The adversarial learning problem can be described as follows. Let $\mathcal{P}$ be the input distribution, $\mathcal{Q}$ be the output distribution, and $f : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathcal{X}$ be the classifier. The local worst-case risk of $f$, denoted by $R_f(\mathcal{P}, \mathcal{Q})$, is defined as

$$R_f(\mathcal{P}, \mathcal{Q}) = \sup_{h \in H} R_f(\mathcal{P}, \mathcal{Q}, h),$$

where $H$ is the hypothesis class. The adversarial expected risk over a distribution $\mathcal{P}$ is given by

$$R^e(\mathcal{P}, \mathcal{Q}) = \mathbb{E}_{z \sim \mathcal{P}} R_f(\mathcal{P}, \mathcal{Q}, h(z)).$$

The adversarial expected risk over a distribution $\mathcal{P}$ is equivalent to the standard expected risk under a new distribution $\tilde{\mathcal{P}}$. The adversarial learning problem is to find a classifier $f$ that minimizes the local worst-case risk.

Proposed method

- Define a mapping $\mathcal{T}_\mathcal{P} : \mathcal{Z} \rightarrow \mathcal{Z}$
- Let $\mathcal{P}^f = \mathcal{T}_\mathcal{P} \mathcal{P}$, the pushforward of $\mathcal{P}$ by $\mathcal{T}_\mathcal{P}$
- The relationship between the local worst-case risk and adversarial expected risk is as follows.

$$R_f(\mathcal{P}, \mathcal{Q}) \leq R_f(\mathcal{P}^f, \mathcal{Q}^f), \quad \forall h \in H.$$

Local worst-case risk bound

- Assume that for any function $f \in \mathcal{F}$ and any $z \in \mathcal{Z}$, there exists $h_f$ such that $f(z) - f(z') \leq h_f(z')$ for any $z'$
- Let $\lambda_{\max}(f) = \sup_{z \in \mathcal{Z}} |f(z) - f(z')|$
- Strong duality result for local worst-case risk by Gao and Kleywegt [2]

$$R_f(\mathcal{P}, \mathcal{Q}) = \min_{h \in H} \{\lambda \cdot \text{diam}(\mathcal{Z}) + R_f(\mathcal{P}, h(z))\}.$$

MINIMAX LEARNING

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

$$W_2(P, Q) = \inf_{\pi \in \Pi(P, Q)} \mathbb{E}[\|z - z'\|_2^2]^{1/2},$$

where $\Pi(P, Q)$ 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 $\mathcal{A}$ at $P$:

$$R_{\mathcal{A}}(P) = \sup_{h \in H} R_f(\mathcal{P}, \mathcal{Q}, h),$$

where $W_2(P, Q) = \|h(P) - h(Q)\|_1$ is the $\rho$-Wasserstein ball of radius $\rho > 0$ centered at $P$.

MAIN RESULTS

Motivation

- The adversarial expected risk over a distribution $\mathcal{P}$ is equivalent to the standard expected risk under a new distribution $\tilde{\mathcal{P}}$.
- We can show that all these new distributions locate within a $\rho$-Wasserstein ball centered at $\mathcal{P}$.

Adversarial risk bounds

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

$$R_f(\mathcal{P}, \mathcal{Q}) \leq \sum_{i=1}^n f(x_i) + \lambda_{\max}(f) \sqrt{\frac{2M}{N}} \sqrt{\text{diam}(\mathcal{Z})} + M \sqrt{\frac{|\mathcal{F}|}{2n}}$$

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 $1 - \delta$

$$R_f(\mathcal{P}, \mathcal{Q}) \leq \sum_{i=1}^n f(x_i) + \lambda_{\max}(f) \sqrt{\frac{2M}{N}} \sqrt{\text{diam}(\mathcal{Z})} + (2\rho + 1) + (1 + \rho) \sqrt{\frac{|\mathcal{F}|}{2n}}$$

Neural networks

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

$$R_f(\mathcal{P}, \mathcal{Q}) \leq \sum_{i=1}^n f(x_i) + \lambda_{\max}(f) \sqrt{\frac{2M}{N}} \sqrt{\text{diam}(\mathcal{Z})} + (2\rho + 1) + (1 + \rho) \sqrt{\frac{|\mathcal{F}|}{2n}}$$

REMARKS

There are two data dependent terms $\sum_{i=1}^n f(x_i)$ and $\lambda_{\max}(f)$ in our bound, suggesting the following optimization problem for adversarial robustness.

$$\min_{f \in \mathcal{F}} \sum_{i=1}^n f(x_i) + \lambda_{\max}(f) \sqrt{\frac{2M}{N}} \sqrt{\text{diam}(\mathcal{Z})}$$

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

**REFERENCES**