, \dots, \widetilde{\boldsymbol{x}}_{n*}]$:** $\widetilde{\boldsymbol{X}}_* = \arg\max_{\widetilde{\boldsymbol{X}}\in\mathcal{B}_r^\varepsilon(\boldsymbol{X})} \ell(f(\boldsymbol{A}, \widetilde{\boldsymbol{X}})_i, y_i)$, where $\widetilde{\boldsymbol{X}}_* = [\widetilde{\boldsymbol{x}}_{1*}, \dots, \widetilde{\boldsymbol{x}}_{n*}]$, $f(\cdot)_i \in \mathbb{R}$ is the prediction of node i, y_i is its true label, and $\ell : \mathbb{R} \to \mathbb{R}_+$ is a given loss function.
- **Adversarial loss:**
$$\widetilde{\ell}(f(\boldsymbol{A}, \boldsymbol{X})_i, y_i) := \max_{\widetilde{\boldsymbol{X}}\in\mathcal{B}_r^\varepsilon(\boldsymbol{X})} \ell(f(\boldsymbol{A}, \widetilde{\boldsymbol{X}})_i, y_i).$$
- **Adversarial training error:**
$$\widetilde{\mathcal{L}}_m(f) := \frac{1}{m}\sum_{i=1}^{m} \widetilde{\ell}(f(\boldsymbol{A}, \boldsymbol{X})_i, y_i),$$
- **Adversarial test error:**
$$\widetilde{\mathcal{L}}_u(f) := \frac{1}{n-m}\sum_{i=m+1}^{n} \widetilde{\ell}(f(\boldsymbol{A}, \boldsymbol{X})_i, y_i).$$
- **Generalization gap:**
$$\mathrm{Gen}(f) = \widetilde{\mathcal{L}}_u(f) - \widetilde{\mathcal{L}}_m(f).$$
- **Transductive Rademacher complexity:** Let $\mathcal{F} \subseteq \mathbb{R}^n$, $p \in [0, 0.5]$, and m the number of labeled samples. Let $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)$ be a vector of i.i.d. random variables, where σ_i takes the value $+1$ or -1 with probability p, and 0 with probability $1 - 2p$. Transductive Rademacher Complexity of $\mathcal{F}$ is defined as
$$\mathfrak{R}_{m,n}(\mathcal{F}) \triangleq \left(\frac{1}{m} + \frac{1}{n-m}\right)\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \boldsymbol{\sigma}^T f\right].$$
- **The hypothesis class of GCNs:**
$$\mathcal{F} = \left\{\boldsymbol{H}^{(L)} = \phi(g(\boldsymbol{A})\cdots\phi(g(\boldsymbol{A})\boldsymbol{X}\boldsymbol{W}^{(1)})\cdots\boldsymbol{W}^{(L)}) : \|\boldsymbol{W}^{(l)}\|_2, \|\boldsymbol{W}^{(l)}\|_p \leq \omega, l \in [L]\right\}, \qquad (1)$$
and the propagation procedure can be written as
$$\boldsymbol{H}^{(0)} = \boldsymbol{X}, \quad \boldsymbol{H}^{(l)} = \phi(g(\boldsymbol{A})\boldsymbol{H}^{(l-1)}\boldsymbol{W}^{(l)}), \quad l \in [L] \qquad (2)$$
where ω denotes the maximum bound over the $\|\cdot\|_2, \|\cdot\|_p$ of $\boldsymbol{W}^{(l)}$, $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d_{l-1}\times d_l}$ is a layer-specific weight matrix, d_l is the width of l-th layer, $d_0 = d$, $\phi(\cdot)$ is the ReLU function, and the graph filter $g(\boldsymbol{A}) : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ is a function of the adjacency matrix $\boldsymbol{A}$, such as

$g(\boldsymbol{A}) = \boldsymbol{A} + \boldsymbol{I}_n$	the graph with self-loops
$g(\boldsymbol{A}) = \boldsymbol{D}^{-1}\boldsymbol{A}$	the random-walk graph
$g(\boldsymbol{A}) = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$	the symmetric normalized graph

where $\boldsymbol{I}_n$ is the identity matrix, and $\boldsymbol{D}$ is the degree matrix defined by $\boldsymbol{D}_{i,i} = \sum_{j\in[n]} \boldsymbol{A}_{i,j}$.

Generalization Error Bound for Binary Classification

Assume that the range of the loss function ℓ is $[0, 1]$. With probability at least $1 - \delta$ for all $f \in \mathcal{F}$, we then have

$$\mathrm{Gen}(f) \leq \mathcal{O}(\max\{\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{n-m}}\}) + Q_{m,n}(\sqrt{2\log(2)L} + 1)\|g(\boldsymbol{A})\|_\infty^L \omega^L\big(B_{p^*}\|\boldsymbol{X}\|_{2,p^*} + \varepsilon s(r^*, p, d)\big)$$

where $B_{p*} = \sqrt{2\log(2d)}$, if $p = 1$; $B_{p*} = \sqrt{2}[\frac{\Gamma(\frac{1+p^*}{2})}{\sqrt{\pi}}]^{\frac{1}{p^*}}$, if $p \in (1, 2]$; $B_{p*} = 1$, if $p \in [2, +\infty)$, $Q_{m,n} = \sqrt{\frac{2n}{m(n-m)}}$, and $s(r^*, p, d) = d^{\max\{0, \frac{1}{r^*} - \frac{1}{p}\}}$.

- The upper bound above has an unavoidable dimension dependency, i.e., $s(r^*, p, d)$, which arises from the mismatch between the p-norm on the weight $\boldsymbol{W}^{(1)}$ and the r-norm in the adversarial node set $\mathcal{B}_r^\varepsilon(\boldsymbol{X})$. One could avoid such a dimension dependency by applying a perturbation-dependent norm regularizer on the weight matrix.
- The generalization error might increase exponentially with the number of layers L. It is worth noting that if $\omega = \mathcal{O}(1/\|g(\boldsymbol{A})\|_\infty)$ or selecting an appropriate graph filter, one can significantly weaken depth dependency and tighten the bound.

Generalization Error Bound for Multi-class Classification

Assume that the range of the loss function ℓ is $[0, 1]$. With probability at least $1 - \delta$ for all $f \in \mathcal{F}$, for any fixed $\gamma > 0$, $\mathrm{Gen}(f)$ is upper-bounded by

$$\mathcal{O}(\max\{\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{n-m}}\}) + Q_{m,n}\frac{4K}{\gamma}(\sqrt{\log(2)L} + 1)\|g(\boldsymbol{A})\|_\infty^L \omega^L\left(B_{p^*}\|\boldsymbol{X}\|_{2,p^*} + \varepsilon s(r^*, p, d)\right)$$

where $B_{p*} = \sqrt{2\log(2d)}$, if $p = 1$; $B_{p*} = \sqrt{2}[\frac{\Gamma(\frac{1+p^*}{2})}{\sqrt{\pi}}]^{\frac{1}{p^*}}$, if $p \in (1, 2]$; $B_{p*} = 1$, if $p \in [2, +\infty)$, $Q_{m,n} = \sqrt{\frac{2n}{m(n-m)}}$, and $s(r^*, p, d) = d^{\max\{0, \frac{1}{r^*} - \frac{1}{p}\}}$.

- The convergence rate of $\mathcal{O}(K)$ in the number of classes K is comparable with the existing generalization bounds for traditional multi-class classification tasks. Letting $\varepsilon = 0$, we also obtain the high-probability generalization bound of GCNs for multi-class classification.

Explicit Bounds for GCN Models

- **SGC.** $\boldsymbol{H}^{(L)} = \mathrm{Softmax}(g(\boldsymbol{A})^L\boldsymbol{X}\boldsymbol{W}^{(1)}\cdots\boldsymbol{W}^{(L)})$.
For any $\delta \in (0, 1)$, with probability $1 - \delta$,
$$\mathrm{Gen}(f) \leq \mathcal{O}(\max\{\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{n-m}}\}) + Q_{m,n}\frac{2K}{\gamma}\|g(\boldsymbol{A})^L\|_\infty\omega^L(B_{p^*}\|\boldsymbol{X}\|_{2,p^*} + \varepsilon s(r^*, p, d)).$$
It is worth noting that $\|g(\boldsymbol{A})^L\|_\infty \leq \|g(\boldsymbol{A})\|_\infty^L$, thereby alleviating the negative impact of perturbation-relevant term and leading to a tighter generalization bound.
- **Residual GCN.** $\boldsymbol{H}^{(l)} = \phi(g(\boldsymbol{A})\boldsymbol{H}^{(l-1)}W^{(l)}) + \boldsymbol{H}^{(l-1)}, \quad \boldsymbol{H}^{(0)} = \phi(\boldsymbol{X}\boldsymbol{W}^{(0)})$.
For any $\delta \in (0, 1)$, with probability $1 - \delta$,
$$\begin{aligned}\mathrm{Gen}(f) \leq &\mathcal{O}\big(\max\{\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{n-m}}\}\big)\\ &+ Q_{m,n}\frac{4K}{\gamma}(\sqrt{\log(2)L} + 1)\|g(\boldsymbol{A})\|_\infty^L\omega(\omega + 1)^L(B_{p^*}\|\boldsymbol{X}\|_{2,p^*} + \varepsilon s(r^*, p, d)).\end{aligned}$$
The above bound suggests that residual connections partially alleviate over-smoothing.
- **GCNII.** $\boldsymbol{H}^{(l)} = \phi(((1-\alpha)g(\boldsymbol{A})\boldsymbol{H}^{(l-1)} + \alpha\boldsymbol{H}^{(l)})((1-\beta)\boldsymbol{I}_n + \beta\boldsymbol{W}^{(l)})), \quad \boldsymbol{H}^{(0)} = \phi(\boldsymbol{X}\boldsymbol{W}^{(0)})$.
For any $\delta \in (0, 1)$ and $\alpha, \beta \in (0, 1)$, with probability $1 - \delta$,
$$\begin{aligned}\mathrm{Gen}(f) \leq &\mathcal{O}(\max\{\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{n-m}}\}) + Q_{m,n}\frac{4K}{\gamma}(\sqrt{\log(2)L} + 1)(B_{p^*}\|\boldsymbol{X}\|_{2,p^*} + \varepsilon s(r^*, p, d))\\ &\times \omega^2\Big((1-\alpha)\|g(\boldsymbol{A})\|_\infty^L(1 - \beta + \beta\omega)^L + \alpha(1-\alpha)\sum_{l=0}^{L}\|g(\boldsymbol{A})\|_\infty^l(1 - \beta + \beta\omega)^l\Big).\end{aligned}$$
As α increases, the perturbation term will significantly weaken the dependency on depth, thereby reducing the generalization error. Additionally, if β approaches zero and $\|g(\boldsymbol{A})\|_\infty \leq 1$, this bound is independent of the number of layers.

Partial Experimental Results

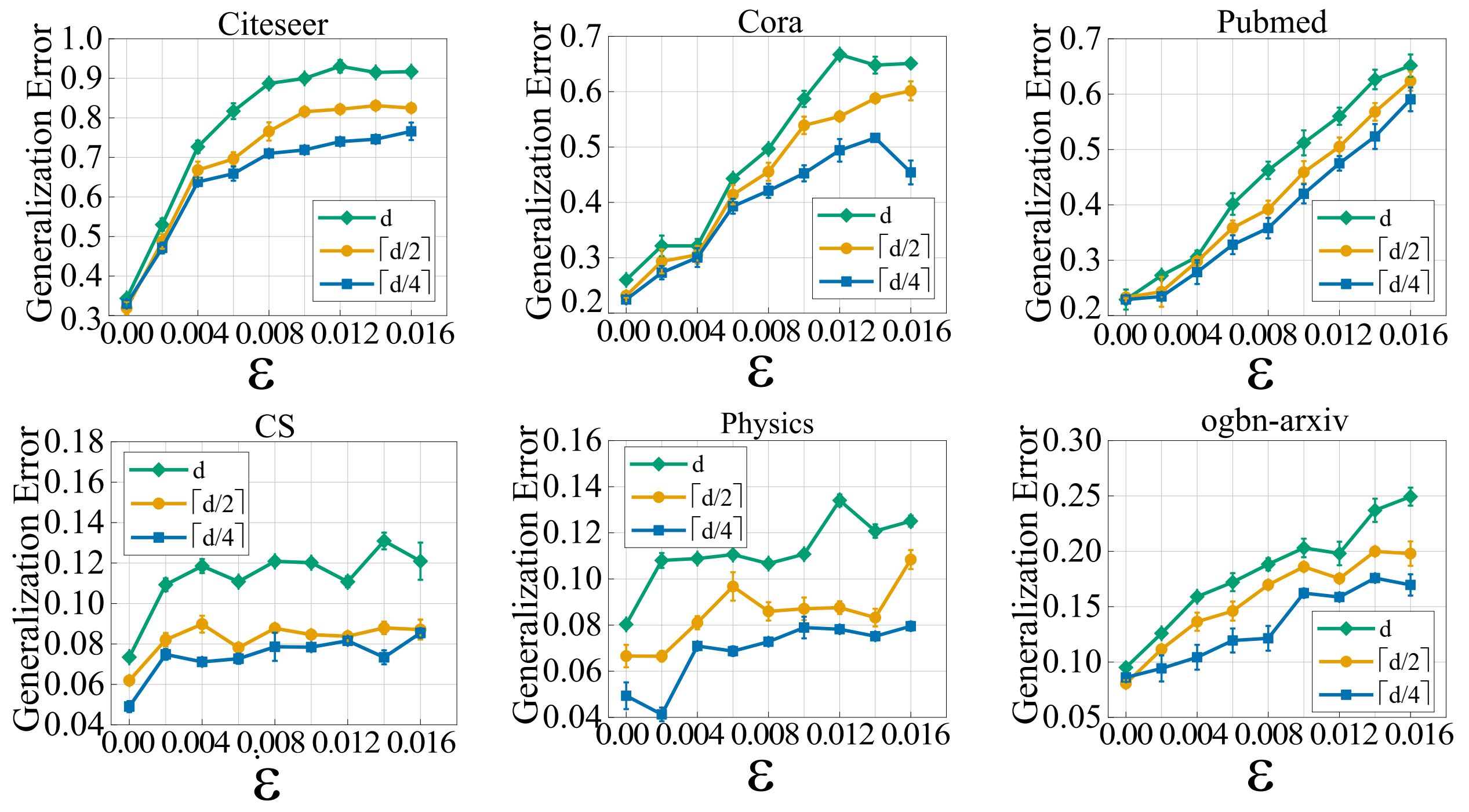

Figure 1. The empirical generalization error (mean value and standard deviation) with different feature dimensions.

Note. The empirical generalization error decreases steadily with the dimension, which implies that low-dimensional feature projection can help reduce the generalization error.

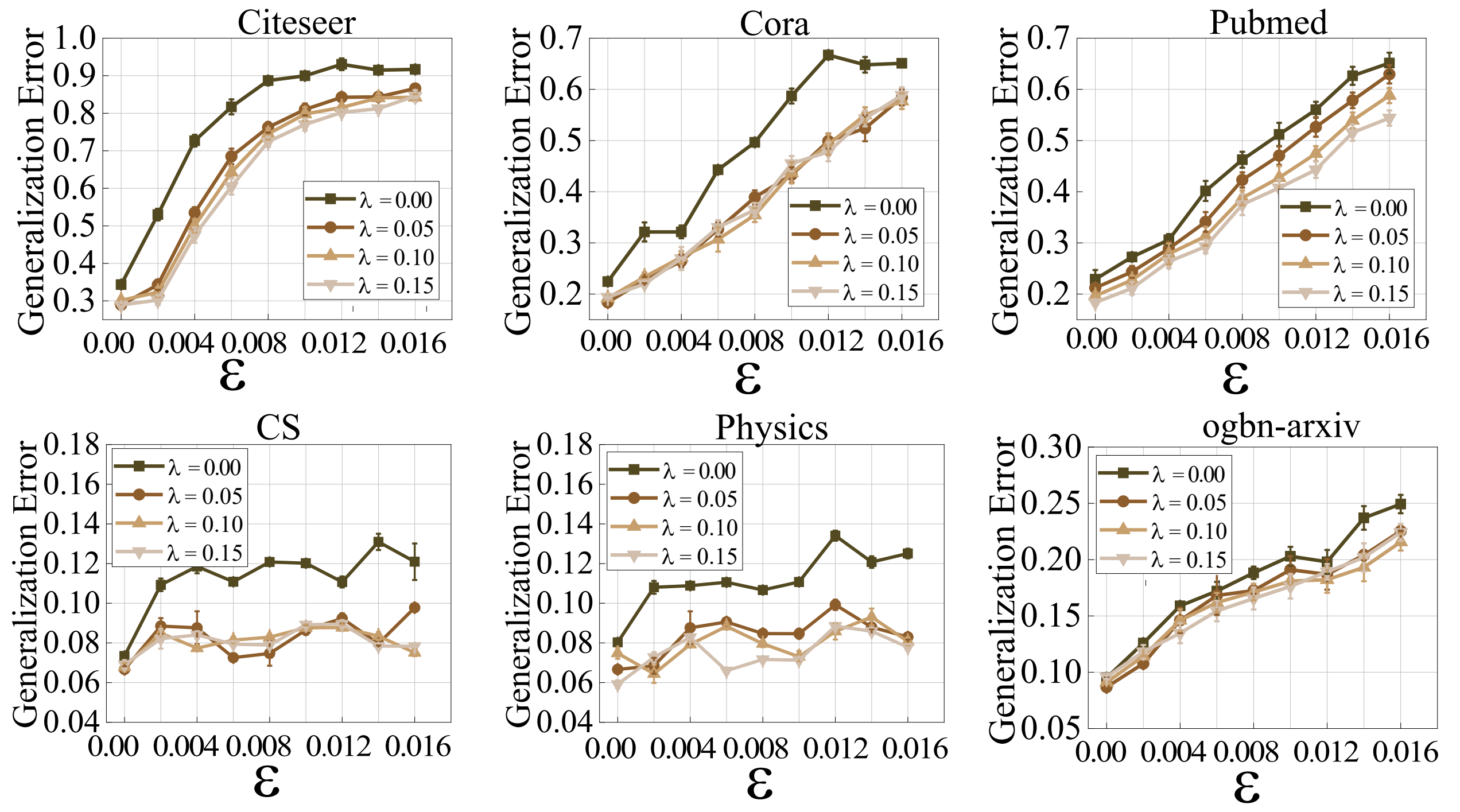

Figure 2. The empirical generalization error (mean value and standard deviation) of models trained with ℓ_1 regularization for different regularization parameters (i.e., λ).

Note. The empirical generalization error of the regularized model is smaller than that without (i.e., $\lambda = 0$), which demonstrating the importance of an appropriate regularizer to achieve good generalization performance.

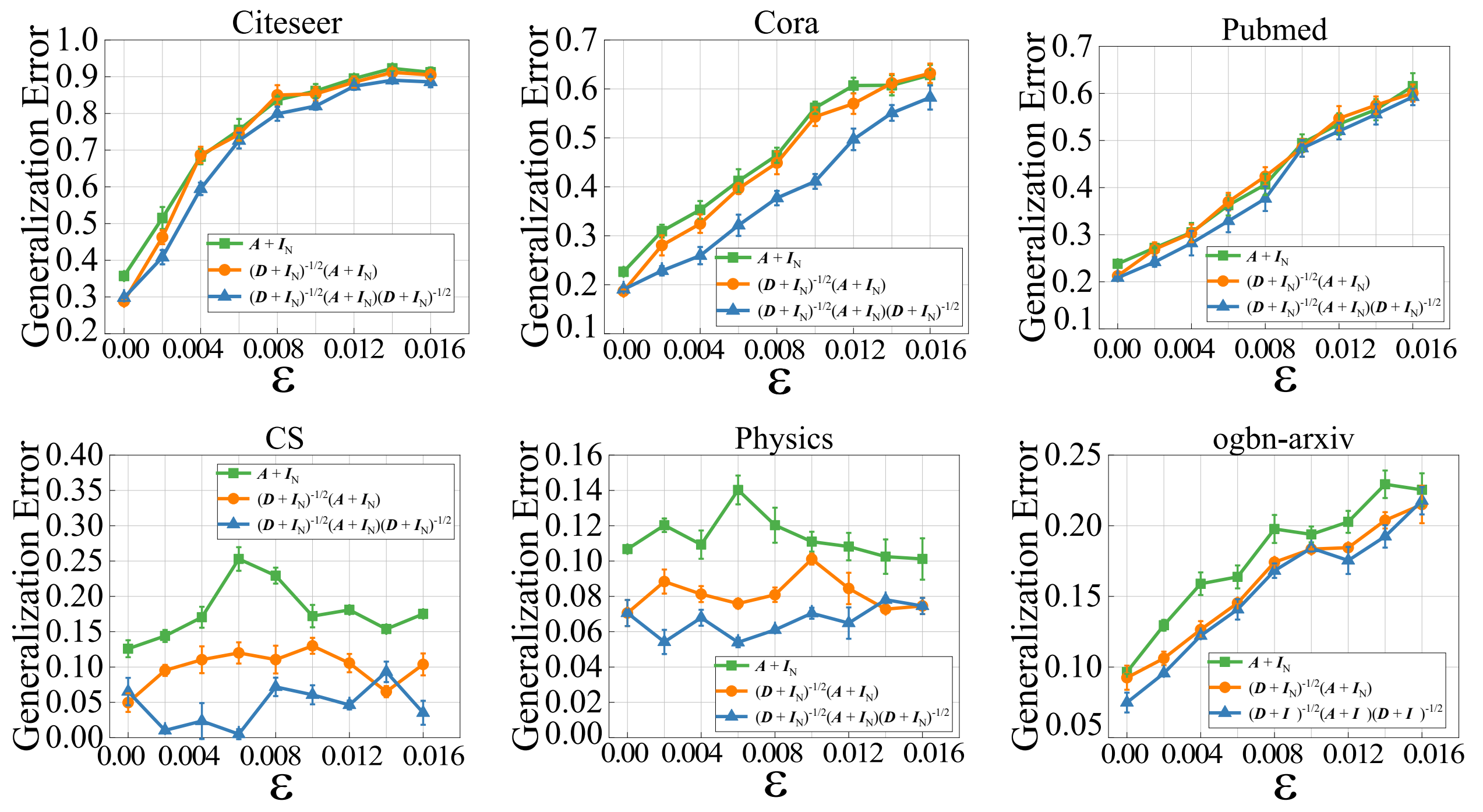

Figure 3. The empirical generalization error (mean value and standard deviation) with graph filters, where depth is set to 6. ε denotes the maximum allowable perturbation.

Note. As shown in Figure 3, the graph with self-loops has larger empirical generalization errors than the normalized graphs. Hence, we argue that normalizing the graph matrix can facilitate the adversarial generalization of GCNs.