Distributionally Robust Optimization

Runyu Tang

March 2, 2023

Robust for deterministic constraints:
$g(x,\xi) \le 0, \forall \xi \to \sup_{\xi} g(x,\xi) \le 0.$
Robust for stochastic constraints:
$\mathbb{E}^{\mathbb{P}}[g(x,\xi)] \le 0, \forall \mathbb{P} \to \sup_{\mathbb{P}} \mathbb{E}^{\mathbb{P}}[g(x,\xi)] \le 0.$

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}^{\mathbb{P}}\left[f(\boldsymbol{x},\boldsymbol{\xi})\right] \\ \text{s.t.} \quad & g_j(\boldsymbol{x},\boldsymbol{\xi}) \le 0, \quad \text{almost surely } \forall j, \mathbb{P}\in\mathcal{F}, \\ & \boldsymbol{x}\in\mathcal{X}, \end{aligned}$$

where $\mathcal{F}$ is the ambiguity set

Commonly used ambiguity sets:

- Moment-based ambiguity set
- Distance-based ambiguity set
  - ϕ-divergence
  - Wasserstein distance

Moment-based ambiguity set:

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P} \left| \begin{array}{l} \mathbb{P}\left(\tilde{\boldsymbol{\xi}} \in \Xi\right) = 1, \\ (\mathbb{E}[\tilde{\boldsymbol{\xi}}] - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\mathbb{E}[\tilde{\boldsymbol{\xi}}] - \boldsymbol{\mu}_0) \leq \gamma_1, \\ \mathbb{E}_{\mathbb{P}}[(\tilde{\boldsymbol{\xi}} - \boldsymbol{\mu}_0)(\tilde{\boldsymbol{\xi}} - \boldsymbol{\mu}_0)^T] \preceq \gamma_2 \boldsymbol{\Sigma}_0. \end{array} \right. \right\}$$

- Ξ: nonempty support, closed and convex,
- $\boldsymbol{\mu}_0 \in \mathbb{R}^n$: first moment,
- $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{n \times n}$: covariance matrix.

$$\mathcal{F}(\Xi, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \gamma_1, \gamma_2)$$

Delage and Ye (2010), Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems, *Operations Research.*

Moment-based ambiguity set:
Inner moment problem:

$$
\begin{aligned}
\Psi(\boldsymbol{x};\gamma_1,\gamma_2) = \max_{\mathbb{P}\in\mathcal{F}} \quad & \mathbb{E}^{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] \\
\Leftrightarrow \max_{F} \quad & \int f(\boldsymbol{x},\boldsymbol{\xi})\mathrm{d}F(\boldsymbol{\xi}) \\
\text{s.t.} \quad & \int \mathrm{d}F(\boldsymbol{\xi}) = 1, \\
& \int (\boldsymbol{\xi}-\boldsymbol{\mu}_0)(\boldsymbol{\xi}-\boldsymbol{\mu}_0)^\top \mathrm{d}F(\boldsymbol{\xi}) \preceq \gamma_2\boldsymbol{\Sigma}_0, \\
& \int \begin{bmatrix} \boldsymbol{\Sigma}_0 & (\boldsymbol{\xi}-\boldsymbol{\mu}_0) \\ (\boldsymbol{\xi}-\boldsymbol{\mu}_0) & \gamma_1 \end{bmatrix} \mathrm{d}F(\boldsymbol{\xi}) \succeq 0
\end{aligned}
$$

Moment-based ambiguity set:

Dual of the inner moment problem:

$$
\begin{aligned}
\min_{r,\mathbf{Q},\mathbf{P},\mathbf{p},s} \quad & \left(\gamma_2 \boldsymbol{\Sigma}_0 - \boldsymbol{\mu}_0 \boldsymbol{\mu}_0^\top\right) \cdot \mathbf{Q} + r + (\boldsymbol{\Sigma}_0 \cdot \mathbf{P}) - 2\boldsymbol{\mu}_0^\top \mathbf{p} + \gamma_1 s \\
\text{s.t.} \quad & \boldsymbol{\xi}^\top \mathbf{Q} \boldsymbol{\xi} - 2\boldsymbol{\xi}^\top (\mathbf{p} + \mathbf{Q}\boldsymbol{\mu}_0) + r - f(\mathbf{x}, \boldsymbol{\xi}) \geqslant 0, \forall \boldsymbol{\xi}, \\
& \mathbf{Q} \succeq 0, \\
& \begin{bmatrix} \mathbf{P} & \mathbf{p} \\ \mathbf{p}^\top & s \end{bmatrix} \succeq 0,
\end{aligned}
$$

Moment-based ambiguity set:
Inner moment problem:

$$
\begin{aligned}
&\max_{\mathbb{P}\in\mathcal{F}} && \mathbb{E}^{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] \\
\Leftrightarrow &\min_{\boldsymbol{Q},\boldsymbol{q},r,t} && r+t \\
&\text{s.t.} && r \geq f(\boldsymbol{x},\boldsymbol{\xi}) - \boldsymbol{\xi}^\top \boldsymbol{Q}\boldsymbol{\xi} - \boldsymbol{\xi}^\top \boldsymbol{q}, \forall \boldsymbol{\xi} \\
& && t \geq (\gamma_2 \boldsymbol{\Sigma}_0 + \boldsymbol{\mu}\boldsymbol{\mu}^\top)\cdot \boldsymbol{Q} + \boldsymbol{\mu}^\top \boldsymbol{q} + \sqrt{\gamma_1}\|\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{q}+2\boldsymbol{Q}\boldsymbol{\mu})\|, \\
& && \boldsymbol{Q} \succeq 0,
\end{aligned}
$$

which can be solved to any precision ϵ in time polynomial in $\log(1/\epsilon)$ and the size of the problem (under some assumptions: $f(\boldsymbol{x},\boldsymbol{\xi})$ is concave in $\boldsymbol{\xi}$).

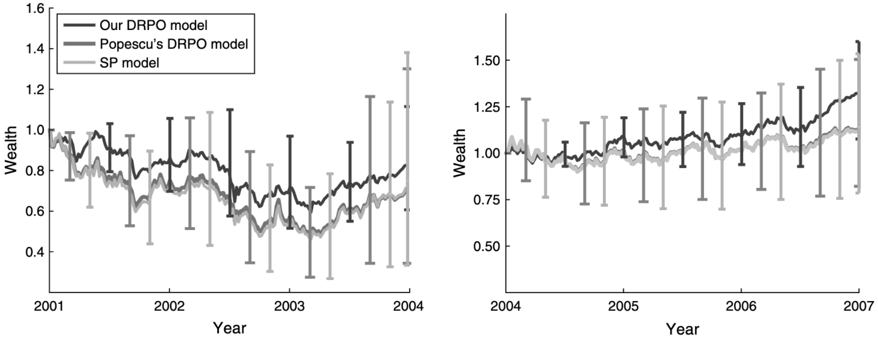
Under some assumptions (f concave in ξ and convex in x, support set is closed and bounded, ...), given a set of $\{\boldsymbol{\xi}_i\}_{i=1}^M$ of M samples, for any $\delta > 0$, let
$\hat{\boldsymbol{\mu}} = \frac{1}{M}\sum_{i=1}^M \boldsymbol{\xi}_i$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{i=1}^M (\boldsymbol{\xi}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{\xi}_i - \hat{\boldsymbol{\mu}})^\top$,
$\bar{\gamma}_1 = \frac{\bar{\beta}(\bar{\delta}/2)}{1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2)}$, $\bar{\gamma}_2 = \frac{1+\bar{\beta}(\bar{\delta}/2)}{1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2)}$,
where $\bar{\alpha}(\bar{\delta}/4) = O(1/\sqrt{M})$, $\bar{\beta}(\bar{\delta}/2) = O(1/M)$.

Then, if M is large enough, with probability greater than $1-\delta$ over the choice of $\{\boldsymbol{\xi}_i\}_{i=1}^M$, we have that any optimal solution of the DRSP formed using these samples will satisfy the constraint:

$$\mathbb{E}[f(\boldsymbol{x}^*, \boldsymbol{\xi})] \leq \Psi(\boldsymbol{x}^*; \bar{\gamma}_1, \bar{\gamma}_2),$$

where $\mathbb{E}$ is the expectation w.r.t the true distribution of $\boldsymbol{\xi}$.

Figure 1. Comparison of wealth evolution in 300 experiments conducted over the years 2001–2007.



Note. For each approach, the figures indicate periodically the 10th and 90th percentiles of the distribution of accumulated wealth.

At any given day of the experiment, the algorithms are allowed to use a period of 30 days from the most recent history to assign the portfolio. In Popescu(2007), the mean and covariance matrix of the distribution is assumed to be equal to the empirical estimates measured on the last 30 days.

4.2. Generalized-Moment Ambiguity Set

Wiesemann et al. (2014) formally introduce the following generalized-moment ambiguity set that is based on a convex function $\boldsymbol{\phi}: \mathbb{R}^{I_u} \mapsto \mathbb{R}^{I_v}$:

$$\mathcal{G} = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathbb{R}^{I_u}) \;\middle|\; \begin{array}{l} \tilde{\boldsymbol{u}} \sim \mathbb{P} \\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{u}}] \in \mathcal{Q} \\ \mathbb{E}_{\mathbb{P}}[\boldsymbol{\phi}(\tilde{\boldsymbol{u}})] \le \boldsymbol{\sigma} \\ \mathbb{P}[\tilde{\boldsymbol{u}} \in \mathcal{U}] = 1 \end{array} \right\}.$$

ity, among others. Based on the lifting and projection theorem (Wiesemann et al. 2014, theorem 5), it holds that $\Pi_{\tilde{\boldsymbol{u}}}\mathcal{F} = \mathcal{G}$, where

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathbb{R}^{I_u+I_v} \times \{1\}) \;\middle|\; \begin{array}{l} ((\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}), \tilde{s}) \sim \mathbb{P} \\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{u}} \mid \tilde{s} = 1] \in \mathcal{Q} \\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{v}} \mid \tilde{s} = 1] \le \boldsymbol{\sigma} \\ \mathbb{P}[(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) \in \mathcal{Z} \mid \tilde{s} = 1] = 1 \\ \mathbb{P}[\tilde{s} = 1] = 1 \end{array} \right\}$$

with $\mathcal{Z} = \{(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) \mid \boldsymbol{u} \in \mathcal{U}, \boldsymbol{v} \ge \boldsymbol{\phi}(\boldsymbol{u})\}$. That is to say, a generalized-moment ambiguity set can be mapped into an event-wise ambiguity set with only one scenario, that is, $S = 1$.

https://xiongpengnus.github.io/rsome/

Papers if you are interested:

- Delage and Ye (2010), Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research* 58(3):595-612
- Wolfram Wiesemann, Daniel Kuhn, Melvyn Sim (2014) Distributionally Robust Convex Optimization. *Operations Research* 62(6):1358-1376.
- Bertsimas, Dimitris, Melvyn Sim, and Meilin Zhang. 2019. Adaptive distributionally robust optimization. *Management Science* 65(2) 604-618.
- Chen, Zhi, Melvyn Sim, Peng Xiong. 2020. Robust stochastic optimization made easy with RSOME. *Management Science* 66(8) 3329-3339.

Distance-based ambiguity set:

$$\mathcal{B}(r) = \left\{ \mathbb{P} \in \mathcal{P} : d(\mathbb{P}, \hat{\mathbb{P}}) \leq r \right\}$$

- $\hat{\mathbb{P}}$: reference distribution,
- $r > 0$: radius of the ambiguity set.
- $d(\mathbb{P}, \hat{\mathbb{P}})$: distance between $\mathbb{P}$ and $\hat{\mathbb{P}}$.

Can be data driven! We can use the empirical (discrete) distribution as the reference distribution.

$$\hat{\mathbb{P}} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\hat{\xi}_i}.$$

ϕ**-divergence** (or f-divergence)):

$$d_\phi(\mathbb{P}, \hat{\mathbb{P}}) \begin{cases} = \sum \hat{p}\, \phi(\frac{p}{\hat{p}}), \text{ where } p \text{ and } \hat{p} \text{ are densities of } \mathbb{P}, \hat{\mathbb{P}} \\ = \int_\Omega \phi\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\hat{\mathbb{P}}}\right) \mathrm{d}\hat{\mathbb{P}} \end{cases}$$

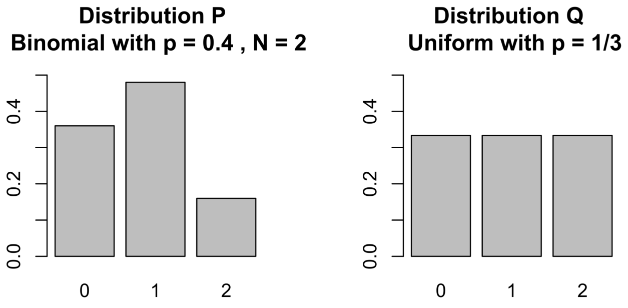
Kullback-Leibler divergence (relative entropy): $\phi(t) = t\log(t) - t + 1$:

$$d_{KL}(\mathbb{P}, \hat{\mathbb{P}}) = \sum p \log \frac{p}{\hat{p}}.$$

Hellinger distance: $\phi(t) = (\sqrt{t} - 1)^2$:

$$d_H(\mathbb{P}, \hat{\mathbb{P}}) = \sum(\sqrt{p} - \sqrt{\hat{p}})^2.$$

KL divergence



$$d(P,Q) = \frac{9}{25}\ln\left(\frac{9/25}{1/3}\right) + \frac{12}{25}\ln\left(\frac{12/25}{1/3}\right) + \frac{4}{25}\ln\left(\frac{4/25}{1/3}\right) \approx 0.085$$

$$d(Q,P) = \frac{1}{3}\ln\left(\frac{1/3}{9/25}\right) + \frac{1}{3}\ln\left(\frac{1/3}{12/25}\right) + \frac{1}{3}\ln\left(\frac{1/3}{4/25}\right) \approx 0.097$$

The ϕ-divergence based DRO:

$$\min_{\mathbf{x}} \max_{\mathbb{P}} \mathbb{E}_{\mathbb{P}}[f(\mathbf{x}, \tilde{\boldsymbol{\xi}})]$$
$$\text{s.t. } d_\phi(\mathbb{P}, \hat{\mathbb{P}}) \le \theta$$

Consider a discrete case:

$$\min_{\mathbf{x}} \max_{\mathbb{P}} \sum_{i}^{N} p_i[f(\mathbf{x}, \tilde{\boldsymbol{\xi}}_i)]$$
$$\text{s.t. } \sum_{i}^{N} q_i \phi(\frac{p_i}{q_i}) \le \theta \qquad (\alpha)$$
$$\sum_{i}^{N} p_i = 1 \qquad (\lambda)$$

By duality:

$$\min_{\mathbf{x},\lambda,\alpha} \lambda + \alpha \sum_{i}^{N} q_i \phi^* \left(\frac{f(\mathbf{x}, \tilde{\boldsymbol{\xi}}_i) - \lambda}{\alpha} \right) + \alpha\theta.$$

which is a finite-dimension convex optimization problem.

If we use the KL divergence and $f(\mathbf{x}, \boldsymbol{\xi}) = c(\boldsymbol{\xi})^\top x + g(\boldsymbol{\xi})$, we have:

$$\min_{\mathbf{x},\lambda,\alpha} \lambda + \alpha \sum_{i}^{N} q_i \exp \left(\frac{c(\tilde{\boldsymbol{\xi}}_i)^\top \mathbf{x} + g(\tilde{\boldsymbol{\xi}}_i) - \lambda}{\alpha} \right) + \alpha(\theta - 1).$$

where $\phi^*_{KL}(s) = e^s - 1$.
Exponential cone programming! can be solved by MOSEK
`https://docs.mosek.com/modeling-cookbook/expo.html`.

Papers if you are interested:

- Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, Gijs Rennen, (2012) Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science* 59(2):341-357.
- Güzin Bayraksan, David K. Love. (2015) Data-Driven Stochastic Programming Using Phi-Divergences. *In INFORMS Tutorials in Operations Research*. Published online: 26 Oct 2015; 1-19
- Bart P. G. Van Parys, Peyman Mohajerin Esfahani, Daniel Kuhn (2020) From Data to Decisions: Distributionally Robust Optimization Is Optimal. *Management Science* 67(6):3387-3402.

KL-divergence based DRO is the least conservative data-driven predictors and prescriptors whose out-of-sample disappointment decays at a rate no less than some prescribed threshold $r > 0$. (Van Parys et al., 2020)

Definition 1.

For any $p \in [1, \infty]$, the Wasserstein distance between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ is defined as:

$$W_p(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \left(\int_{\Omega \times \Omega} \|x - y\|^p \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p}.$$

where $\|\cdot\|$ is a norm on $\mathbb{R}^m$ and $\Pi(\mathbb{P}, \mathbb{Q})$ is the set of all probability measures on $\Omega \times \Omega$ with marginals $\mathbb{P}$ and $\mathbb{Q}$, respectively.

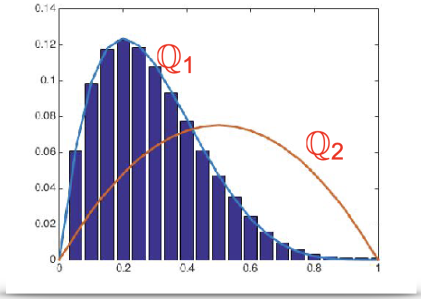
Wasserstein distance is a **metric**:

- nonnegative, symmetric, subadditive,
- it vanishes only if $\mathbb{P} = \mathbb{Q}$.
- it is finite whenever $\mathbb{P}$ and $\mathbb{Q}$ have finite p-th order moments.

Looking at the discrete case:

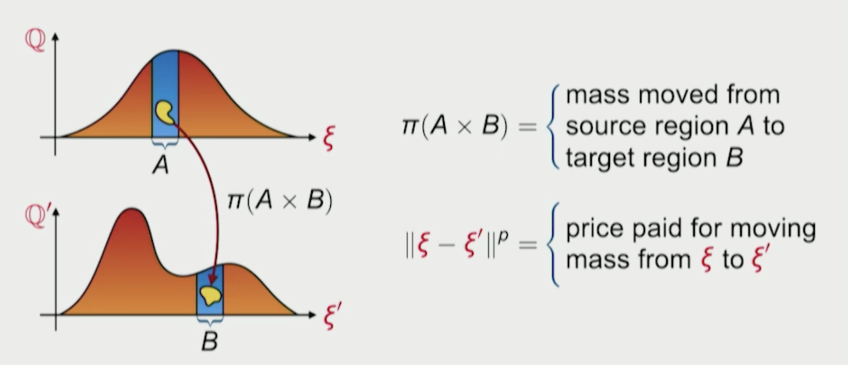
$$
\begin{aligned}
W_p(\mathbb{P}, \mathbb{Q}) = \min_{\Pi} & \sum_{k,l} \|\xi_k, \xi_l\| \Pi_{kl}. \\
\text{s.t.} & \sum_{l} \pi_{kl} = \mathbb{P}_k, \forall k \\
& \sum_{k} \pi_{kl} = \mathbb{Q}_l, \forall l
\end{aligned}
$$

It links to the optimal transport problem!



$d_{\mathrm{W}}(\mathbb{Q}_1, \mathbb{Q}_2) =$ minimum cost of moving $\mathbb{Q}_1$ to $\mathbb{Q}_2$

Link to optimal transport problem:



Dual Kantorovich Problem

Definition 2.

For any $p \in [1, \infty]$, the Wasserstein distance between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ is defined as:

$$W_p^p(\mathbb{P}, \mathbb{Q}) = \sup \int_\Omega \psi(x)\mathbb{P}(\mathrm{d}x) - \int_\Omega \phi(y)\mathbb{Q}(\mathrm{d}y).$$

$$\text{s.t. } \psi(x) - \phi(y) \leq \|x - y\|^p, \forall x, y \in \Omega.$$

Nominal distribution: In the absence of any structural information, it is convenient to set $\hat{\mathbb{P}}^N$ to the discrete empirical distribution: the uniform distribution on the N training samples $\{\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_N\}$,

$$\hat{\mathbb{P}}^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{\xi}_i},$$

where δ_ξ is the Dirac delta function centered at ξ.
Then, the Wasserstein DRO is

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}^{\mathbb{P}}\left[f(\boldsymbol{x}, \boldsymbol{\xi})\right] \\ \text{s.t.} \quad & \mathcal{F} = \{\mathbb{P} \mid W_p(\mathbb{P}, \hat{\mathbb{P}}^N) \leq \theta\}. \end{aligned}$$

Definition: $\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N) = \left\{ \mathbb{Q} \in \mathcal{P}(\Xi) : W_p(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon \right\}$

Contains every $\mathbb{Q}$ obtainable by re-shaping $\widehat{\mathbb{P}}_N$ at a cost of at most ε

Worst-case risk: $\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) = \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[\ell(\xi)]$

Worst-case optimal risk: $\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \mathcal{L}) = \inf_{\ell \in \mathcal{L}} \mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell)$

$$
\begin{aligned}
\min_{\boldsymbol{x}} \quad & \sup \int_{\xi} f(\boldsymbol{x}, \boldsymbol{\xi}) \mathbb{P}(\mathrm{d}\boldsymbol{\xi}) \\
\text{s.t.} \quad & W_p(\mathbb{P}, \hat{\mathbb{P}}^N) \leq \theta.
\end{aligned}
$$

Let's first focus on the inner sup problem:

$$
\begin{aligned}
\sup \quad & \int_{\xi} f(\boldsymbol{x}, \boldsymbol{\xi}) \mathbb{P}(\mathrm{d}\boldsymbol{\xi}) \\
\text{s.t.} \quad & \inf_{\pi \in \Pi(\mathbb{P}, \hat{\mathbb{P}}^N)} \left(\int_{\Omega \times \Omega} \|\xi - \hat{\xi}\|^p \pi(\mathrm{d}\xi, \mathrm{d}\hat{\xi}) \right)^{1/p} \leq \theta.
\end{aligned}
$$

Looking at the discrete case [This part may not rigorous enough, please refer to the paper for more details.]:

$$
\begin{aligned}
\sup_{\mathbb{P}} \quad & \sum_k f(\boldsymbol{x}, \boldsymbol{\xi})\mathbb{P}_k \\
\text{s.t.} \quad & \min_{\pi} \sum_{k,l} \|\xi_k - \xi_l\|^p \pi_{kl} \leq \theta. \\
& \sum_l \pi_{kl} = \mathbb{P}_k, \forall k \\
& \sum_k \pi_{kl} = \hat{\mathbb{P}}_l^N, \forall l
\end{aligned}
$$

$$
\begin{aligned}
\sup_{\mathbb{P}} \quad & \sum_k f(\boldsymbol{x}, \boldsymbol{\xi}) \mathbb{P}_k \\
\text{s.t.} \quad & \sum_{k,l} \|\xi_k - \hat{\xi}_l\|^p \pi_{kl} \leq \theta. \\
& \sum_l \pi_{kl} = \mathbb{P}_k, \forall k \\
& \sum_k \pi_{kl} = \frac{1}{N}, \forall l
\end{aligned}
\Rightarrow
\begin{aligned}
\sup_{\mathbb{P}} \quad & \sum_k \sum_l f(\boldsymbol{x}, \boldsymbol{\xi}) \pi_{kl} \\
\text{s.t.} \quad & \sum_{k,l} \|\xi_k - \hat{\xi}_l\|^p \pi_{kl} \leq \theta. \\
& \sum_k \pi_{kl} = \frac{1}{N}, \forall l
\end{aligned}
$$

$$
\begin{aligned}
\inf_{\lambda, s_i} \quad & \lambda\theta + \frac{1}{N}\sum_{i}^{N} s_i \\
\text{s.t.} \quad & s_i + \lambda\|\xi_k - \hat{\xi}_l\|^p \geq f(\boldsymbol{x}, \xi_k), \forall l \leq N, \forall k \\
& \lambda \geq 0.
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \inf_{\lambda, s_i} \quad & \lambda\theta + \frac{1}{N}\sum_{i}^{N} s_i \\
\text{s.t.} \quad & \max_{\xi \in \Xi} f(\boldsymbol{x}, \xi) - \lambda\|\xi - \hat{\xi}_i\|^p \leq s_i, \forall i \leq N \\
& \lambda \geq 0.
\end{aligned}
$$

$$\inf_{\lambda, s_i} \ \lambda\theta + \frac{1}{N}\sum_{i}^{N} s_i$$

$$\text{s.t. } [-f]^*(z_i - v_i) + \sigma_{\Xi}(v_i) - z_i^T \hat{\xi}_i + \psi(q)\lambda \|\frac{z_i}{\lambda}\|_*^q \leq s_i, \forall i \leq N.$$

where $\psi(q) = (q-1)^{q-1}/q^q$ for $q > 1$ and $\psi(1) = 1$, $\|\cdot\|_*$ represents dual norm and f^* denotes the conjugate function. $(f^*(y) = \sup\{xy - f(x), \forall x\})$

Finite convex program

Assumptions:

- $p = 1$.
- $\xi \in \Xi$ satisfies $C\xi \leq D$
- $f(\xi)$ is piecewise linear $f(\xi) = \max\{a\xi + b, 0\}$.

Then, we have

$$\max_{\xi \in \Xi} f(\xi) - \lambda\|\xi - \hat{\xi}_i\| = \max_{\xi \in \Xi} \max\{a\xi + b, 0\} - \lambda\|\xi - \hat{\xi}_i\| \leq s_i, \forall i$$

$$\Rightarrow \begin{cases} \max\limits_{\xi \in \Xi} a\xi + b - \lambda\|\xi - \hat{\xi}_i\| \leq s_i, \forall i \\ \max\limits_{\xi \in \Xi} -\lambda\|\xi - \hat{\xi}_i\| \leq s_i, \forall i \end{cases}$$

$$
\begin{aligned}
&\max_{\xi\in\Xi} a\xi + b - \lambda\|\xi - \hat{\xi}_i\| \le s_i, \forall i \\
&\max_{\xi\in\Xi} a\xi + b - \max_{||v_i||_*\le\lambda} v_i(\xi - \hat{\xi}_i) \le s_i, \forall i && \text{(Dual Norm)} \\
\min_{||v_i||_*\le\lambda} &\max_{\xi\in\Xi} a\xi + b - v_i(\xi - \hat{\xi}_i) \le s_i, \forall i && \text{(Change Seq of min max)}
\end{aligned}
$$

$$
\begin{cases}
\max_{\xi\in\Xi} a\xi + b - v_i(\xi - \hat{\xi}_i) \le s_i \\
||v_i||_* \le \lambda
\end{cases}
\tag{1}
$$

$$\max_{\xi \in \Xi} a\xi + b - v_i(\xi - \hat{\xi}_i) = \max_{\xi \in \Xi} (a - v_i)\xi + b - v_i\hat{\xi}_i$$

$$\max_{\xi \in \Xi} (a - v_i)\xi + b - v_i\hat{\xi}_i$$

$$\text{s.t. } C\xi \leq D$$

$$\min_u Du + b + v_i\hat{\xi}_i$$

$$\text{s.t.} Cu \geq a - v_i$$

$$\Rightarrow \min_u Du + b + (a - Cu)\hat{\xi}_i = a\hat{\xi}_i + b + u(D - C\hat{\xi}_i)$$

To sum up:

$$
\begin{aligned}
\sup_{\xi\in\Xi} \quad & \int_\xi \max\{a\xi+b,0\}\mathbb{P}(\mathrm{d}\boldsymbol{\xi}) \\
\text{s.t.} \quad & W_p(\mathbb{P},\hat{\mathbb{P}}^N)\le\theta.
\end{aligned}
\iff
\begin{aligned}
\min_{\lambda,s_i,u,v} \quad & \lambda\theta+\frac{1}{N}\sum_i^N s_i \\
\text{s.t.}\; & a\hat{\xi}_i+b+u(D-C\hat{\xi}_i)\le s_i, \forall i \\
& |a-Cu|\le\lambda \\
& v(D-C\hat{\xi}_i)\le s_i, \forall i \\
& |Cv|\le\lambda
\end{aligned}
$$

where $\Xi=\{\xi|C\xi\le D\}$.

Merits of Wasserstein DRO:

- Fidelity: DRO are more "honest" than their nominal counterparts, as they acknowledge the presence of distributional uncertainty.
- Tractability: finite convex program (when $p = 1$ finite LP)
- Performance guarantee:
- Regularization by Robustification:

Performance guarantee:

Theorem 18 (Concentration Inequalities I). *Suppose that $\widehat{\mathbb{P}}_N$ is the empirical distribution, whereas $p \neq m/2$, and the unknown true distribution $\mathbb{P}$ is light-tailed in the sense that there exist $\alpha > p$ and $A > 0$ such that $\mathbb{E}^{\mathbb{P}}[\exp(\|\xi\|^{\alpha})] \leq A$. Then, there are constants $c_1, c_2 > 0$ that depend on $\mathbb{P}$ only through α, A,* and *m such that for any $\eta \in (0, 1]$, the concentration inequality $\mathbb{P}^N[\mathbb{P} \in \mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)] \geq 1 - \eta$ holds whenever ε exceeds*

$$\varepsilon_{p,N}(\eta) = \begin{cases} \left(\dfrac{\log(c_1/\eta)}{c_2 N}\right)^{\min\{p/m, 1/2\}} & \text{if } N \geq \dfrac{\log(c_1/\eta)}{c_2}, \\ \left(\dfrac{\log(c_1/\eta)}{c_2 N}\right)^{p/\alpha} & \text{if } N < \dfrac{\log(c_1/\eta)}{c_2}. \end{cases} \tag{26}$$

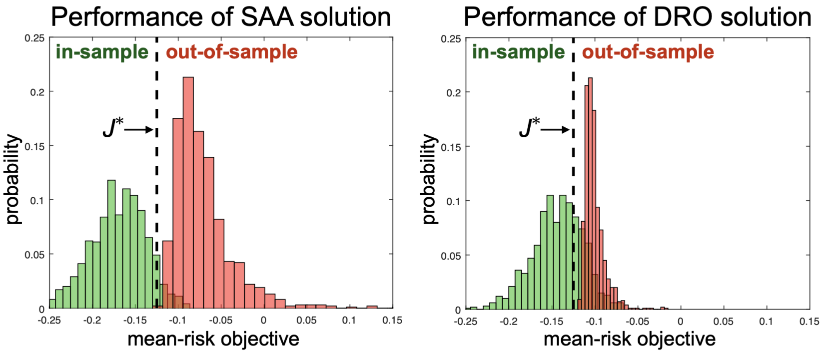
Variation Regularization:
For Wasserstein DRO, for a broad class of loss functions, possibly non-convex and non-smooth, with high probability, the Wasserstein DRO is asymptotically equivalent to variation regularization problem. [Not rigorous here, please refer to the paper for more details.]

$$\min_{x} \mathbb{E}_{\xi\sim\hat{\mathbb{P}}^N}[f(x,\xi)] + \rho\mathcal{V}(f)$$

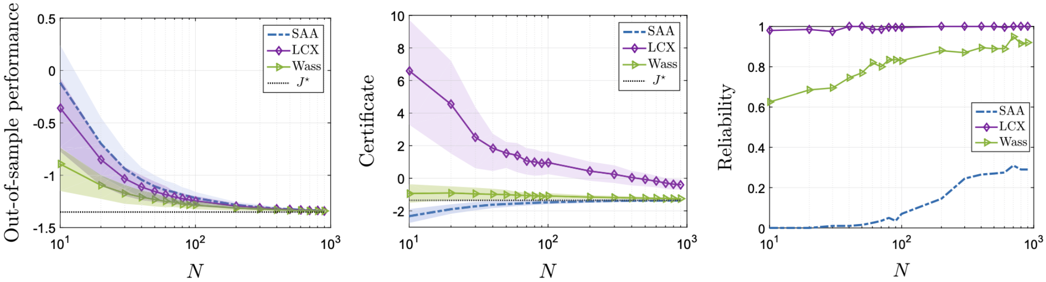
DEFINITION 1 (VARIATION). Let $q \in [1,\infty]$ and f be a continuous function on $\mathcal{Z}$. When $q \in [1,\infty)$, assume ∇f exists $\mathbb{Q}$-almost everywhere. The *variation* of f with respect to $\mathbb{Q}$ is defined as

$$\mathcal{V}_{\mathbb{Q},q}(f) := \begin{cases} \| \|\nabla f\|_* \|_{\mathbb{Q},q}, & q \in [1,\infty), \\ \mathbb{Q}\text{-}\operatorname{ess\,sup}_{z\in\mathcal{Z}} \sup_{\tilde{z}\neq z} \dfrac{(f(\tilde{z})-f(z))_+}{\|\tilde{z}-z\|}, & q=\infty. \end{cases}$$

Rui Gao, Xi Chen, Anton J. Kleywegt (2022) Wasserstein Distributionally Robust Optimization and Variation Regularization. Operations Research 0(0).



- in-sample: optimistic bias
- out-of-sample: pessimistic bias
- **DRO reduces bias & post-decision disappointment**



where LCX is linear-convex ordering (LCX)-based goodness-of-fit test. from [Bertsimas, Gupta and Kallus(2014) Robust SAA.]

$$\min_x \mathbb{E}^{\mathbb{P}}\left[cx - p\min(x, D)\right] = \mathbb{E}^{\mathbb{P}}\left[(c-p)x + p(x-\tilde{D})^+\right]$$

When the demand $\{\tilde{D}_t\}$ is i.i.d. process with distribution $\mathbb{P}$, the optimal solution is $x^* = \inf\left\{y : F(y) < \frac{p-c}{p}\right\}$.

We can use SAA, moment-based DRO, Wasserstein DRO or KL-divergence based DRO to solve this problem.

Let's do this!

- $c = 5, p = 7$.
- Demand distribution: $D \sim \text{Binomial}(10, 0.5) + 1$.
- Sample size: $N = 50$
- Generate the sample $\{\hat{D}_t\}$.

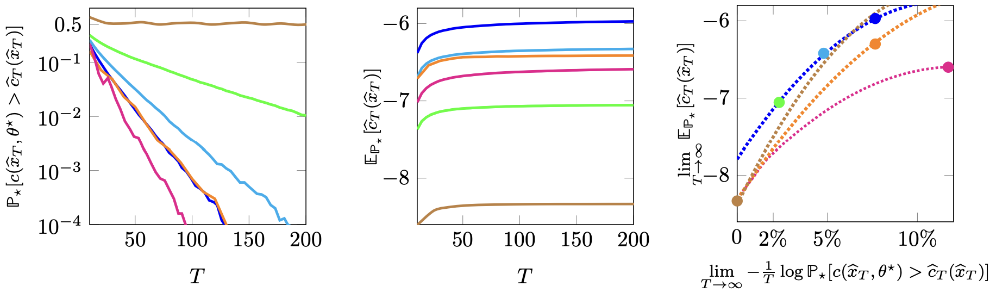
SAA:

$$\hat{x}^*_{\text{SAA}} = \arg\min_x \frac{1}{N} \sum_{t=1}^{N} \left[(c-p)x + p(x - \hat{D}_i)^+ \right]$$

DRO:

$$\hat{x}^*_{\text{DRO}} = \arg\min_x \max_{\mathbb{P} \in \mathcal{F}} \mathbb{E}^{\mathbb{P}} \left[(c-p)x + p(x - \hat{D}_i)^+ \right]$$

You may try the Julia language for optimization.



(a) Out-of-sample disappointment versus training sample size T

(b) In-sample cost versus training sample size T

(c) Asymptotic in-sample cost versus decay rate of out-of-sample disappointment

Sutter, T., Van Parys, B. P., & Kuhn, D. (2021). A general framework for optimal data-driven optimization. arXiv preprint arXiv:2010.06606v2.

Materials if you are interested

- Lectures
  - Daniel Kuhn's talk at DTU
  - Daniel Kuhn's talk at INFORMS 2019
  - Wenzao Su's summer school
- People
  - Daniel Kuhn `https://www.epfl.ch/labs/rao/`
  - Jose Blanchet `https://web.stanford.edu/~jblanche/`
  - Melvyn Sim `https://bizfaculty.nus.edu.sg/faculty-details/?profId=127`
  - Gao Rui, Chen Zhi
- Misc
  - RSOME: Robust Stochastic Optimization Made Easy
  - `https://github.com/Operations-Research-Science/Ebook-An_introduction_to_robust_optimization`

- Bertsimas, D., and Sim, J. (2004). The price of robustness. *Management Science*, 50(1), 1-13.
- Delage, E., and Ye, Y. (2010). Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research*.
- Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh. Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning. In INFORMS TutORials in Operations Research. 2019.

- Relationship with regularization
- Multi-period robust optimization
- Chance constraints
- Robust satisficing
- Statistical properties
- ...

Data driven robust satisficing:

$$\begin{aligned} \kappa_\tau = \min \quad & k \\ \text{s.t.} \quad & \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] - \tau \leq k\Delta(\mathbb{P},\hat{\mathbb{P}}), \forall \mathbb{P} \\ & \boldsymbol{x} \in \mathcal{X}, k > 0 \end{aligned}$$

which is equivalent to

$$\begin{aligned} \kappa_\tau = \min \quad & k \\ \text{s.t.} \quad & \frac{1}{N}\sum y_i \leq \tau \\ & y_i \geq \sup_{\xi_i} \left\{ f(x,\xi_i) - k\|\xi_i - \hat{\xi}_i\| \right\} \forall i \\ & \boldsymbol{x} \in \mathcal{X}, k > 0 \end{aligned}$$

- Daniel Zhuoyu Long, Melvyn Sim, Minglong Zhou, (2022) Robust Satisficing. *Operations Research* 0(0).
- Rui Gao (2022) Finite-Sample Guarantees for Wasserstein Distributionally Robust Optimization: Breaking the Curse of Dimensionality. *Operations Research* 0(0).
- Rui Gao, Xi Chen, Anton J. Kleywegt (2022) Wasserstein Distributionally Robust Optimization and Variation Regularization. *Operations Research* 0(0).
- Sutter, T., Van Parys, B. P., & Kuhn, D. (2021). A general framework for optimal data-driven optimization. arXiv preprint arXiv:2010.06606v2.
- Chen, Li and Sim, Melvyn and Zhang, Xun and Zhou, Minglong, Robust Explainable Prescriptive Analytics (May 11, 2022). SSRN

- Zhaowei Hao, Long He, Zhenyu Hu, & Jun Jiang (2020), Robust Vehicle Pre-Allocation with Uncertain Covariates. *Production and Operations Management*, 29: 955-972
- Jose Blanchet, Lin Chen, Xun Yu Zhou (2022) Distributionally Robust Mean-Variance Portfolio Selection with Wasserstein Distances. *Management Science* 68(9):6382-6410.
- Luying Sun, Weijun Xie, Tim Witten (2022) Distributionally Robust Fair Transit Resource Allocation During a Pandemic. *Transportation Science* 0(0).
- Long He, Sheng Liu, Zuo-Jun Max Shen (2022), Smart urban transport and logistics: A business analytics perspective. *Production and Operations Management*, 31, 3771-3787.