Exact Generalization Guarantees
For (Regularized) Wasserstein Distributionally Robust Models

Waïss Azizian, Franck Iutzeler, Jérôme Malick

NeurIPS 2023

Université Grenoble Alpes

MIAI Grenoble Alpes

Introduction

Basic task of Statistical Learning: learn from a finite number of samples from a true distribution

Goal of generalization guarantees:

relate risk w.r.t. samples to true risk

Our Theorem (Informal)

With high probability,

$$\textit{robust risk w.r.t. samples} \geq \textit{true risk}$$

- *For general classes of models*
- *Not overly pessimistic*
- *No curse of dimensionality*

Notations: standard and robust models in ML

- $f_\theta(\xi)$ the loss induced by a model parametrized by θ
- ξ uncertain variable (e.g., data point $\xi = (x, y)$)
- $\hat{P}_n$ empirical distribution with samples $\xi_1, \ldots, \xi_n$ of the true distribution P

$$\min_{\theta \in \Theta} \mathbb{E}_{\xi \sim \hat{P}_n}[f_\theta(\xi)] = \frac{1}{n} \sum_{i=1}^{n} f_\theta(\xi_i)$$

$\rightarrow$ Over-confident decisions and sensitive to distribution shifts

Wasserstein distributionally robust optimization (WDRO)

$$\min_{\theta \in \Theta} \sup_{Q : W_2(\hat{P}_n, Q) \leq \rho} \mathbb{E}_{\xi \sim Q}[f_\theta(\xi)]$$

where W_2 is the optimal transport cost between Q and Q'

- Robust version of ERM against distributions Q satisfying $W_2(\hat{P}_n, Q) \leq \rho$

Exact Generalization for WDRO

- Robust risk:

$$\widehat{\mathcal{R}}_{\rho^2}(f_\theta) = \sup_{Q \in \mathcal{P}(\Xi): W_2(\hat{P}_n, Q) \leq \rho} \mathbb{E}_{\xi \sim Q}[f_\theta(\xi)] .$$

- Direct generalization guarantees (Esfahani and Kuhn, 2018):

$$\text{if } W_2(\hat{P}_n, P) \leq \rho \quad \text{then} \quad \underbrace{\widehat{\mathcal{R}}_{\rho^2}(f_\theta)}_{\text{can compute from } \hat{P}_n} \geq \underbrace{\mathbb{E}_{\xi \sim P}[f_\theta(\xi)]}_{\text{cannot access}}$$

- Limitations:

→ It requires $\rho \propto 1/n^{1/d}$ where d is the dimension of ξ (Fournier and Guillin, 2015)

→ Not optimal: $\rho \propto 1/\sqrt{n}$ suffices asymptotically (Blanchet et al., 2022), in particular cases (Shafieezadeh-Abadeh et al., 2019) or with error terms (Gao, 2022).

Main Contribution: Exact Generalization for WDRO

Setting

- Θ, Ξ compact, f_θ smooth
- Covers many examples: logistic regression, smooth kernels, smooth neural networks,...

Theorem

For $\delta \in (0, 1)$, for ρ small enough and for any n, if

$$\rho \geq \mathcal{O}\left(\sqrt{\frac{\log 1/\delta}{n}}\right)$$

Generalization guarantee: w.p. $1 - \delta$, for all $\theta \in \Theta$,

$$\widehat{\mathcal{R}}_{\rho^2}(f_\theta) \geq \mathbb{E}_{\xi \sim P}\left[f_\theta(\xi)\right]$$

More in the paper

See the paper and come to the (virtual) poster for details, refinements and extensions :-)

Our results also:

- Allow for bigger ρ
- Capture distribution shifts
- Provide an upper-bound on the robust risk
- Extend to entropy-regularized formulation