Rademacher Complexity for Adversarially Robust Generalization

Dong Yin

joint work with

Kannan Ramchandran, Peter Bartlett
UC Berkeley

ICML, 2019

Introduction

Machine learning models are vulnerable to adversarial perturbations.

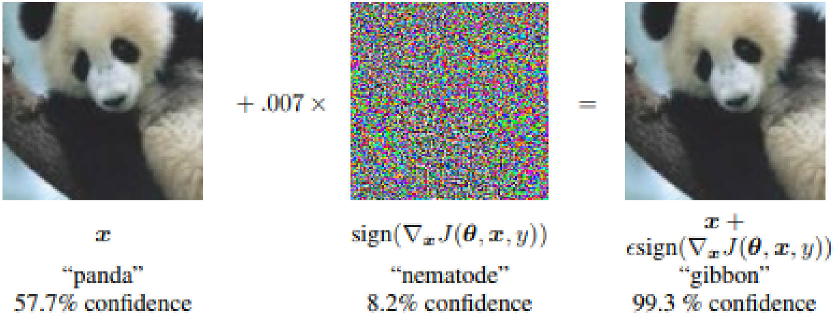

Figure: Adding invisible perturbations to the images can lead the model to wrong predictions with high confidence (Goodfellow et al. 2015)

Introduction

- Adversarial training: currently the most effective approach to training models robust to adversarial perturbations (Madry et al, 2017).
- Natural training:

$$\min_{f \in \mathcal{F}} \mathbb{E} \ell(f(\mathbf{x}), y)$$

- Adversarial training:

$$\min_{f \in \mathcal{F}} \mathbb{E} \max_{\mathbf{x}' \in \mathbb{B}_{\mathbf{x}}^{\infty}(\epsilon)} \ell(f(\mathbf{x}'), y)$$

Introduction

Adversarially robust generalization can be hard.

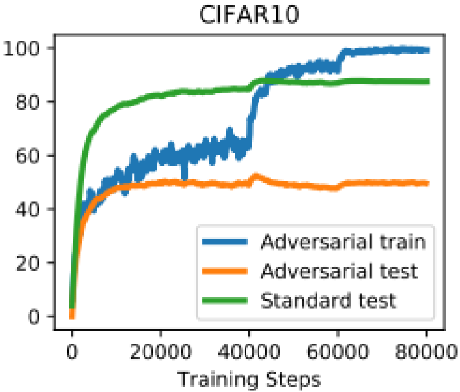

Figure: Model can achieve 96% adversarial training accuracy whereas the adversarial test accuracy is only 47% (Madry et al. 2017, Schmidt et al. 2018)

Introduction

- How can we better understand adversarially robust generalization?
- This paper: Rademacher complexity analysis.

Preliminaries

- Feature-label space $\mathcal{X} \times \mathcal{Y}$.
- Hypothesis class $\mathcal{F}$.
- Loss function $\ell(f(\mathbf{x}), y)$, $f \in \mathcal{F}$, $\ell_{\mathcal{F}} := \{\ell(f(\cdot), \cdot) : f \in \mathcal{F}\}$.
- Empirical risk: $R_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i)$
- Population risk $R(f) := \mathbb{E}[\ell(f(\mathbf{x}), y)]$
- Rademacher complexity

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}) := \frac{1}{n} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(\mathbf{x}_i) \right],$$

Preliminaries

Theorem

(Bartlett and Mendelson, 2002, Mohri et al. 2012) Suppose that $\ell(f(\mathbf{x}), y) \in [0, 1]$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following holds for all $f \in \mathcal{F}$,

$$R(f) \leq R_n(f) + 2\mathfrak{R}_{\mathcal{S}}(\ell_{\mathcal{F}}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Preliminaries

- Adversarial loss function $\widetilde{\ell}(f(\mathbf{x}), y) := \max_{x' \in \mathbb{B}_{\mathbf{x}}^{\infty}(\epsilon)} \ell(f(\mathbf{x}'), y)$, $f \in \mathcal{F}$
- Adversarial empirical risk:

$$\widetilde{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \widetilde{\ell}(f(\mathbf{x}_i), y_i).$$

- Adversarial population risk

$$\widetilde{R}(f) := \mathbb{E}[\widetilde{\ell}(f(\mathbf{x}), y)],$$

- Adversarial Rademacher complexity

$$\mathfrak{R}_{\mathcal{S}}(\widetilde{\ell}_{\mathcal{F}}), \text{ where } \widetilde{\ell}_{\mathcal{F}} := \{\widetilde{\ell}(f(\cdot), \cdot) : f \in \mathcal{F}\}$$

Preliminaries

Corollary

For any $\delta \in (0,1)$, *with probability at least* $1-\delta$, *the following holds for all* $f \in \mathcal{F}$,

$$\widetilde{R}(f) \leq \widetilde{R}_n(f) + 2\mathfrak{R}_{\mathcal{S}}(\widetilde{\ell}_{\mathcal{F}}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

How do we compare $\mathfrak{R}_{\mathcal{S}}(\ell_{\mathcal{F}})$ and $\mathfrak{R}_{\mathcal{S}}(\widetilde{\ell}_{\mathcal{F}})$?

Main Results

Binary linear classifier

Theorem

Let $\mathcal{F} := \{\langle \mathbf{x}, \mathbf{w} \rangle : \|\mathbf{w}\|_p \le W\}$ *and* $\widetilde{\mathcal{F}} := \{\min_{\mathbf{x}' \in \mathbb{B}_{\mathbf{x}}^{\infty}(\epsilon)} y \langle \mathbf{w}, \mathbf{x}' \rangle : \|\mathbf{w}\|_p \le W\}$*. Suppose that* $\frac{1}{p} + \frac{1}{q} = 1$*. Then, there exists a universal constant* $c \in (0,1)$ *such that*

$$\max\{\mathfrak{R}_{\mathcal{S}}(\mathcal{F}), c\epsilon W \frac{d^{\frac{1}{q}}}{\sqrt{n}}\} \le \mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}) \le \mathfrak{R}_{\mathcal{S}}(\mathcal{F}) + \epsilon W \frac{d^{\frac{1}{q}}}{\sqrt{n}}.$$

- **Tight** upper and lower bounds.
- Adversarial Rademacher complexity is **never smaller** than its natural counterpart.
- **Unavoidable dimension dependence** in adversarial Rademacher complexity (unless $p = 1$).

Main Results

- Multi-class linear classifiers: similar dimension dependence also exists in the margin-based risk bound.
- Lower bound of adversarial Rademacher complexity for neural networks: existence of dimension dependence.
- Risk bound on the adversarial loss for one-hidden layer ReLU network via SDP surrogate loss (Raghunathan et al. 2018).

Experiments

MNIST, Linear classifier

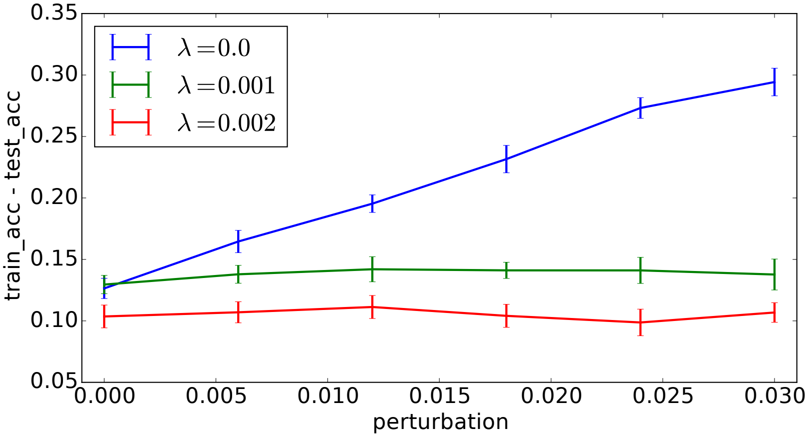

Figure: ℓ_1 regularization reduces adversarial generalization gap.

Experiments

MNIST, Linear classifier

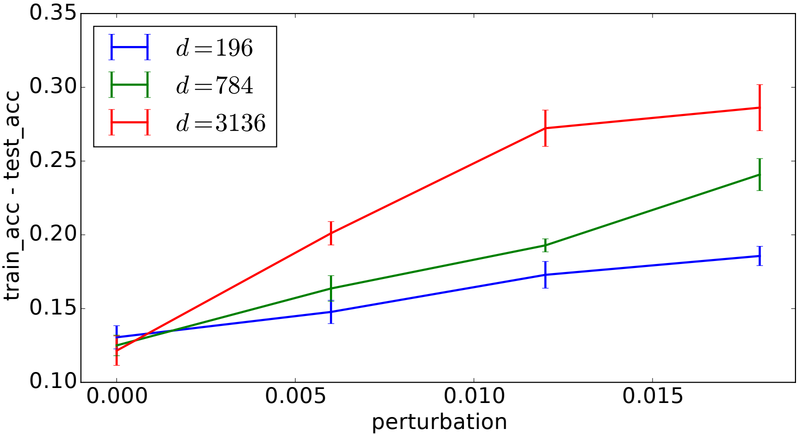

Figure: Adversarial generalization gap becomes larger in higher dimensions.

Experiments

MNIST, four layer CNN

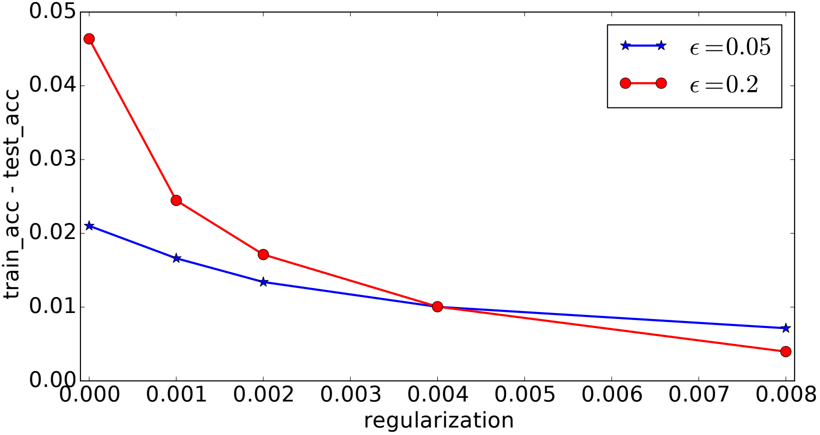

Figure: ℓ_1 regularization reduces adversarial generalization gap.

Thank you

Poster: Pacific Ballroom 207