Theoretical Analysis of Adversarial Learning: A Minimax Approach

Zhuozhuo Tu[1], Jingwei Zhang[2,1], Dacheng Tao[1]

[1]The University of Sydney

[2]The Hong Kong University of Science and Technology

NeurIPS 2019

Overview

Goal: theoretically understand the generalization performance in the presence of adversaries.

$$\mathbb{E}_{(x,y)\sim P}[\max_{x'\in N(x)} l(h(x'),y)]$$

Our method applies to

- General adversaries encompassing all l_q-bounded adversaries.
- Multi-class problems and popular loss functions such as the hinge loss and ramp loss.

Standard vs Adversarial Expected Risk

Standard vs Adversarial Expected Risk

Standard expected risk:

$$R_P(h) = \mathbb{E}_{(x,y)\sim P}[l(h(x), y)].$$

Standard vs Adversarial Expected Risk

Standard expected risk:

$$R_P(h) = \mathbb{E}_{(x,y)\sim P}[l(h(x), y)].$$

The adversarial expected risk:

$$R_P(h, \mathcal{B}) = \mathbb{E}_{(x,y)\sim P}[\max_{x' \in N(x)} l(h(x'), y)],$$

where $N(x) = \{x' : x' - x \in \mathcal{B}\}$.

Standard vs Adversarial Expected Risk

Standard expected risk:

$$R_P(h) = \mathbb{E}_{(x,y)\sim P}[l(h(x), y)].$$

The adversarial expected risk:

$$R_P(h, \mathcal{B}) = \mathbb{E}_{(x,y)\sim P}[\max_{x' \in N(x)} l(h(x'), y)],$$

where $N(x) = \{x' : x' - x \in \mathcal{B}\}$.

The relationship between the two notions of expected risk?

Proposed Methodology

Challenge: The inner maximization problem is usually very hard to solve directly.

Proposed Methodology

Challenge: The inner maximization problem is usually very hard to solve directly.

Trick: (1) Introduce a transport map $T_h : \mathcal{Z} \to \mathcal{Z}$ such that

$$R_P(h, \mathcal{B}) = R_{P'}(h)$$

Proposed Methodology

Challenge: The inner maximization problem is usually very hard to solve directly.

Trick: (1) Introduce a transport map $T_h : \mathcal{Z} \to \mathcal{Z}$ such that

$$R_P(h, \mathcal{B}) = R_{P'}(h)$$

(2) Prove that all these distributions P' locate within a Wasserstein ball centered at P

$$W_p(P, P') \leq \epsilon_{\mathcal{B}}$$

Proposed Methodology

Challenge: The inner maximization problem is usually very hard to solve directly.

Trick: (1) Introduce a transport map $T_h : \mathcal{Z} \to \mathcal{Z}$ such that

$$R_P(h, \mathcal{B}) = R_{P'}(h)$$

(2) Prove that all these distributions P' locate within a Wasserstein ball centered at P

$$W_p(P, P') \leq \epsilon_{\mathcal{B}}$$

(3) The adversarial expected risk is upper bounded as

$$R_P(h, \mathcal{B}) \leq R_{\epsilon_{\mathcal{B}},1}(P, h), \quad \forall h \in \mathcal{H}$$

Main Theorem

Main Theorem

Assume that for any function $f \in \mathcal{F}$ and any $z \in \mathcal{Z}$, there exists $\lambda_{f,z}$ such that $f(z') - f(z) \leq \lambda_{f,z} d_{\mathcal{Z}}(z, z')$ for any $z' \in \mathcal{Z}$.

Main Theorem

Assume that for any function $f \in \mathcal{F}$ and any $z \in \mathcal{Z}$, there exists $\lambda_{f,z}$ such that $f(z') - f(z) \leq \lambda_{f,z} d_{\mathcal{Z}}(z, z')$ for any $z' \in \mathcal{Z}$.

Theorem (Adversarial Risk Bounds)

Under the assumption, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for any $f \in \mathcal{F}$:

$$R_P(f, \mathcal{B}) \leq \frac{1}{n} \sum_{i=1}^{n} f(z_i) + \lambda^{+}_{f,P_n} \epsilon_{\mathcal{B}} + \frac{24\mathfrak{C}(\mathcal{F})}{\sqrt{n}} + \frac{12\sqrt{\pi}}{\sqrt{n}} \Lambda_{\epsilon_{\mathcal{B}}} \cdot diam(Z) + M\sqrt{\frac{log(\frac{1}{\delta})}{2n}},$$

where $\lambda^{+}_{f,P_n} := \inf\{\lambda : \mathbb{E}_{P_n}(\sup_{z' \in \mathcal{Z}}\{f(z') - \lambda d_{\mathcal{Z}}(z, z') - f(z)\}) = 0\}$

Main Theorem

Assume that for any function $f \in \mathcal{F}$ and any $z \in \mathcal{Z}$, there exists $\lambda_{f,z}$ such that $f(z') - f(z) \leq \lambda_{f,z} d_{\mathcal{Z}}(z, z')$ for any $z' \in \mathcal{Z}$.

Theorem (Adversarial Risk Bounds)

Under the assumption, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for any $f \in \mathcal{F}$:

$$R_P(f, \mathcal{B}) \leq \frac{1}{n} \sum_{i=1}^{n} f(z_i) + \lambda^{+}_{f,P_n} \epsilon_{\mathcal{B}} + \frac{24\mathfrak{C}(\mathcal{F})}{\sqrt{n}} + \frac{12\sqrt{\pi}}{\sqrt{n}} \Lambda_{\epsilon_{\mathcal{B}}} \cdot diam(Z) + M\sqrt{\frac{log(\frac{1}{\delta})}{2n}},$$

where $\lambda^{+}_{f,P_n} := \inf\{\lambda : \mathbb{E}_{P_n}(\sup_{z' \in \mathcal{Z}}\{f(z') - \lambda d_{\mathcal{Z}}(z, z') - f(z)\}) = 0\}$

Example Bounds

We illustrate the application of our method to two commonly-used models: SVMs and neural networks.

Example Bounds

We illustrate the application of our method to two commonly-used models: SVMs and neural networks.

Corollary

- Support Vector Machines

$$R_P(f,\mathcal{B}) \leq \frac{1}{n}\sum_{i=1}^{n} f(z_i) + \lambda^{+}_{f,P_n}\epsilon_{\mathcal{B}} + \frac{144}{\sqrt{n}}\Lambda r\sqrt{d} + \frac{12\sqrt{\pi}}{\sqrt{n}}\Lambda_{\epsilon_{\mathcal{B}}}\cdot(2r+1) + (1+\Lambda r)\sqrt{\frac{log(\frac{1}{\delta})}{2n}},$$

where $\lambda^{+}_{f,P_n} \leq \max_i\{2y_i w\cdot x_i, ||w||_2\}$

- Neural Networks

$$R_P(f,\mathcal{B}) \leq \frac{1}{n}\sum_{i=1}^{n} f(z_i) + \lambda^{+}_{f,P_n}\epsilon_{\mathcal{B}} + \frac{288}{\gamma\sqrt{n}}\prod_{i=1}^{L}\rho_i s_i BW\left(\sum_{i=1}^{L}\left(\frac{b_i}{s_i}\right)^{\frac{1}{2}}\right)^2 + \frac{12\sqrt{\pi}}{\sqrt{n}}\Lambda_{\epsilon_{\mathcal{B}}}\cdot(2B+1) + \sqrt{\frac{log(1/\delta)}{2n}},$$

where $\lambda^{+}_{f,P_n} \leq \max_j\{\frac{2}{\gamma}\prod_{i=1}^{L}\rho_i||A_i||_\sigma, \frac{1}{\gamma}(\mathcal{M}(\mathcal{H}_{\mathcal{A}}(x_j),y_j) + \max\mathcal{H}_{\mathcal{A}}(x_j) - \min\mathcal{H}_{\mathcal{A}}(x_j))\}$

Thank You

Welcome to our poster session for further discussions!
Today, 05:00 PM–07:00 PM
East Exhibition Hall B + C #238