Knowledge-Guided Wasserstein Distributionally Robust Optimization

Zitao Wang[1] Ziyuan Wang[2] Molei Liu[3] Nian Si[4]

[1]Department of Statistics, Columbia University
[2]Department of Industrial Engineering and Management Sciences, Northwestern University
[3]Department of Biostatistics, Peking University Health Science Center
[3]Beijing International Center for Mathematical Research, Peking University
[4]Department of Industrial Engineering and Decision Analytics, HKUST

June 14, 2025

Wasserstein DRO

Given a cost function $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to \overline{\mathbb{R}}_+$, the *Wasserstein Transport Cost* between two measures $\mathbb{P}$ and $\mathbb{Q}$ is

$$\mathcal{D}_c(\mathbb{P}, \mathbb{Q}) \coloneqq \inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \int c(U, V)\, d\pi, \quad \text{s.t. } U \sim \mathbb{P}, V \sim \mathbb{Q}.$$

Wasserstein DRO

Given a cost function $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to \overline{\mathbb{R}}_+$, the *Wasserstein Transport Cost* between two measures $\mathbb{P}$ and $\mathbb{Q}$ is

$$\mathcal{D}_c(\mathbb{P}, \mathbb{Q}) := \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int c(U, V)\, d\pi, \text{ s.t. } U \sim \mathbb{P}, V \sim \mathbb{Q}.$$

The *Wasserstein distributionally robust optimization* (WDRO) framework solves the minimax stochastic program:

$$\inf_\beta \sup_{\mathbb{P} \in \mathcal{B}_\delta(\mathbb{P}_N^*; c)} \mathbb{E}_{\mathbb{P}}[\ell(X, Y; \beta)]$$

Wasserstein DRO

Given a cost function $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to \overline{\mathbb{R}}_+$, the *Wasserstein Transport Cost* between two measures $\mathbb{P}$ and $\mathbb{Q}$ is

$$\mathcal{D}_c(\mathbb{P}, \mathbb{Q}) := \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int c(U, V)\, d\pi, \text{ s.t. } U \sim \mathbb{P}, V \sim \mathbb{Q}.$$

The *Wasserstein distributionally robust optimization* (WDRO) framework solves the minimax stochastic program:

$$\inf_{\beta} \sup_{\mathbb{P} \in \mathcal{B}_\delta(\mathbb{P}_N^*; c)} \mathbb{E}_{\mathbb{P}}[\ell(X, Y; \beta)]$$

where $\mathcal{B}_\delta(\mathbb{P}_N^*; c)$ is an *ambiguity set* of candidate measures for $\mathbb{P}^*$, constructed as a δ-ball around the empirical measure $\mathbb{P}_N^*$:

$$\mathcal{B}_\delta(\mathbb{P}_N^*; c) := \{\mathbb{P} \in \mathcal{P}(\mathbb{R}^{d+1}) | \mathcal{D}_c(\mathbb{P}, \mathbb{P}_N^*) \leq \delta\}$$

WDRO Linear Regression

It is shown that using the cost function

$$c_{q,0}\big((x,y),(u,v)\big) = \|x-u\|_q^2 + \infty \cdot |y-v|$$

equates WDRO linear regression with p-norm regularization on RMSE.

WDRO Linear Regression

It is shown that using the cost function

$$c_{q,0}\big((x,y),(u,v)\big) = \|x-u\|_q^2 + \infty \cdot |y-v|,$$

equates WDRO linear regression with p-norm regularization on RMSE.

Theorem 1 ((Blanchet, Kang, & Murthy, 2019, Theorem 1))
For any $q \in [1,\infty]$ we have

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{P}:\mathcal{B}_\delta(c_{q,0})} \mathbb{E}_{\mathbb{P}}\left[(Y-\beta^\intercal X)^2\right] = \inf_{\beta \in \mathbb{R}^d} \left\{ \sqrt{\mathrm{MSE}_N(\beta)} + \sqrt{\delta}\|\beta\|_p \right\}^2,$$

with (p,q) such that $p^{-1}+q^{-1}=1$.

The Knowledge-Guided Cost

With a prior knowledge θ, we control the extent of perturbation along the direction of θ. The knowledge-guided cost function associated to the q-norm is

$$c_{q,\lambda}(x-u) = \|x-u\|_q^2 + \lambda \cdot (\theta^\intercal (x-u))^2 + \infty \cdot |y-v|.$$

The Knowledge-Guided Cost

With a prior knowledge θ, we control the extent of perturbation along the direction of θ. The knowledge-guided cost function associated to the q-norm is

$$c_{q,\lambda}(x-u) = \|x-u\|_q^2 + \lambda \cdot (\theta^\intercal (x-u))^2 + \infty \cdot |y-v|.$$

We call it

1. Strong-transferring if $\lambda = \infty$,
2. Weak-transferring if $\lambda < \infty$.

Proposition 1

We have the following upper bound for strong-transferring:

$$\inf_{\beta} \sup_{\mathbb{P} \in \mathcal{B}_\delta(c_{q,\infty})} \mathbb{E}_{\mathbb{P}} \left[(Y - \beta^\intercal X)^2 \right] \leq \inf_{\alpha \in \mathbb{R}} \mathbb{E}_{\mathbb{P}_N^*} \left[(Y - (\alpha\theta)^\intercal X)^2 \right]$$

Tractable Reformulation of Knowledge-Guided WDRO

With data $\mathbf{y} \in \mathbb{R}^N$ and $\mathbf{X} \in \mathbb{R}^{N \times d}$, and an accessible learner $\theta \in \mathbb{R}^d$, we study the strong-transferring estimator that solves

$$\underset{\beta}{\operatorname{argmin}} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2 + \delta \cdot \min_{\kappa \in \mathbb{R}} \|\beta - \kappa\theta\|_p \right\},$$

Tractable Reformulation of Knowledge-Guided WDRO

With data $\mathbf{y} \in \mathbb{R}^N$ and $\mathbf{X} \in \mathbb{R}^{N \times d}$, and an accessible learner $\theta \in \mathbb{R}^d$, we study the strong-transferring estimator that solves

$$\underset{\beta}{\operatorname{argmin}} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2 + \delta \cdot \min_{\kappa \in \mathbb{R}} \|\beta - \kappa\theta\|_p \right\},$$

and for $p = q = 2$, its weak-transferring counterpart:

$$\underset{\beta}{\operatorname{argmin}} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2 + \delta \cdot \|\beta\|_{\Psi_\lambda} \right\},$$

with $\Psi_\lambda = \mathbb{I}_d - \frac{1}{\|\theta\|_2^2 + \lambda^{-1}} \theta\theta^\intercal$.

Tractable Reformulation of Knowledge-Guided WDRO

With data $\mathbf{y} \in \mathbb{R}^N$ and $\mathbf{X} \in \mathbb{R}^{N\times d}$, and an accessible learner $\theta \in \mathbb{R}^d$, we study the strong-transferring estimator that solves

$$\operatorname*{argmin}_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2 + \delta \min_{\kappa\in\mathbb{R}} \|\beta - \kappa\theta\|_p \right\},$$

and for $p = q = 2$, its weak-transferring counterpart:

$$\operatorname*{argmin}_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2 + \delta\|\beta\|_{\Psi_\lambda} \right\},$$

with $\Psi_\lambda = \mathbb{I}_d - \dfrac{1}{\|\theta\|_2^2 + \lambda^{-1}}\theta\theta^\intercal$.

The hyperparameters are

1. $\delta \in [0, \infty]$ controls the regularization strength;
2. $\lambda \in [0, \infty]$ measures our confidence in the prior knowledge θ.

Feasibility Set

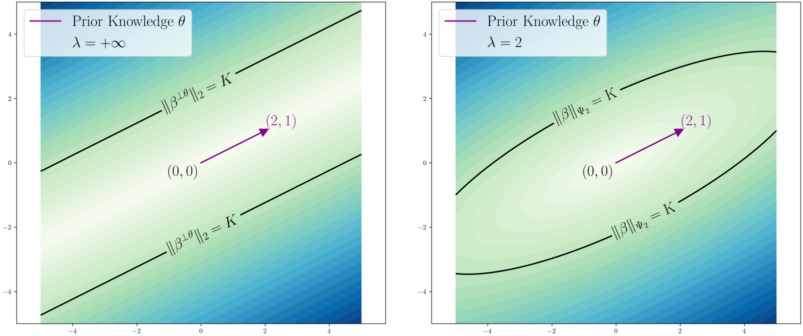

Figure 1: Feasibility Sets: Left – Strong-Transferring Regularizer ($p = q = 2$, $\lambda = \infty$); Right – Weak-Transferring Regularizer ($p = q = 2$, $\lambda = 2$).

Thank you!

Bibliography I

Blanchet, J., Kang, Y., & Murthy, K. (2019). Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, *56*(3), 830-857.