Bridging the Gap: Rademacher Complexity in Robust and Standard Generalization

Jiancong Xiao JCXIAO@UPENN.EDU
University of Pennsylvania
Ruoyu Sun SUNRUOYU@CUHK.EDU.CN
The Chinese University of Hong Kong, Shenzhen
Qi Long QLONG@UPENN.EDU
University of Pennsylvania
Weijie J Su SUW@WHARTON.UPENN.EDU
University of Pennsylvania

Editors: Shipra Agrawal and Aaron Roth

Adversarial Rademacher complexity was introduced to bound adversarially robust generalization (Khim and Loh, 2018; Yin et al., 2019). However, despite the dedication of numerous works (Awasthi et al., 2020; Gao and Wang, 2021; Xiao et al., 2022; Mustafa et al., 2022) to this problem, achieving a satisfactory bound remains an elusive goal. Existing works on deep neural networks (DNNs) either apply to a surrogate loss or yield bounds that are notably looser compared to their standard counterparts.[1]

This paper presents upper bounds for adversarial Rademacher complexity of DNNs that matches the best-known upper bounds in standard settings, as established in the work of Bartlett et al. (2017). The dependency on width and dimension improve from at least $\mathcal{O}(\sqrt{m})$ or $\mathcal{O}(\sqrt{d})$ to $\mathcal{O}(\ln(dm))$. This provides a new insight on understanding robust generalization: the complexity of standard and robust generalization is nearly identical.

To state the bound, some notation is necessary. The notation mainly follows the work of Bartlett et al. (2017). The networks will use L fixed activation functions $(\sigma_1, \cdots, \sigma_L)$, where σ_i is ρ_i-Lipschitz and $\sigma_i(0) = 0$. Let $\ell(\cdot, y)$ be a ρ-Lipshitz function with respect to the first argument and takes values in $[0, 1]$. Given L weight matrices $W = (W_1, \cdots, W_L)$ with $W_l \in \mathbb{R}^{m_l \times m_{l-1}}$, let the deep neural networks be $f(x) = \sigma_L W_L \sigma_{L-1}(W_{L-1} \cdots \sigma_1(W_1 x) \cdots)$. The network output $f(x) \in \mathbb{R}^{m_L}$ (with $m_0 = d$ and $m_L = k$) is converted to a class label in $\{1, \cdots, k\}$ by taking the $\arg\max$ over components, with an arbitrary rule for breaking ties. Whenever input data $x_1, \cdots, x_n \in \mathbb{R}^d$ are given with $\|x_i\|_2 \le B$, collect them as columns of a matrix $X \in \mathbb{R}^{d \times n}$. Let $\mathcal{B}(x)$ be arbitrary perturbation set around x. For example, for ℓ_p attack, we denote $\mathcal{B}_\varepsilon^p(x) = \{x' \mid \|x - x'\|_p \le \varepsilon\}$. The ℓ_p norm $\|\cdot\|_p$ is always computed entry-wise. Thus, for a matrix, $\|\cdot\|_2$ corresponds to the Frobenius norm. Finally, let $\|\cdot\|_\sigma$ denote the spectral norm.

Theorem 1 *Let nonlinearities $(\sigma_1, \cdots, \sigma_L)$ be given as above. Let the network $f : \mathbb{R}^d \to \mathbb{R}^k$ with weight matrices $W = (W_1, \cdots, W_L)$ have spectral norm bounds $(s_1, \cdots, s_L)$ and ℓ_1-norm bounds $(a_1, \cdots, a_L)$. Then for $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ drawn i.i.d. from any probability distribution $\mathcal{D}$ over $\mathbb{R}^d \times \{1, \cdots, k\}$, with probability at least $1 - \delta$ over $\mathcal{S}$, the adversarially robust generalization gap satisfies*

$$\mathbb{E}_{\mathcal{D}} \max_{x' \in \mathcal{B}(x)} \ell(f(x'), y) - \mathbb{E}_{\mathcal{S}} \max_{x' \in \mathcal{B}(x)} \ell(f(x'), y) \le \tilde{\mathcal{O}}\left(\frac{\tilde{B} \rho \prod_{i=1}^L \rho_i s_i}{\sqrt{n}} \left(\sum_{i=1}^L \frac{a_i^{2/3}}{s_i^{2/3}} \right)^{3/2} + \sqrt{\frac{\ln(1/\delta)}{n}} \right),$$

where $\tilde{B}$ is the magnitude of adversarial examples, i.e., $\|x'\|_2 \le \tilde{B}$, $\forall x' \in \mathcal{B}(x)$ and $x \in \{x_i\}_{i=1}^n$.

1. Extended abstract. Full version appears as [arXiv reference, 2406.05372].

References

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pages 431–441. PMLR, 2020.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Qingyi Gao and Xiao Wang. Theoretical investigation of generalization bounds for adversarial learning of deep neural networks. *Journal of Statistical Theory and Practice*, 15(2):1–28, 2021.

Justin Khim and Po-Ling Loh. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Waleed Mustafa, Yunwen Lei, and Marius Kloft. On the generalization analysis of adversarial learning. In *International Conference on Machine Learning*, pages 16174–16196. PMLR, 2022.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, and Zhi-Quan Luo. Adversarial rademacher complexity of deep neural networks. *arXiv preprint arXiv:2211.14966*, 2022.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pages 7085–7094. PMLR, 2019.