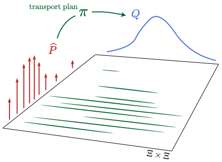
Universal generalization guarantees for Wasserstein distributionally robust models



Tam Le and Jérôme Malick

ICLR 2025 Spotlight

Standard machine learning framework:

- Family of loss functions $\mathcal{F}$
- $\xi_1, \ldots \xi_n$, i.i.d samples from (unknown) population distribution P

Then learn with **empirical risk minimization** (ERM)

$$\underset{f \in \mathcal{F}}{\text{Minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} f(\xi_i) = \mathbb{E}_{\xi \sim \widehat{P}}[f(\xi)]$$

[1] Wasserstein distributionally robust optimization: Theory and applications in machine learning, Kuhn et al. 2019

Standard machine learning framework:

- Family of loss functions $\mathcal{F}$
- $\xi_1, \ldots \xi_n$, i.i.d samples from (unknown) population distribution P

Then learn with **empirical risk minimization** (ERM)

$$\underset{f \in \mathcal{F}}{\text{Minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} f(\xi_i) = \mathbb{E}_{\xi \sim \widehat{P}}[f(\xi)]$$

Sensible to overfitting, data corruption, biases ...

to *distribution shifts* between training and new test data Q

$$\mathbb{E}_{\xi \sim \widehat{P}}[f(\xi)] \neq \mathbb{E}_{\xi \sim Q}[f(\xi)]$$

[1] Wasserstein distributionally robust optimization: Theory and applications in machine learning, Kuhn et al. 2019

Standard machine learning framework:

- Family of loss functions $\mathcal{F}$
- $\xi_1, \dots \xi_n$, i.i.d samples from (unknown) population distribution P

Then learn with **empirical risk minimization** (ERM)

$$\underset{f \in \mathcal{F}}{\text{Minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} f(\xi_i) = \mathbb{E}_{\xi \sim \widehat{P}}[f(\xi)]$$

Sensible to overfitting, data corruption, biases ...

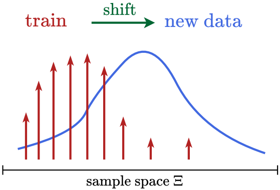
to *distribution shifts* between training and new test data Q

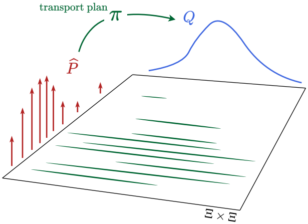
$$\mathbb{E}_{\xi \sim \widehat{P}}[f(\xi)] \neq \mathbb{E}_{\xi \sim Q}[f(\xi)]$$

Our focus: Wasserstein distributionally robust optimization (WDRO)[1]

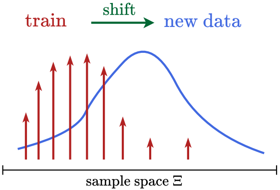
[1] Wasserstein distributionally robust optimization: Theory and applications in machine learning, Kuhn et al. 2019

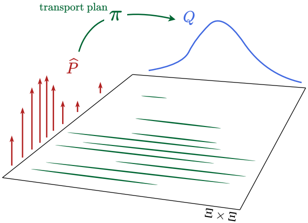
WDRO framework: Distribution shift as transport of distribution





WDRO framework: Distribution shift as transport of distribution





Optimal transport cost. (Wasserstein distance)

$$W_c(\widehat{P}, Q) = \inf \left\{ \mathbb{E}_{(\xi,\zeta)\sim\pi}[c(\xi,\zeta)] \;:\; \substack{\pi \text{ distribution} \\ [\pi]_1 = \widehat{P}, [\pi]_2 = Q} \right\}$$

Ambiguity set around $\widehat{P}$, with **radius** ρ

$$\left\{ Q \in \mathcal{P}(\Xi) \;:\; W_c(\widehat{P}, Q) \leq \rho \right\}$$

Wasserstein robustness and generalization

Instead of ERM, learn with WDRO [Esfahani and Kuhn 2018; Zhao and Guan 2018 . . .]

$$\operatorname*{Minimize}_{f \in \mathcal{F}} \quad \max_{Q : W_c(\widehat{P}, Q) \leq \rho} \quad \mathbb{E}_{\xi \sim Q}[f(\xi)] := \widehat{R}_\rho(f)$$

Wasserstein robustness and generalization

Instead of ERM, learn with WDRO [Esfahani and Kuhn 2018; Zhao and Guan 2018 ...]

$$\underset{f\in\mathcal{F}}{\text{Minimize}} \quad \max_{Q:W_c(\widehat{P},Q)\leq\rho} \mathbb{E}_{\xi\sim Q}[f(\xi)] := \widehat{R}_\rho(f)$$

Exact generalization bound of WDRO: If ρ is high enough, then

$$P \in \{Q \ : \ W_c(\widehat{P},Q) \leq \rho\},$$

leading to the **exact bound**

$$\boxed{\widehat{R}_\rho(f) = \max_{Q:W_c(\widehat{P},Q)\leq\rho} \mathbb{E}_Q[f] \qquad \geq \qquad \mathbb{E}_P[f]}$$

Wasserstein robustness and generalization

Instead of ERM, learn with WDRO [Esfahani and Kuhn 2018; Zhao and Guan 2018 ...]

$$\underset{f \in \mathcal{F}}{\text{Minimize}} \quad \max_{Q : W_c(\widehat{P}, Q) \leq \rho} \quad \mathbb{E}_{\xi \sim Q}[f(\xi)] := \widehat{R}_\rho(f)$$

Exact generalization bound of WDRO: If ρ is high enough, then

$$P \in \{Q \ : \ W_c(\widehat{P}, Q) \leq \rho\},$$

leading to the **exact bound**

$$\boxed{\widehat{R}_\rho(f) = \max_{Q : W_c(\widehat{P}, Q) \leq \rho} \mathbb{E}_Q[f] \qquad \geq \qquad \mathbb{E}_P[f]}$$

Choice of ρ?

Out-of-the-box guarantees and dimension curse

Natural choice: an estimate of $\boxed{W_c(\widehat{P}, P) \underset{n\to\infty}{\longrightarrow} 0}$.

Fournier and Guillin 2015: with high probability,

$$W_c(\widehat{P}, P) \leq \frac{C}{n^{\frac{1}{d}}} \quad \text{data dimension!}$$

Can we do better, in the specific context of WDRO?

Toward dimension-free rates

Generalization guarantees with $\boxed{\rho > O(1/\sqrt{n})}$ have been studied a lot, But only with approximate bounds or specific settings.

- *Azizian, Iutzeler, Malick 2023*: **exact bound** but smooth losses, square distance cost, growth assumptions...
- *An & Gao 2021*: smooth losses and/or approximate terms
- *Shapiro & Blanchet 2021*: smooth losses, asymptotics
- *Shafieezadeh-Abadeh et al. 2019*: linear models ...
- ...

Toward dimension-free rates

Generalization guarantees with $\boxed{\rho > O(1/\sqrt{n})}$ have been studied a lot, But only with approximate bounds or specific settings.

- *Azizian, Iutzeler, Malick 2023*: **exact bound** but smooth losses, square distance cost, growth assumptions...
- *An & Gao 2021*: smooth losses and/or approximate terms
- *Shapiro & Blanchet 2021*: smooth losses, asymptotics
- *Shafieezadeh-Abadeh et al. 2019*: linear models ...
- ...

Generalization guarantees for Wasserstein robust models [L. & Malick 2025]

Ξ compact, c and $f \in \mathcal{F}$ are continuous [...]. For $\rho \geq \frac{K}{\sqrt{n}}$ with probability $1 - \delta$,

$$\boxed{\widehat{R}_\rho(f) \geq \mathbb{E}_P[f]} \qquad \forall f \in \mathcal{F}$$

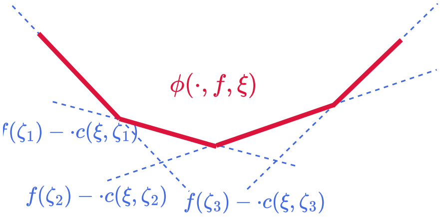
K depends on δ, $c(\cdot,\cdot)$, the (compact) parameter and sample spaces, P and $\mathcal{F}$.

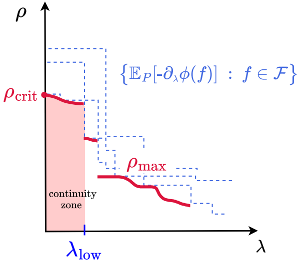
Nonsmooth analysis based proof

Duality formula [Blanchet and Murthy 2019; Gao and Kleywegt 2016]

$$\widehat{R}_\rho(f) = \inf_{\lambda \geq 0} \lambda\rho + \mathbb{E}_{\widehat{P}}[\phi(\lambda, f)]$$

Nonsmooth analysis to derive a dual lower bound.





Main takeaways

- **Exact generalization bound** for Wasserstein distributionally robust optimization.
- **Wide setting** thanks to nonsmooth analysis tools
- General proof scheme; extension to **regularized versions**

$$\phi_{\epsilon,\tau}(\lambda, f, \xi) = (\epsilon + \lambda\tau) \log \mathbb{E}_{\pi_0(\cdot|\xi)} \left[\exp\left(\frac{f - \lambda c(\xi, \cdot)}{\epsilon + \lambda\tau} \right) \right]$$

Thank you!