Exact Generalization Guarantees for Wasserstein Distributionally Robust Models

Waïss Azizian[1,2], Franck Iutzeler[1,2], Jérôme Malick[1,2,3]

[1] Univ. Grenoble Alpes, [2] LJK, [3] CNRS,

Contributions

Generalization bounds for WDRO
- Robust objective $\implies$ exact upper-bound on the true risk w.h.p.
- No curse of dimensionality and for general classes of models
- Cover distribution shifts at testing

Distributionally Robust Optimization (DRO)

Empirical Risk Minimization (ERM):
- θ model parameter, ξ uncertain variable (e.g., data point $\xi = (x, y)$)
- $f_\theta(\xi)$ the loss induced by a model parametrized by θ
- $\xi_1, \dots, \xi_n$ samples of the true distribution P

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

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

Distributionally Robust Optimization (DRO) to mitigate these issues

$$\min_{\theta\in\Theta} \sup_{\mathrm{Q}\in\mathcal{U}(\mathrm{P}_n)} \mathbb{E}_{\xi\sim\mathrm{Q}}[f_\theta(\xi)]$$

- $\mathcal{U}(\mathrm{P}_n)$ neighborhood of P_n in probability space.

Wasserstein Distributionally Robust Optimization

A popular choice

$$\mathcal{U}(\mathrm{P}_n) = \{\mathrm{Q}\in\mathcal{P}(\Xi) : W_2(\mathrm{P}_n, \mathrm{Q}) \le \rho\}$$

with the Wasserstein distance

$$W_2^2(\mathrm{Q}, \mathrm{Q}') := \inf_{\pi\in\mathcal{P}(\Xi\times\Xi), \pi_1=\mathrm{Q}, \pi_2=\mathrm{Q}'} \mathbb{E}_{(\xi,\zeta)\sim\pi}\left[\frac{1}{2}\|\xi-\zeta\|^2\right]$$

$\mathcal{P}(\Xi\times\Xi)$ probability distributions on $\Xi\times\Xi$, and π_1 and π_2 the marginals of π.

Wasserstein Distributionally Robust Optimization (WDRO)

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

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

if P satisfies $W_2(\mathrm{P}_n, \mathrm{P}) \le \rho$, then

$$\underbrace{\widehat{\mathcal{R}}_{\rho^2}(f_\theta)}_{\text{can compute \& optimize}} \ge \underbrace{\mathbb{E}_{\xi\sim\mathrm{P}}[f_\theta(\xi)]}_{\text{cannot access}}\,.$$

$\to$ But it requires $\rho \propto 1/n^{1/d}$ where $\xi\in\mathbb{R}^d$ (Fournier and Guillin, 2015)

$\to$ 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).

Generalization Guarantees

Setting
- $(\theta, \xi) \in \Theta\times\Xi \mapsto f_\theta(\xi)$ C^2 with Θ and $\Xi\subset\mathbb{R}^d$ compact
- P supported on the interior of Ξ, and for all θ, $\mathrm{P}(\nabla_\xi f_\theta(\xi) = 0) < 1$

Theorem 1

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

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

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

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

Distribution shifts: w.p. $1-\delta$, for all $\theta\in\Theta$ and Q s.t.

$$W_2^2(\mathrm{P}, \mathrm{Q}) \le \rho\left(\rho - \mathcal{O}\left(\sqrt{\frac{\log 1/\delta}{n}}\right)\right) \quad \textit{it holds} \quad \widehat{\mathcal{R}}_{\rho^2}(f_\theta) \ge \mathbb{E}_{\xi\sim\mathrm{Q}}[f_\theta(\xi)]$$

Additional assumptions
- f_θ grows quadratically near its maximums uniformly in $\theta\in\Theta$.
- $\{f_\theta : \theta\in\Theta\}$ is relatively compact for $D(f, g) := \|f-g\|_\infty + D_{\mathrm{H}}(\arg\max f, \arg\max g)$.

Theorem 2

The conclusions of Theorem 2 hold for all ρ satisfying

$$\mathcal{O}\left(\sqrt{\frac{\log 1/\delta}{n}}\right) \le \rho \le \frac{\rho_c}{2} - \mathcal{O}\left(\sqrt{\frac{\log 1/\delta}{n}}\right)$$

where

$$\rho_c^2 = \inf_{\theta\in\Theta} \mathbb{E}_{\xi\sim\mathrm{P}}\left[\tfrac{1}{2} d(\xi, \arg\max f_\theta)^2\right]$$

The critical radius ρ_c:

If $\rho \gg \rho_c$, there is some $\theta\in\Theta$ s.t.

$$\rho^2 \gg \mathbb{E}_{\xi\sim\mathrm{P}}\left[\tfrac{1}{2} d(\xi, \arg\max f_\theta)^2\right]$$

and so there exists Q supported on $\arg\max f_\theta$ that satisfies $W_2(\mathrm{P}, \mathrm{Q}) \ll \rho$. Hence, the RHS of *(Distribution shifts)* is equal to $\max_\Xi f_\theta$.

Idea of proof

Strong duality:

$$\widehat{\mathcal{R}}_{\rho^2}(f_\theta) = \inf_{\lambda\ge 0} \lambda\rho^2 + \mathbb{E}_{\xi\sim\mathrm{P}_n}\left[\sup_{\zeta\in\Xi}\left\{f_\theta(\zeta) - \tfrac{\lambda}{2}\|\zeta-\xi\|_2^2\right\}\right]$$

Concentration:

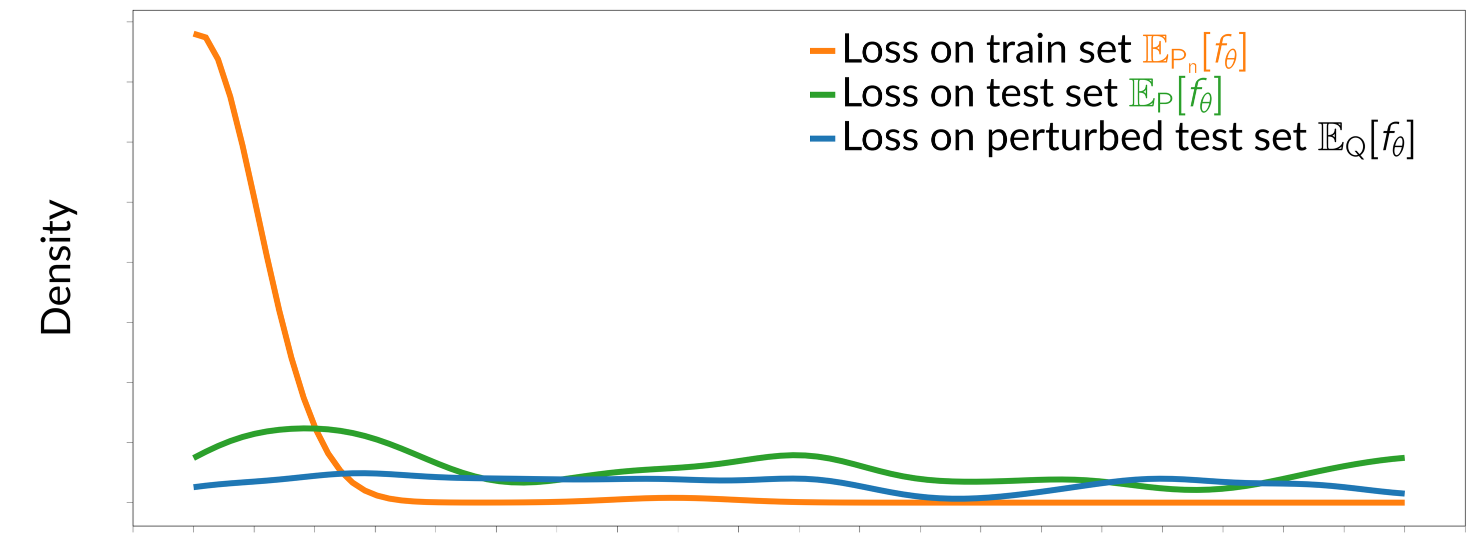
$\lambda^{-1}\mathbb{E}_{\xi\sim\mathrm{P}_n}\left[\sup_{\zeta\in\Xi}\left\{f_\theta(\zeta) - \tfrac{\lambda}{2}\|\zeta-\xi\|_2^2\right\}\right]$ concentrates with error $\mathcal{O}\left(\frac{1}{\lambda}\sqrt{\frac{\log 1/\delta}{n}}\right)$

Bound on dual multiplier:

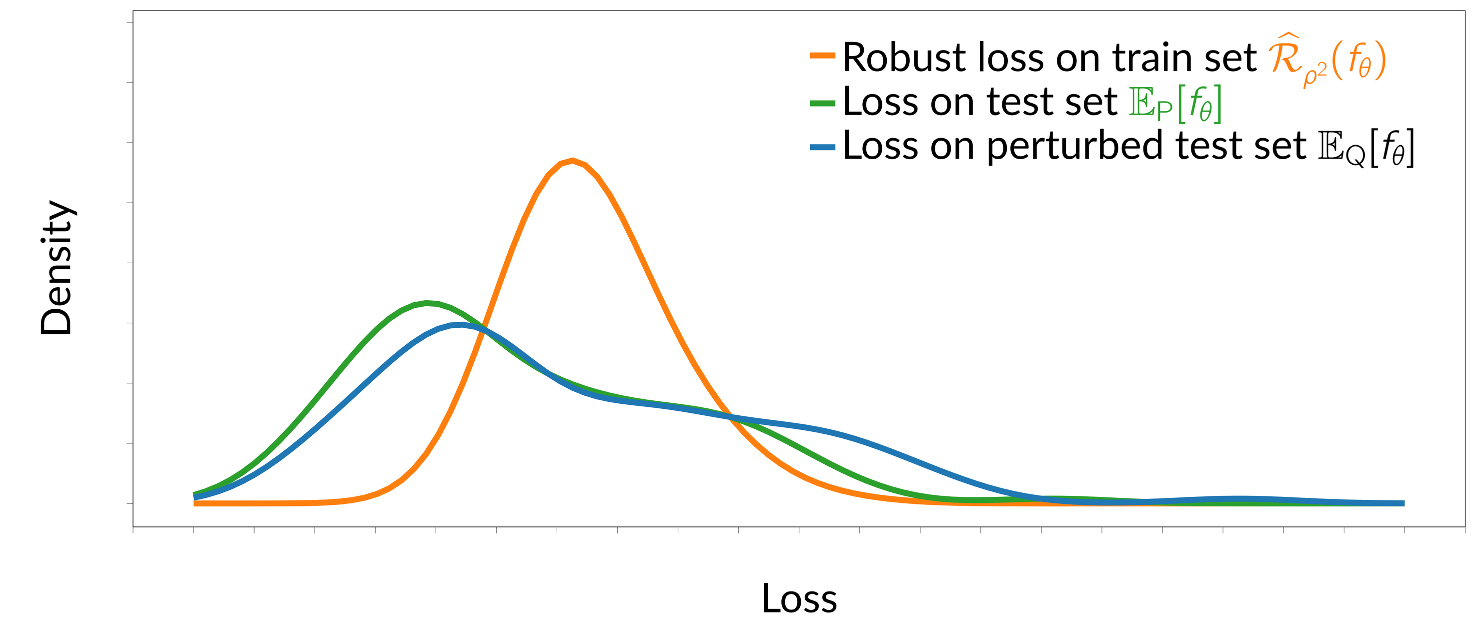
λ bounded away from 0, as $\propto 1/\rho$, for admissible ρ.

Illustration: Logistic Regression

Smoothed histogram of losses with the ERM predictor



Smoothed histogram of losses with the WDRO predictor



Extension: entropy-regularized WDRO

Inspired by OT, regularized WDRO (Wang et al., Azizian et al., 2023)

$$\sup\left\{\mathbb{E}_{\xi\sim\pi_2}[f(\xi)] - \varepsilon\,\mathrm{KL}(\pi\,|\pi_\sigma^n) : \pi,\ \pi_1 = \mathrm{P}_n,\ \mathbb{E}_{(\xi,\zeta)\sim\pi}\left[\tfrac{1}{2}\|\xi-\zeta\|^2\right] \le \rho^2\right\}$$
$$= \inf_{\lambda\ge 0} \lambda\rho^2 + \mathbb{E}_{\xi\sim\mathrm{P}_n}\left[\log\left(\mathbb{E}_{\zeta\sim\pi_\sigma(\cdot|\xi)}\left[e^{\frac{f_\theta(\zeta)-\lambda\|\xi-\zeta\|_2^2/2}{\varepsilon}}\right]\right)\right]$$

with prior $\pi_\sigma^n \propto \mathrm{P}_n(\,d\xi)e^{-\frac{\|\xi-\zeta\|_2^2}{2\sigma^2}}\,d\zeta$

$\to$ Similar results as Theorem 1 hold for regularized risks.

References

W. Azizian, F. Iutzeler, and J. Malick. Exact generalization guarantees for (regularized) Wasserstein distributionally robust models. *arXiv 2023*.

W. Azizian, F. Iutzeler, and J. Malick. Regularization for Wasserstein distributionally robust optimization. *ESAIM: COCV*, 2023.

J. Blanchet, K. Murthy, and N. Si. Confidence regions in wasserstein distributionally robust estimation. *Biometrika*, 2022.

P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric. *Mathematical Programming*, 2018.

N. Fournier and A. Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 2015.

R. Gao. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 2022.

S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani. Regularization via Mass Transportation. *JMLR*, 2019.

J. Wang, R. Gao, and Y. Xie. Sinkhorn distributionally robust optimization. *arXiv 2021*.