Sinkhorn Distributionally Robust Optimization

Jie Wang[†], Rui Gao[‡], Yao Xie[†]

† H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology

‡ Department of Information, Risk, and Operations Management, University of Texas at Austin

Contributions

- Distributionally robust optimization with entropic regularized Wasserstein distance (Sinkhorn distance).
- Ambiguity set contains only absolutely continuous distributions.
- Computationally efficient first-order optimization algorithm.

Decision-Making Under Uncertainty

- Objective: Find decision θ to minimize the risk

$$\mathcal{R}(\theta;\mathbb{P}) = \mathbb{E}_{\mathbb{P}}\big[f_\theta(z)\big].$$

- Available Information:

Structual : $\mathbb{P}$ is supported on $\Omega \subseteq \mathbb{R}^d$

Statistical : $\hat{x}_1,\ldots,\hat{x}_n \sim \mathbb{P}$

Supply Chain Mgmt. Portfolio Mgmt. Machine Learning

- Sample Average Approximation (SAA):

$$\inf_{\theta\in\Theta} \quad \Big\{\mathcal{R}(\theta;\widehat{\mathbb{P}}_n) \triangleq \mathbb{E}_{\widehat{\mathbb{P}}_n}\big[f_\theta(z)\big]\Big\},$$

where $\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\hat{x}_i}$.

- Distributionally Robust Optimization (DRO):

$$\inf_{\theta\in\Theta} \quad \left\{\sup_{\mathbb{P}\in\mathcal{P}} \; \mathbb{E}_{\mathbb{P}}\big[f_\theta(z)\big]\right\},$$

where $\mathcal{P} = \big\{\mathbb{P}:\; W(\widehat{\mathbb{P}},\mathbb{P}) \le \rho\big\}$.

- Facts about Wasserstein DRO:
  - For WDRO with n-point nominal distribution, the worst-case distribution is supported on $n+1$ points.
  - Finite-dimensional convex reformulation is available if the objective is a pointwise maximum of finitely many concave functions.
  - Some cases the same performance as SAA.

Sinkhorn Robust Formulation

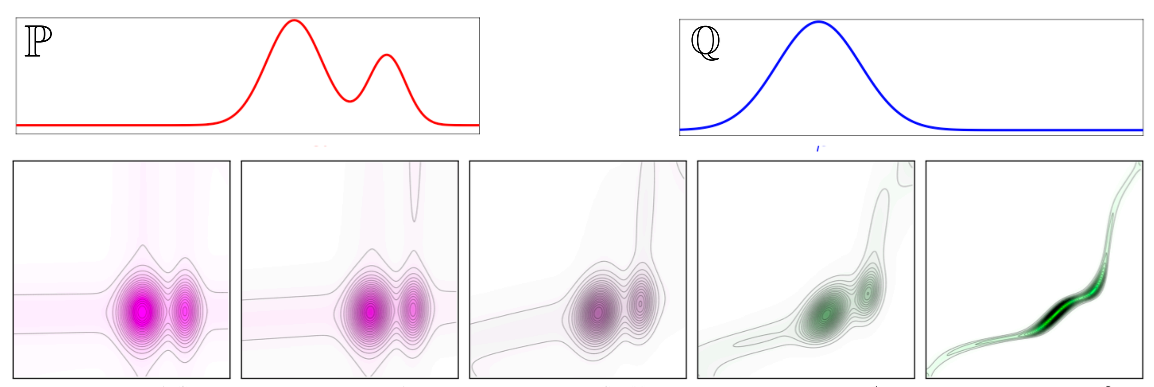
- Sinkhorn Distance [Cuturi 2013]:

$$\mathcal{W}_\varepsilon(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \Big\{\mathbb{E}_{(X,Y)\sim\gamma}[c(X,Y)] + \varepsilon H(\gamma \mid \mathbb{P}\otimes\nu)\Big\}.$$

Relative Entropy between γ and $\mathbb{P}\otimes\nu$:

$$H(\gamma \mid \mathbb{P}\otimes\nu) = \int \log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\mathbb{P}(x)\,\mathrm{d}\nu(y)}\right)\mathrm{d}\gamma(x,y).$$

- Visualization of Transport Mapping γ for Varying ε:



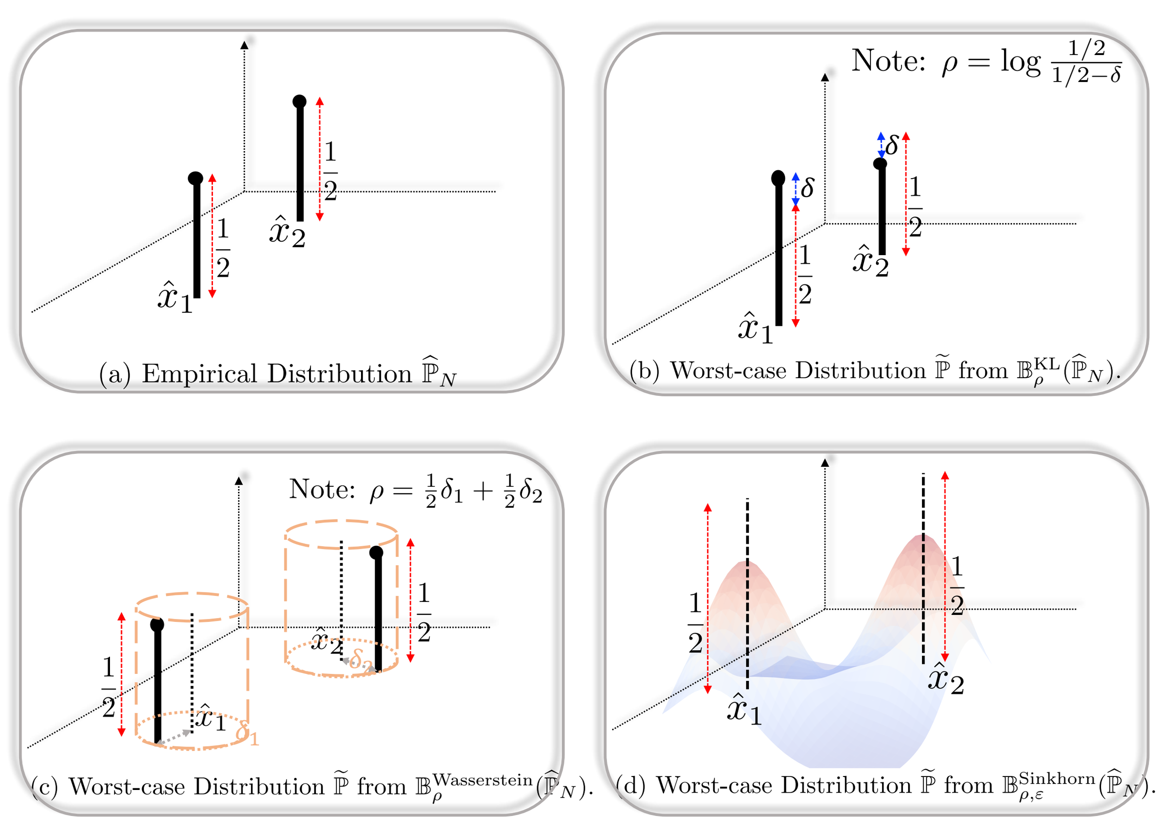
$\varepsilon = 10$ $\quad\varepsilon = 1$ $\quad\varepsilon = 0.5$ $\quad\varepsilon = 10^{-1}$ $\quad\varepsilon = 10^{-2}$

- Sinkhorn DRO:

$$V^* = \inf_{\theta\in\Theta} \sup_{\mathbb{P}\in\mathbb{B}_{\rho,\varepsilon}(\widehat{\mathbb{P}})} \mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)],$$

where $\mathbb{B}_{\rho,\varepsilon}(\widehat{\mathbb{P}}) = \big\{\mathbb{P}:\; \mathcal{W}_\varepsilon(\widehat{\mathbb{P}},\mathbb{P}) \le \rho\big\}$.

- Visualization of Worst-Case Distributions:



(a) Empirical Distribution $\widehat{\mathbb{P}}_N$
(b) Worst-case Distribution $\widetilde{\mathbb{P}}$ from $\mathbb{B}^{\mathrm{KL}}_{\rho}(\widehat{\mathbb{P}}_N)$
(c) Worst-case Distribution $\widetilde{\mathbb{P}}$ from $\mathbb{B}^{\mathrm{Wasserstein}}_{\rho}(\widehat{\mathbb{P}}_N)$
(d) Worst-case Distribution $\widetilde{\mathbb{P}}$ from $\mathbb{B}^{\mathrm{Sinkhorn}}_{\rho,\varepsilon}(\widehat{\mathbb{P}}_N)$

Remark:

(i) Most DRO models give discrete distributional estimate;

(ii) Sinkhorn DRO model gives continuous estimate.

Theorem: Strong Dual Reformulation

Assume that

- $\nu\{z:\; 0 \le c(x,z) < \infty\} = 1$ for $\widehat{\mathbb{P}}$-almost every x;
- $\int e^{-c(x,z)/\varepsilon}\,\mathrm{d}\nu(z) < \infty$ for $\widehat{\mathbb{P}}$-almost every x;
- $\mathcal{Z}$ is a measurable space;
- Function $f:\; \mathcal{Z} \to \mathbb{R}\cup\{\infty\}$ is measurable.

Then $V_{\mathrm{P}} = V_{\mathrm{D}}$:

$$V_{\mathrm{P}} = \sup_{\mathbb{P}} \; \Big\{\mathbb{E}_{z\sim\mathbb{P}}[f(z)]:\quad W_\varepsilon(\widehat{\mathbb{P}},\mathbb{P}) \le \rho\Big\},$$

$$V_{\mathrm{D}} = \inf_{\lambda>0} \; \lambda\bar{\rho} + \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\lambda\varepsilon\log\Big(\mathbb{E}_{z\sim\mathbb{Q}_x}\Big[e^{f(z)/(\lambda\varepsilon)}\Big]\Big)\Big],$$

where

$$\bar{\rho} = \rho + \varepsilon\int_\Omega \log\left(\int_\Omega e^{-c(x,z)/\varepsilon}\,\mathrm{d}\nu(z)\right)\mathrm{d}\widehat{\mathbb{P}}(x),$$

$$\mathrm{d}\mathbb{Q}_x(z) = \frac{e^{-c(x,z)/\varepsilon}}{\int_\Omega e^{-c(x,u)/\varepsilon}\,\mathrm{d}\nu(u)}\,\mathrm{d}\nu(z).$$

Proof Sketch of Strong Duality

1. First show the weak duality result $V_{\mathrm{P}} \le V_{\mathrm{D}}$.
2. Construct primal feasible solution $\widetilde{\mathbb{P}}$ with

$$V_{\mathrm{P}} \ge \mathbb{E}_{z\sim\widetilde{\mathbb{P}}}[f(z)] = V_{\mathrm{D}}.$$

Geometry of Worst-Case Distribution:

- For each $x\in\mathrm{supp}(\widehat{\mathbb{P}})$, optimal transport maps it to a (conditional) distribution γ_x:

$$\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)} = \alpha_x\cdot\exp\Big(\big(f(z) - \lambda^* c(x,z)\big)/(\lambda^*\varepsilon)\Big).$$

- Worst-case distribution $\widetilde{\mathbb{P}} = \int \gamma_x\,\mathrm{d}\widehat{\mathbb{P}}(x)$.

Algorithm for Robust Learning

$$V^* = \min_{\lambda\ge 0} \; \{\lambda\bar{\rho} + V(\lambda)\},$$

where $V(\lambda) \triangleq \min_{\theta\in\Theta} \; \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\lambda\varepsilon\log\Big(\mathbb{E}_{\mathbb{Q}_x}\Big[e^{f_\theta(z)/(\lambda\varepsilon)}\Big]\Big)\Big]$

Bisection Search on λ: Estimating $V(\lambda)$ up to accuracy $O(\delta)$ for $O(\mathrm{Poly}(\log\frac{1}{\delta}))$ times to find δ-optimal solution of V^*.

Stochastic Approximation for Solving $V(\lambda)$

- Goal: to solve the optimization

$$\min_{\theta\in\Theta} \; \Big\{F(\theta) \triangleq \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\lambda\varepsilon\log\Big(\mathbb{E}_{z\sim\mathbb{Q}_x}\Big[e^{f_\theta(z)/(\lambda\varepsilon)}\Big]\Big)\Big]\Big\}.$$

- Biased Stochastic Mirror Descent (BSMD): For $t = 1,\ldots,T$,

$$\begin{cases} v(\theta_t) \leftarrow \text{(biased) gradient estimate of } F(\theta_t) \\ \theta_{t+1} \leftarrow \mathbf{Prox}_{\theta_t}(\gamma v(\theta_t)) \end{cases}$$

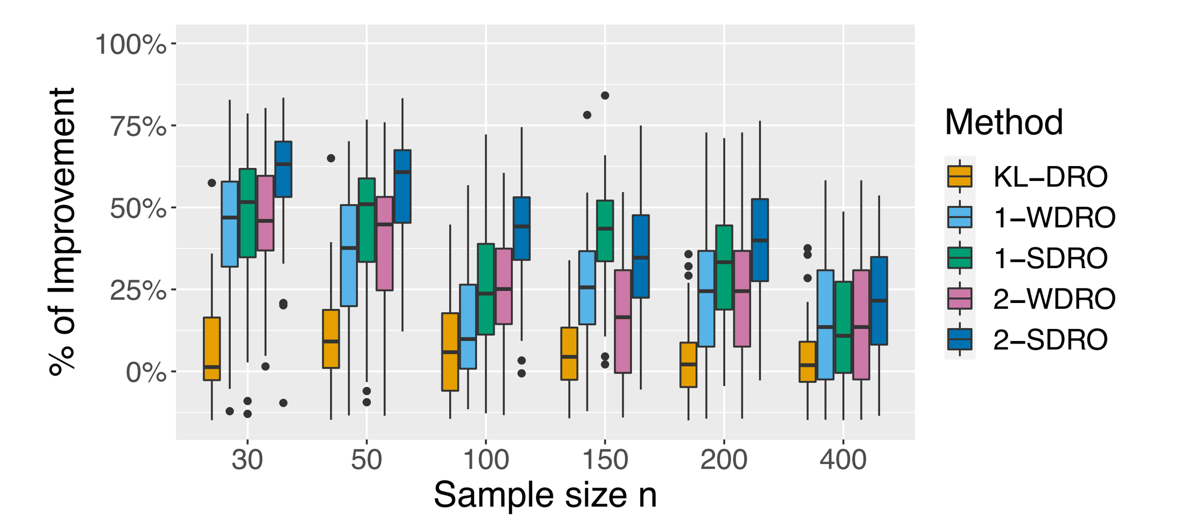
Remark: Gradient estimators should optimally balance the bias-variance trade-off.

- Complexity of finding δ-optimal solution or δ-critical point:

Estimators	Convex Nonsmooth	Convex Smooth	Nonconvex Smooth
Vanilla SGD	$O(\delta^{-3})$	$O(\delta^{-3})$	$O(\delta^{-6})$
V-MLMC	N/A	$\tilde{O}(\delta^{-2})$	$\tilde{O}(\delta^{-4})$
RT-MLMC	N/A	$\tilde{O}(\delta^{-2})$	$\tilde{O}(\delta^{-4})$

Mean-Risk Portfolio Optimization

Performance for Varying Sample Size n



Performance for Varying Data Dimension D