PAC-Bayesian Spectrally-Normalized Bounds for Adversarially Robust Generalization

Jiancong Xiao , Ruoyu Sun , Zhi-Quan Luo
The Chinese University of Hong Kong, Shenzhen, China

Nov 2023

Background: Adversarial Defense

Optimization problem of adversarial defense (given n samples)

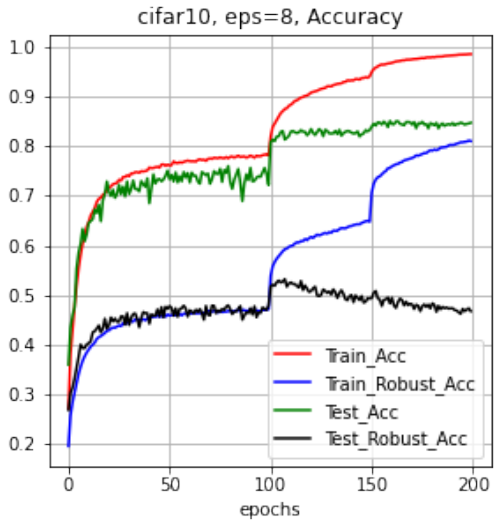
$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} \max_{\|x_i - x_i'\|_p \leq \epsilon} \ell(f_w(x_i'), y_i), \tag{1}$$

Adversarial Training:

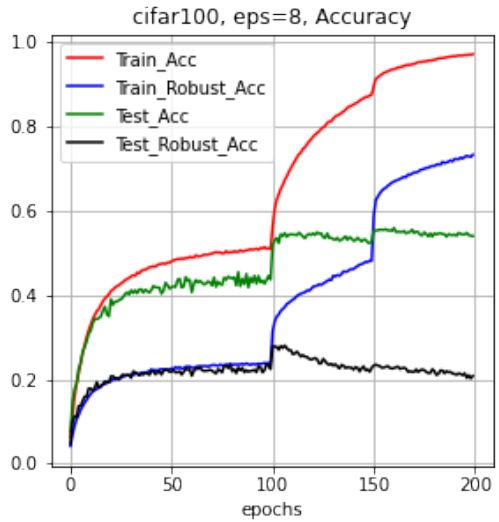
- Training data = clean data + adversarial data
- SOTA defense
- 65% robust accuracy on CIFAR-10 (DeepMind)
- Far from satisfactory

Background: Adversarial Defense

Robust Overfitting / Robust Generalization:



(a)



(b)

- Standard training (only clean data): good generalization
- Adversarial training: poor robust generalization

Why robust generalization gap is large?

Overview

Norm-based Complexity:

- Let n be the number of samples and the training samples x is bounded by B. Let f be a d-layer feedforward network. Then, with high probability, we have

$$\text{Generalization} \leq \mathcal{O}(B\Pi_{i=1}^{d}\|W_i\|/\sqrt{n}).$$

- As $n \to +\infty$, bound $\to 0$

Extension to Robust settings:

- Unsolve problem
- Previous work have tried Rademacher complexity, Covering number, Pac-bayes approach
- No satisfactory solution

Overview

For General Audience:

- Main Results (Informal): Let n be the number of samples and the training samples x is bounded by B. ϵ is the attack intensity. Let f be a d-layer feedforward network. Then, with high probability, we have

$$\text{Robust Generalization} \leq \mathcal{O}((B+\epsilon)\Pi_{i=1}^{d}\|W_i\|/\sqrt{n}).$$

- As $\epsilon \to 0$, reduce to standard generalization bound
- As $n \to +\infty$, bound $\to 0$
- Attack intensity $\times$ Norm-Based Complexity $\approx\to$ Robust Overfitting or Generalization

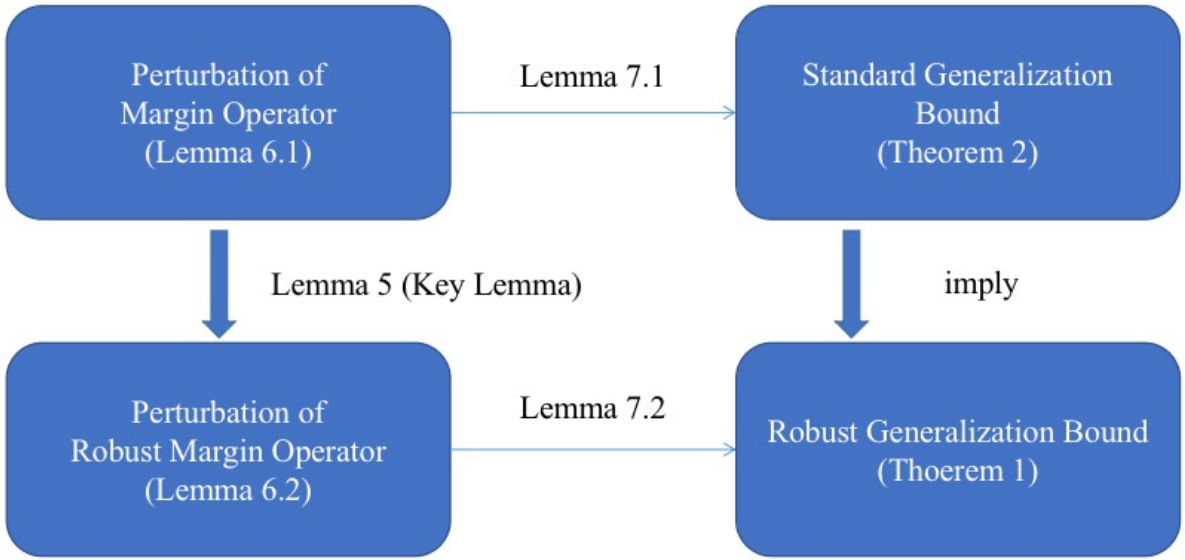
For Theory Researchers:

- Mathematical difficulty over the past few years

Pac-Bayesian Bound

Pac-Bayesian Approach

- [Neyshabur et al., 2017] provided a simpler proof for (standard) generalization bound
- Unclear how to extend to robust setting [Farnia et al., 2018]
- Restructure the proof of [Neyshabur et al., 2017] to incorporate IAE (Key Lemma)



Thank you!

References I

Farnia, F., Zhang, J. M., and Tse, D. (2018).
Generalizable adversarial training via spectral normalization.
arXiv preprint arXiv:1811.07457.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. (2017).
A pac-bayesian approach to spectrally-normalized margin bounds for neural networks.
arXiv preprint arXiv:1707.09564.