Theoretical Analysis of Adversarial Learning: A Minimax Approach

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

[1]School of Computer Science, The University of Sydney [2]Department of Computer Science and Engineering, HKUST

THE UNIVERSITY OF SYDNEY
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

MAIN CONTRIBUTIONS

- Propose a general method for analyzing the risk bound in the presence of adversaries. Our method is general in several respects. First, the adversary we consider is general and encompasses all l_q bounded adversaries. Second, our method can be applied to multi-class problems and commonly used loss functions such as the hinge loss and ramp loss.
- Prove a new bound for the local worst-case risk under a weak version of Lipschitz condition.
- Derive the adversarial risk bounds for SVMs and deep neural networks. Our bounds have two data-dependent terms, suggesting that minimizing the sum of the two terms can help achieve adversarial robustness.

ADVERSARIAL LEARNING

The adversarial learning problem can be described as follows.

- The learner receives n training examples denoted by $S = ((x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n))$ drawn i.i.d. from P and tries to select a hypothesis $h \in \mathcal{H}$ that has a small expected risk.
- However, in the presence of adversaries, there will be imperceptible perturbations to the input of examples, which are called adversarial examples.
- We assume that the adversarial examples are generated by adversarially choosing an example from neighborhood $N(x) = \{x' : x' - x \in \mathcal{B}\}$ where $\mathcal{B}$ is a nonempty set. The radius of the adversary is defined as $\epsilon_{\mathcal{B}} := \sup_{x \in \mathcal{B}} d_{\mathcal{X}}(x, 0)$

To measure the learner's performance in the presence of adversaries, we define the adversarial expected risk of a hypothesis $h \in \mathcal{H}$ as

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

If $\epsilon_{\mathcal{B}} = 0$, then the adversarial expected risk will reduce to the standard expected risk without an adversary.

Since the true distribution is usually unknown, we instead consider adversarial empirical risk.

$$R_{P_n}(h, \mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n}\left[\max_{x' \in N(x_i)} l(h(x'), y_i)\right].$$

MINIMAX LEARNING

- Wasserstein distance between two probability measures $P, Q \in \mathcal{P}_p(\mathcal{Z})$ is defined as

$$W_p(P, Q) := \inf_{M \in \Gamma(P,Q)} (\mathbb{E}_{(z,z')\sim M}[d^p_{\mathcal{Z}}(z, z')])^{1/p},$$

  where $\Gamma(P, Q)$ denotes the collection of all measures on $\mathcal{Z} \times \mathcal{Z}$ with marginals P and Q on the first and second factors, respectively.
- The local worst-case risk of h at P,

$$R_{\epsilon,p}(P, h) := \sup_{Q \in B^W_{\epsilon,p}(P)} R_Q(h),$$

  where $B^W_{\epsilon,p}(P) := \{Q \in \mathcal{P}_p(Z) : W_p(P, Q)) \leq \epsilon\}$ is the p-Wasserstein ball of radius $\epsilon \geq 0$ centered at P.

MAIN RESULTS

Motivation

- The adversarial expected risk over a distribution P is equivalent to the standard expected risk under a new distribution P'.
- We can show that all these new distributions locate within a Wasserstein ball centered at P.
- By considering the worst case within this Wasserstein ball, the original adversarial learning problem can be reduced to a minimax problem, and we can use the minimax approach to derive the adversarial risk bound.

Proposed method

- Define a mapping $T_h : \mathcal{Z} \to \mathcal{Z}$

$$z = (x, y) \to (x^*, y),$$

  where $x^* = \arg\max_{x' \in N(x)} l(h(x'), y)$.
- Let $P' = T_h \# P$, the pushforward of P by T_h, we have

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

- Therefore, the relationship between local worst-case risk and adversarial expected risk is as follows.

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

Local worst-case risk bound

- 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 $z' \in \mathcal{Z}$.
- Let $\lambda^+_{f,P_n} := \inf\{\lambda : \psi_{f,P_n}(\lambda) = 0\}$ where $\psi_{f,P_n}(\lambda) := \mathbb{E}_{P_n}(\sup_{z' \in \mathcal{Z}}\{f(z') - \lambda d_{\mathcal{Z}}(z, z') - f(z)\})$.
- Strong duality result for local worst-case risk by Gao & Kleywegt [2].
  For any upper semicontinuous function $f : \mathcal{Z} \to \mathbb{R}$ and for any $P \in \mathcal{P}_p(\mathcal{Z})$,

$$R_{\epsilon_{\mathcal{B}},1}(P, f) = \min_{\lambda \geq 0}\{\lambda\epsilon_{\mathcal{B}} + \mathbb{E}_P[\varphi_{\lambda,f}(z)]\},$$

  where $\varphi_{\lambda,f}(z) := \sup_{z' \in \mathcal{Z}}\{f(z') - \lambda \cdot d_{\mathcal{Z}}(z, z')\}$.

Lemma 1. Fix some $f \in \mathcal{F}$. Define $\bar{\lambda}$ via

$$\bar{\lambda} := \arg\min_{\lambda \geq 0}\{\lambda\epsilon_{\mathcal{B}} + \mathbb{E}_{P_n}[\varphi_{\lambda,f}(Z)]\}.$$

Then

$$\bar{\lambda} \in \begin{cases} [0, \frac{M}{\epsilon_{\mathcal{B}}}] & if\ \epsilon_{\mathcal{B}} \geq \frac{M}{\lambda^+_{f,P_n}} \\ [\lambda^-_{f,P_n}, \lambda^+_{f,P_n}] & if\ \epsilon_{\mathcal{B}} < \frac{M}{\lambda^+_{f,P_n}} \end{cases},$$

where $\lambda^-_{f,P_n} := \sup\{\lambda : \psi_{f,P_n}(\lambda) = \lambda^+_{f,P_n} \cdot \epsilon_{\mathcal{B}}\}$ if the set $\{\lambda : \psi_{f,P_n}(\lambda) = \lambda^+_{f,P_n} \cdot \epsilon_{\mathcal{B}}\}$ is nonempty, otherwise $\lambda^-_{f,P_n} := 0$.

Lemma 2. Under the assumptions, for any $f \in \mathcal{F}$, we have

$$R_{\epsilon_{\mathcal{B}},1}(P, f) - R_{\epsilon_{\mathcal{B}},1}(P_n, f) \leq \frac{24\mathfrak{C}(\mathcal{F})}{\sqrt{n}} + \frac{12\sqrt{\pi}}{\sqrt{n}}\Lambda_{\epsilon_{\mathcal{B}}} \cdot diam(Z) + M\sqrt{\frac{log(\frac{1}{\delta})}{2n}}$$

with probability at least $1 - \delta$.

Adversarial risk bounds

Theorem 1. Under the assumptions, for any $f \in \mathcal{F}$, we have

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

with probability at least $1 - \delta$.

EXAMPLE BOUNDS

Apply Theorem 1 to two commonly-used models: SVMs and neural networks.

Support vector machines

Corollary 1. In the SVMs setting, for any $f \in \mathcal{F}$, with probability at least $1 - \delta$,

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

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

Neural networks

Corollary 2. In the neural networks setting, for any $f \in \mathcal{F}$, with probability of $1 - \delta$, the following inequality holds

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

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

REMARKS

There are two data dependent terms $1/n\sum_{i=1}^n f(z_i)$ and $\lambda^+_{f,P_n}\epsilon_{\mathcal{B}}$ in our bound, suggesting the following optimization problem for adversarial robustness.

$$\min_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(z_i) + \lambda^+_{f,P_n}\epsilon_{\mathcal{B}}.$$

However, since λ^+_{f,P_n} is computationally intractable in practice, instead of using the exact λ^+_{f,P_n} in the objective function, we may consider the data-dependent upper bound for λ^+_{f,P_n} which is usually easier to obtain and a regularization parameter $\eta \in [0, 1]$ selected via grid search.

REFERENCES

[1] Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal.
Pac-learning in the presence of adversaries.
In *Advances in Neural Information Processing Systems*, pages 230–241, 2018.

[2] Rui Gao and Anton J Kleywegt.
Distributionally robust stochastic optimization with wasserstein distance.
arXiv preprint arXiv:1604.02199, 2016.

[3] Justin Khim and Po-Ling Loh.
Adversarial risk bounds for binary classification via function transformation.
arXiv preprint arXiv:1810.09519, 2018.

[4] Jaeho Lee and Maxim Raginsky.
Minimax statistical learning with wasserstein distances.
In *Advances in Neural Information Processing Systems*, pages 2687–2696, 2018.

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