Minimax Statistical Learning

with Wasserstein distances

Jaeho Lee & Maxim Raginsky

“Minimax” learning

Goal: find the hypothesis minimizing the **worst-case risk**

$$\mathcal{R}_\varrho(P, f) := \sup_{Q \in \Gamma(P,\varrho)} \mathbf{E}_Q\left[f(Z)\right]$$

… $\Gamma(P, \varrho)$ is an **ambiguity set** representing uncertainty, e.g.

- domain drift (mismatch of training & test distribution)
- adversarial attack (enhancing robustness of hypothesis)

“Minimax” learning

Goal: find the hypothesis minimizing the **worst-case risk**

$$\mathcal{R}_\varrho(P, f) := \sup_{Q \in \Gamma(P,\varrho)} \mathbf{E}_Q\left[f(Z)\right]$$

... $\Gamma(P, \varrho)$ is an **ambiguity set** representing uncertainty, e.g.

- domain drift (mismatch of training & test distribution)
- adversarial attack (enhancing robustness of hypothesis)

Approach: find the hypothesis minimizing the **empirical risk**

$$\widehat{f} := \arg\min_{f \in \mathcal{F}} \mathcal{R}_\varrho(P_n, f)$$

"Minimax" learning

Goal: find the hypothesis minimizing the **worst-case risk**

$$\mathcal{R}_\varrho(P,f) := \sup_{Q \in \Gamma(P,\varrho)} \mathbf{E}_Q\left[f(Z)\right]$$

... $\Gamma(P,\varrho)$ is an **ambiguity set** representing uncertainty, e.g.

- domain drift (mismatch of training & test distribution)
- adversarial attack (enhancing robustness of hypothesis)

Approach: find the hypothesis minimizing the **empirical risk**

$$\widehat{f} := \arg\min_{f \in \mathcal{F}} \mathcal{R}_\varrho(P_n, f)$$

Question: what is the speed of convergence

$$\mathcal{R}_\varrho(P,\widehat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}_\varrho(P,f) \to 0 \text{ ?}$$

"Minimax" learning

Goal: find the hypothesis minimizing the **worst-case risk**

$$\mathcal{R}_\varrho(P, f) := \sup_{Q \in \Gamma(P,\varrho)} \mathbf{E}_Q\,[f(Z)]$$

... $\Gamma(P, \varrho)$ is an **ambiguity set** representing uncertainty, e.g.

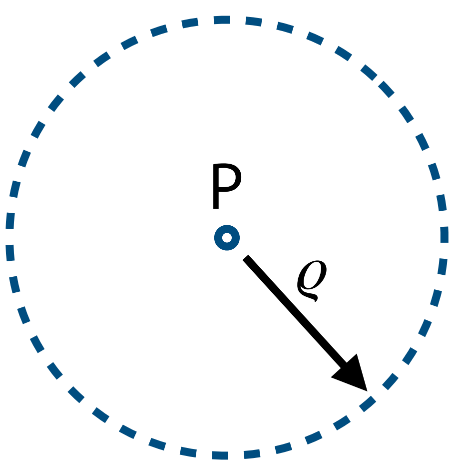
- domain drift (mismatch of training & test distribution)
- adversarial attack (enhancing robustness of hypothesis)

Approach: find the hypothesis minimizing the **empirical risk**

$$\widehat{f} := \arg\min_{f \in \mathcal{F}} \mathcal{R}_\varrho(P_n, f)$$

Question: what is the speed of convergence

$$\mathcal{R}_\varrho(P, \widehat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}_\varrho(P, f) \to 0 \ ?$$



Focus on **1-Wasserstein** ambiguity ball!

$$\Gamma(P, \varrho) = \{\ Q \ : \ W_1(P, Q) \leq \varrho\ \}$$

(we have results for p-Wasserstein balls, too! See Poster#86)

Taming the supremum

Main challenge is to handle the **supremum.**

Taming the supremum

Main challenge is to handle the **supremum.**

$$:= \sup_{z' \in \mathcal{Z}} \{ f(z') - \lambda \cdot (\|z' - z\| - \varrho) \}$$

Trick: (1) write down the **dual form**

$$\mathcal{R}_\varrho(P, f) = \inf_{\lambda \geq 0} \mathbf{E}_P \left[\psi_{\lambda, f}(Z) \right]$$

Taming the supremum

Main challenge is to handle the **supremum.**

Trick: (1) write down the **dual form**

$$:= \sup_{z' \in \mathcal{Z}} \{f(z') - \lambda \cdot (\|z' - z\| - \varrho)\}$$

$$\mathcal{R}_\varrho(P, f) = \inf_{\lambda \geq 0} \mathbf{E}_P \left[\psi_{\lambda, f}(Z)\right]$$

(2) empirical risk minimization is now **joint minimization**

$$\widehat{f} := \arg\min_{\lambda \geq 0, f \in \mathcal{F}} \mathbf{E}_P \left[\psi_{\lambda, f}(Z)\right]$$

Taming the supremum

Main challenge is to handle the **supremum.**

Trick: (1) write down the **dual form**

$$:= \sup_{z' \in \mathcal{Z}} \{f(z') - \lambda \cdot (\|z' - z\| - \varrho)\}$$

$$\mathcal{R}_\varrho(P, f) = \inf_{\lambda \geq 0} \mathbf{E}_P \left[\psi_{\lambda,f}(Z)\right]$$

(2) empirical risk minimization is now **joint minimization**

$$\widehat{f} := \arg\min_{\lambda \geq 0, f \in \mathcal{F}} \mathbf{E}_P \left[\psi_{\lambda,f}(Z)\right]$$

$$:= \Psi_{\Lambda,\mathcal{F}}$$

(3) gauge the complexity of the "**set of all possible** $\psi_{\lambda,f}$"

With high probability,

$$\mathcal{R}_\varrho(P, \widehat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}_\varrho(P, f) = \mathcal{O}\left(\frac{\text{complexity}(\Psi_{\Lambda,\mathcal{F}})}{\sqrt{n}}\right)$$

Result

Theorem) *Under mild assumptions, with high probability,*

$$\mathcal{R}_\varrho(P, \widehat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}_\varrho(P, f) = \mathcal{O}\left(\frac{\text{complexity}(\mathcal{F})}{\sqrt{n}}\right) + \mathcal{O}\left(\frac{1}{\varrho\sqrt{n}}\right)$$

- **vanishes to 0** as the sample size grows.
- does **not** require Lipschitz-type assumptions on f
- similar procedure could be applied for **any ambiguity set** with suitable dual form

Come to poster #86 for...

- applications to **domain adaptation**
- **complementary generalization bound** recovering classic bound as $\varrho \to 0$
- Results on **p-Wasserstein balls**