Sinkhorn Distributionally Robust Optimization

Jie Wang[1], Rui Gao[2], and Yao Xie[1]

[1]Georgia Institute of Technology

[2]The University of Texas at Austin

2022 INFORMS Annual Meeting

Decision-Making Under Uncertainty

$$\begin{aligned}
\text{Risk :} \quad & \mathscr{R}(\theta;\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[f_\theta(z)] \\
\text{Optimal Risk :} \quad & \mathscr{R}(\Theta;\mathbb{P}) = \inf_{\theta\in\Theta} \; \mathbb{E}_{\mathbb{P}}[f_\theta(z)]
\end{aligned}$$

- Available Information:

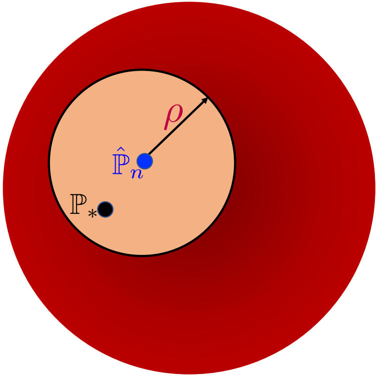
$$\begin{aligned}
\text{Structual :} \quad & \mathbb{P} \text{ is supported on } \Omega \subseteq \mathbb{R}^d \\
\text{Statistical :} \quad & \hat{x}_1, \ldots, \hat{x}_n \sim \mathbb{P}
\end{aligned}$$

- Nominal Problem:

  - Non-parametric estimators: $\hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\hat{x}_i}$.
  - Kernel density estimators: $\hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^n K(\hat{x}_i)$.

Wasserstein DRO

Definition: $\mathscr{P} = \{\mathbb{P} : \ W(\mathbb{P}, \hat{\mathbb{P}}_n) \leq \boldsymbol{\rho}\}$.



Contain each $\mathbb{P}$ such that $W(\mathbb{P}, \hat{\mathbb{P}}_n) \leq \rho$

Worst-case risk : $\sup_{\mathbb{P} \in \mathscr{P}} \mathbb{E}_{\mathbb{P}}[f_{\boldsymbol{\theta}}(z)]$

Robust Optimal Risk : $\inf_{\boldsymbol{\theta} \in \Theta} \sup_{\mathbb{P} \in \mathscr{P}} \mathbb{E}_{\mathbb{P}}[f_{\boldsymbol{\theta}}(z)]$

Limitations of Wasserstein DRO

- Worst-case distribution is discrete:

  For WDRO with n-point nominal distribution, the worst-case distribution is supported on $n+1$ points[1]*.*

- Tractability for limited scenarios:

  Finite-dimensional convex reformulation is available if the objective is a pointwise maximum of finitely many concave functions[2]*.*

- Some cases the same performance as SAA[2].

[1] Rui Gao and Anton J. Kleywegt. "Distributionally Robust Stochastic Optimization with Wasserstein Distance". In: *arXiv preprint arXiv:1604.02199* (Apr. 2016).

[2] Peyman Mohajerin Esfahani and Daniel Kuhn. "Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations". In: *Mathematical Programming* 171.1 (July 2017), pp. 115–166.

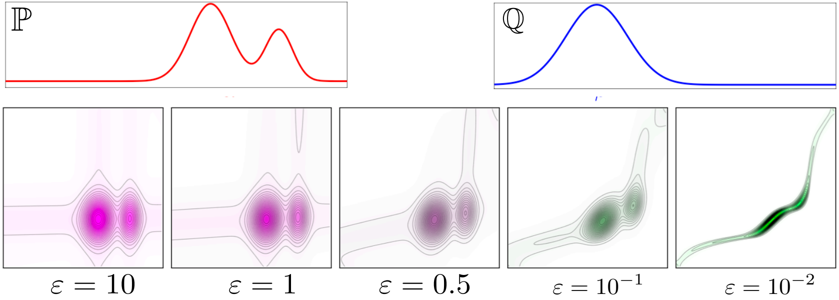
Sinkhorn Distance

- Sinkhorn Distance [Cuturi 2013]:

$$W_\varepsilon(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{\mathbb{E}_{(X,Y)\sim\gamma}[c(X,Y)] + \varepsilon H(\gamma \,|\, \mathbb{P}\otimes\nu)\right\}.$$

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

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



Highlights of Sinkhorn Distance

Probability distance between distributions in $\mathbb{R}^d$ using n samples:

	MMD	Wasserstein	Sinkhorn
Computation	$O(n)$	$\tilde{O}(n^3)$	$\tilde{O}(n^2)$ [Altschuler, Niles-Weed, and Rigollet 2017]
Sample Complexity	$O(n^{-1/2})$	$O(n^{-1/d})$	$O(e^{\kappa/\varepsilon} n^{-1/2} \varepsilon^{-\lfloor d/2 \rfloor})$[Genevay et al. 2019]

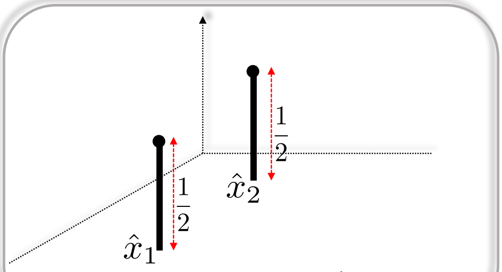
- Fast algorithms for implementation;
- Sharp sample complexity rate;
- Encourage stochastic optimal transport (helpful in some applications, e.g., domain adaptation [Courty, Flamary, and Tuia 2014]).

Main Framework

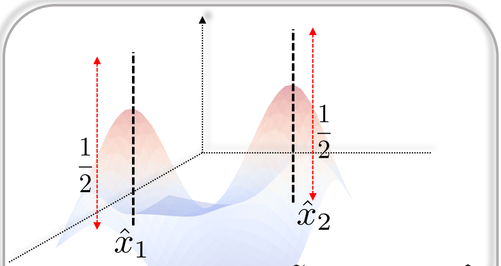
- Sinkhorn DRO:

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

$$\text{where } \mathbb{B}_{\rho,\varepsilon}(\widehat{\mathbb{P}}) = \{\mathbb{P}:\ W_\varepsilon(\widehat{\mathbb{P}},\mathbb{P}) \le \rho\}.$$



(a) Empirical Distribution $\widehat{\mathbb{P}}_N$



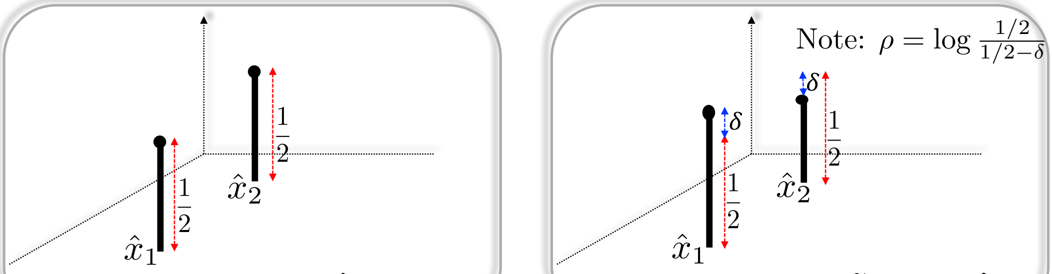
(d) Worst-case Distribution $\widetilde{\mathbb{P}}$ from $\mathbb{B}^{\text{Sinkhorn}}_{\rho,\varepsilon}(\widehat{\mathbb{P}}_N)$.

General DRO Models

- KL-DRO:

$$\inf_{\theta} \sup_{\mathbb{P} \in \mathbb{B}_{\rho}^{\mathsf{KL}}(\widehat{\mathbb{P}})} \mathbb{E}_{z \sim \mathbb{P}}[f_{\theta}(z)],$$

$$\text{where } \mathbb{B}_{\rho}^{\mathsf{KL}}(\widehat{\mathbb{P}}) = \{\mathbb{P} : D_{\mathsf{KL}}(\widehat{\mathbb{P}}, \mathbb{P}) \leq \rho\}.$$



(a) Empirical Distribution $\widehat{\mathbb{P}}_N$

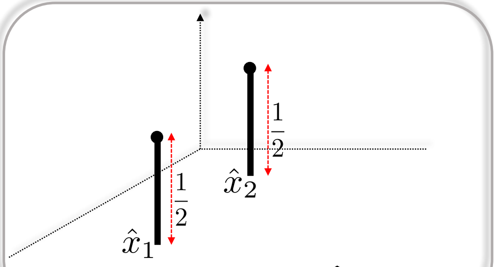
(b) Worst-case Distribution $\widetilde{\mathbb{P}}$ from $\mathbb{B}_{\rho}^{\mathrm{KL}}(\widehat{\mathbb{P}}_N)$.

General DRO Models

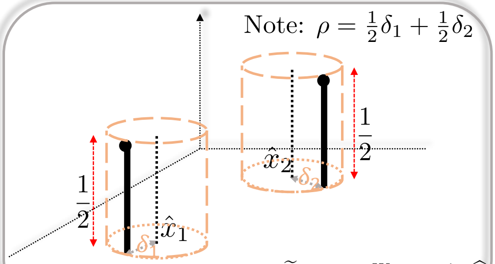
- Wasserstein-DRO:

$$\inf_{\theta} \sup_{\mathbb{P} \in \mathbb{B}_{\rho}^{\text{Wasserstein}}(\widehat{\mathbb{P}})} \mathbb{E}_{z \sim \mathbb{P}}[f_\theta(z)],$$

$$\text{where } \mathbb{B}_{\rho}^{\text{Wasserstein}}(\widehat{\mathbb{P}}) = \{\mathbb{P} : W(\widehat{\mathbb{P}}, \mathbb{P}) \le \rho\}.$$



(a) Empirical Distribution $\widehat{\mathbb{P}}_N$



(c) Worst-case Distribution $\widetilde{\mathbb{P}}$ from $\mathbb{B}_{\rho}^{\text{Wasserstein}}(\widehat{\mathbb{P}}_N)$.

Ongoing Outline

- Sinkhorn DRO:

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

$$\text{where } \mathbb{B}_{\rho,\varepsilon}(\widehat{\mathbb{P}}) = \left\{\mathbb{P}:\ W_\varepsilon(\widehat{\mathbb{P}},\mathbb{P}) \le \rho\right\}.$$

  - Duality Formulation for Sinkhorn DRO
  - First-order Optimization Algorithm
  - Properties and Numerical Results

Tractable Formulation

Assume that

(I) $\nu\{z:\ 0 \le c(x,z) < \infty\} = 1$ for $\widehat{\mathbb{P}}$-almost every x;

(II) The integral $\int e^{-c(x,z)/\varepsilon}\,\mathrm{d}\nu(z) < \infty$ for $\widehat{\mathbb{P}}$-almost every x;

(III) Ω is a measurable space, and the function $f:\ \Omega \to \mathbb{R}\cup\{\infty\}$ is measurable.

Consider the primal

$$V_{\mathrm{P}} = \sup_{\mathbb{P}\in\mathbb{B}_{\rho,\varepsilon}(\widehat{\mathbb{P}})} \mathbb{E}_{z\sim\mathbb{P}}[f(z)], \quad \text{where } \mathbb{B}_{\rho,\varepsilon}(\widehat{\mathbb{P}}) = \left\{\mathbb{P}:\ W_{\varepsilon}(\widehat{\mathbb{P}},\mathbb{P}) \le \rho\right\}. \qquad \texttt{(Sinkhorn DRO)}$$

It admits the **strong dual reformulation**:

$$V_{\mathrm{D}} = \inf_{\lambda>0}\ \lambda\overline{\rho} + \lambda\varepsilon\int_{\Omega} \log\left(\mathbb{E}_{\mathbb{Q}_x}\left[e^{f(z)/(\lambda\varepsilon)}\right]\right)\mathrm{d}\widehat{\mathbb{P}}(x),$$

where

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

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

Interpretation of Worst-case Distribution

$$\tilde{\mathbb{P}} = \arg\max_{\mathbb{P}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[f(z)] : \; W_\varepsilon(\widehat{\mathbb{P}}, \mathbb{P}) \le \rho \right\}$$

- For each $x \in \mathrm{supp}(\widehat{\mathbb{P}})$, optimal transport maps it to a (conditional) distribution γ_x such that
$$\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)} = \alpha_x \cdot \exp\Big(\big(f(z) - \lambda^* c(x,z)\big) / (\lambda^* \varepsilon) \Big),$$
where α_x is the normalizing constant.
- Closed-form expression on $\tilde{\mathbb{P}}$:
$$\frac{\mathrm{d}\tilde{\mathbb{P}}(z)}{\mathrm{d}\nu(z)} = \int \alpha_x \cdot \exp\Big(\big(f(z) - \lambda^* c(x,z)\big) / (\lambda^* \varepsilon) \Big) \, \mathrm{d}\widehat{\mathbb{P}}(x).$$

Worst-case distribution $\tilde{\mathbb{P}}$ support on whole space, while W-DRO is discrete.

Toy Example: Newsvendor

Newsvendor problem: (β: Demand); ($u\min\{\beta,z\}$: Earning); ($k\beta$: Loss).

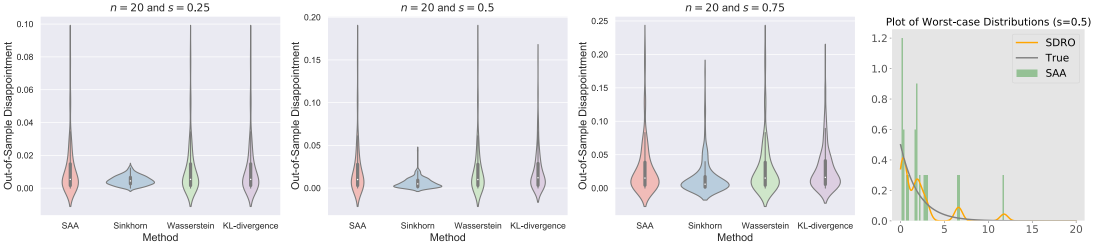
$$\min_{\beta}\ \mathbb{E}_{\mathbb{P}_*}\left[k\beta - u\min\{\beta,z\}\right], \quad k=5, u=7.$$

Performance and Visualization

Newsvendor problem:

$$\min_{\beta} \mathbb{E}_{\mathbb{P}_*}\left[k\beta - u\min\{\beta,\zeta\}\right], \quad k=5, u=7.$$

$\mathbb{P}_* \sim \exp(1/s)$ with $s \in \{0.25, 0.5, 0.75\}$. Access to $n=20$ samples.



Connection of Sinkhorn DRO with Wasserstein DRO

When $\varepsilon \to 0$, the dual objective of Sinkhorn DRO converges into

$$\lambda\rho + \int \text{ess-sup}_{\nu}\ \big(f(\cdot) - \lambda c(x,\cdot)\big)\,\mathrm{d}\widehat{\mathbb{P}}(x).$$

When supp$(\nu) = \Omega$,

Sinkhorn DRO

$$\begin{aligned} \sup_{\mathbb{P}} \quad & \mathbb{E}_{z\sim\mathbb{P}}[f(z)] \\ \text{s.t.} \quad & W_\epsilon(\hat{\mathbb{P}}, \mathbb{P}) \le \rho \end{aligned}$$

$$\xrightarrow{\text{Take } \epsilon \to 0}$$

Wasserstein DRO

$$\begin{aligned} \sup_{\mathbb{P}} \quad & \mathbb{E}_{z\sim\mathbb{P}}[f(z)] \\ \text{s.t.} \quad & W(\hat{\mathbb{P}}, \mathbb{P}) \le \rho \end{aligned}$$

Optimization Algorithm for Sinkhorn DRO

- Based on strong duality,

$$\min_{\theta\in\Theta}\sup_{\mathbb{P}} \left\{\mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)] : \quad W_\varepsilon(\widehat{\mathbb{P}},\mathbb{P}) \le \rho\right\}$$

$$= \min_{\lambda\ge 0}\left\{\lambda\overline{\rho} + \underbrace{\min_{\theta\in\Theta}\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lambda\varepsilon\log\left(\mathbb{E}_{z\sim\mathbb{Q}_x}\left[e^{f_\theta(z)/(\lambda\varepsilon)}\right]\right)\right]}_{V(\lambda)}\right\}$$

- Solve the Monte-Carlo approximated formulation[3]:

$$V(\lambda) \approx \min_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\lambda\varepsilon\log\left(\frac{1}{m}\sum_{j=1}^{m}e^{f_\theta(z_{i,j})/(\lambda\varepsilon)}\right),$$

where $\{\hat{x}_i\}_{i=1}^n \sim \widehat{\mathbb{P}}$ and $\{z_{i,j}\}_{j=1}^m$ are i.i.d. samples generated from $\mathbb{Q}_{\hat{x}_i}$.
- Cons: It requires $\tilde{O}(\delta^{-3})$ samples to obtain δ-optimal solution.

[3] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

Optimization Algorithm for Sinkhorn DRO

- Based on strong duality,

$$\min_{\theta\in\Theta}\sup_{\mathbb{P}} \left\{\mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)] : \quad W_\varepsilon(\widehat{\mathbb{P}},\mathbb{P}) \le \rho\right\}$$

$$= \min_{\lambda\ge 0}\left\{\lambda\overline{\rho} + \underbrace{\min_{\theta\in\Theta}\ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lambda\varepsilon\log\left(\mathbb{E}_{z\sim\mathbb{Q}_x}\left[e^{f_\theta(z)/(\lambda\varepsilon)}\right]\right)\right]}_{V(\lambda)}\right\}$$

- Solve the Monte-Carlo approximated formulation[3]:

$$V(\lambda) \approx \min_{\theta\in\Theta}\ \frac{1}{n}\sum_{i=1}^{n}\lambda\varepsilon\log\left(\frac{1}{m}\sum_{j=1}^{m}e^{f_\theta(z_{i,j})/(\lambda\varepsilon)}\right),$$

where $\{\hat{x}_i\}_{i=1}^{n} \sim \widehat{\mathbb{P}}$ and $\{z_{i,j}\}_{j=1}^{m}$ are i.i.d. samples generated from $\mathbb{Q}_{\hat{x}_i}$.
- **Cons:** It requires $\tilde{O}(\delta^{-3})$ samples to obtain δ-optimal solution.

[3] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

Optimization Algorithm for Sinkhorn DRO: Biased Gradient Update

- Goal: to solve the optimization

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

- Biased gradient update: for each iteration t,
  - Construct a gradient estimate[4] of $F(\theta_t)$, denoted as $v(\theta_t)$;
  - Update $\theta_{t+1}=\textbf{Proximal}_{\theta_t}\big(\gamma_t v(\theta_t)\big)$.

Estimator of solution: randomly selected from (or average over) $\{\theta_t\}_{t=1}^T$

[4] Yifan Hu, Xin Chen, and Niao He. "On the Bias-Variance-Cost Tradeoff of Stochastic Optimization". In: *Advances in Neural Information Processing Systems*. Dec. 2021.

Optimization Algorithm for Sinkhorn DRO: Biased Gradient Update

- Goal: to solve the optimization

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

- Biased gradient update: for each iteration t,
  - Construct a gradient estimate[4] of $F(\theta_t)$, denoted as $v(\theta_t)$;
  - Update $\theta_{t+1} = \textbf{Proximal}_{\theta_t}\left(\gamma_t v(\theta_t)\right)$.

Estimator of solution: randomly selected from (or average over) $\{\theta_t\}_{t=1}^T$

[4] Yifan Hu, Xin Chen, and Niao He. "On the Bias-Variance-Cost Tradeoff of Stochastic Optimization". In: *Advances in Neural Information Processing Systems*. Dec. 2021.

Optimization Algorithm for Sinkhorn DRO: Biased Gradient Update

- Goal: to solve the optimization

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

- Biased gradient update: for each iteration t,
  - Construct a gradient estimate[4] of $F(\theta_t)$, denoted as $v(\theta_t)$;
  - Update $\theta_{t+1} = \textbf{Proximal}_{\theta_t}\big(\gamma_t v(\theta_t)\big)$.

Estimator of solution: randomly selected from (or average over) $\{\theta_t\}_{t=1}^T$

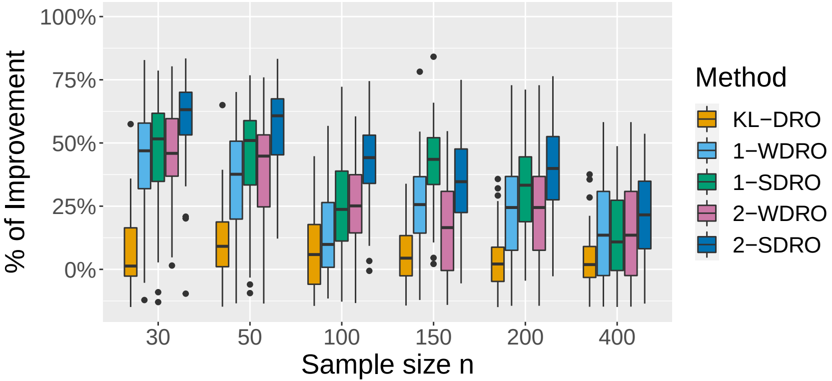
Estimators	**Convex Nonsmooth**	**Convex Smooth**	**Nonconvex Smooth**
Vanilla SGD	$O(\delta^{-3})$	$O(\delta^{-3})$	$O(\delta^{-6})$
V-MLMC	N/A	$\tilde{O}(\delta^{-2})$	$\tilde{O}(\delta^{-4})$
RT-MLMC	N/A	$\tilde{O}(\delta^{-2})$	$\tilde{O}(\delta^{-4})$

[4] Yifan Hu, Xin Chen, and Niao He. "On the Bias-Variance-Cost Tradeoff of Stochastic Optimization". In: *Advances in Neural Information Processing Systems*. Dec. 2021.

Numerical Results

Portfolio Optimization:

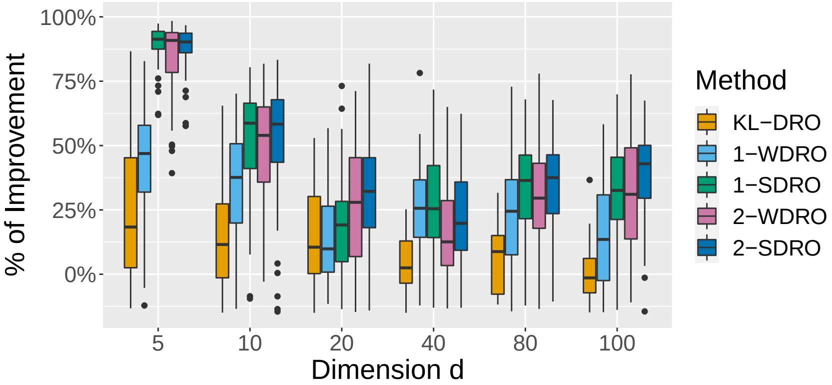
$$\begin{aligned} \inf_x \quad & \mathbb{E}_{\mathbb{P}_*}[-\langle x,\zeta\rangle] + \rho \cdot \mathbb{P}_*\text{-CVaR}_\alpha(-\langle x,\zeta\rangle) \\ \text{s.t.} \quad & x \in \mathscr{X} = \{x \in \mathbb{R}_+^D : \; x^\mathrm{T} 1 = 1\}. \end{aligned}$$



Numerical Results

Portfolio Optimization:

$$\begin{aligned} \inf_{x} \quad & \mathbb{E}_{\mathbb{P}_*}[-\langle x,\zeta\rangle] + \rho\cdot\mathbb{P}_*\text{-CVaR}_\alpha(-\langle x,\zeta\rangle) \\ \text{s.t.} \quad & x\in\mathscr{X} = \{x\in\mathbb{R}_+^D:\ x^{\mathrm{T}}1 = 1\}. \end{aligned}$$



Numerical Simulation Results

Semi-supervised Learning:

- Train classifiers based on data with labels and without labels;
- Two performance measures:
  - Training error for data without labels;
  - Testing error.

	SAA	Sinkhorn	Wasserstein	KL-divergence
Breast Cancer	$.20 \pm .068$ $.19 \pm .073$	$\mathbf{.12 \pm .068}$ $\mathbf{.11 \pm .067}$	$.17 \pm .073$ $.17 \pm .075$	$.19 \pm .038$ $.19 \pm .073$
Magic	$.28 \pm .082$ $.28 \pm .064$	$\mathbf{.25 \pm .091}$ $\mathbf{.25 \pm .074}$	$.27 \pm .077$ $.27 \pm .058$	$.26 \pm .078$ $.27 \pm .066$
QSAR Bio	$.25 \pm .057$ $.25 \pm .062$	$\mathbf{.22 \pm .063}$ $\mathbf{.22 \pm .065}$	$.23 \pm .073$ $.23 \pm .079$	$.25 \pm .037$ $.25 \pm .042$
Spambase	$.19 \pm .038$ $.19 \pm .032$	$\mathbf{.14 \pm .046}$ $\mathbf{.14 \pm .036}$	$.16 \pm .036$ $.16 \pm .028$	$.18 \pm .034$ $.18 \pm .042$

Take Home Message

Sinkhorn DRO is a great notion of DRO models:

- Inherit geometric properties from optimal transport;
- Absolutely continuous worst-case distribution thanks to entropic regularization;
- Improve the out-of-sample performance of Wasserstein DRO;
- Optimization by Monte Carlo approximation and first order method;
- More applications in operations research with Sinkhorn DRO can be explored!

Sinkhorn Distributionally Robust Optimization

To be Submitted to Operations Research – INFORMS PUBs

Online Available: arxiv.org/abs/2109.11926

arXiv

SCAN ME