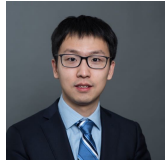
Entropic Regularization for (Wasserstein) Robust Optimization

Speaker: Jie Wang

Date: June 2, 2023



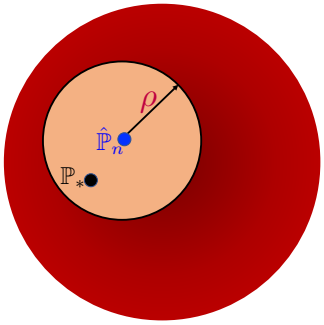
Rui Gao
UT Austin

Yao Xie
Gatech

Wasserstein DRO

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

Distance: $W(\mathbb{P}, \mathbb{Q}) = \min_{\gamma \in \mathcal{P}(\Omega^2)} \left\{ \mathbb{E}_{(\omega,\omega') \sim \gamma}[c(\omega, \omega')] : \ \mathrm{Proj}_{\#1}\gamma = \mathbb{P}, \mathrm{Proj}_{\#2}\gamma = \mathbb{Q} \right\}$.



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

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

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

Tractability of Wasserstein DRO

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

$$=\inf_{\theta\in\Theta,\lambda\ge 0}\left\{\lambda\rho+\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}}\left[f_\theta(z)-\lambda c(x,z)\right]\right]\right\}$$

References	Loss function $f_\theta(z)$	Transport cost	Nominal distribution $\widehat{\mathbb{P}}$	Support $\mathcal{Z}$
[Pflug G et. al 2008, …]	General	General	General	Discrete and finite set
[Esfahani PM et. al 2018, …]	Piecewise concave in z	Norm function	Empirical distribution	Polytope
[Shafieezade et. al 2015, …]	Generalized linear model in (z,θ)	Norm function	Empirical distribution	Whole Euclidean space
[Sinha et. al 2018, …]	$z\mapsto f_\theta(z)-\lambda^* c(x,z)$ is strongly concave	Strongly convex function	General	General

Sinkhorn Distance

- Sinkhorn Distance:

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

  Remark: Sinkhorn distance does not satisfy definition of "distance function".

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

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

Historical Review:

- Originally proposed by [Wilson' 62].
- Convergence of algorithm for the first time by [Sinkhorn'64].
- Operation complexity analysis and GPGPU parallel by [Cuturi'13].

Sinkhorn Distance

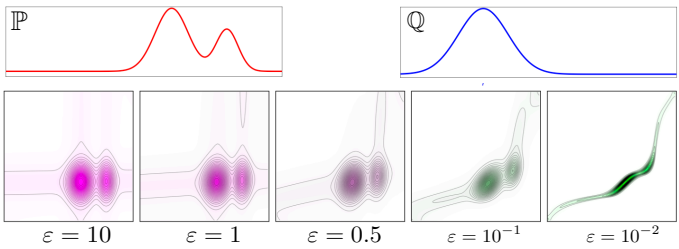
- Sinkhorn Distance:

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

Remark: Sinkhorn distance does not satisfy definition of "distance function".

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

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

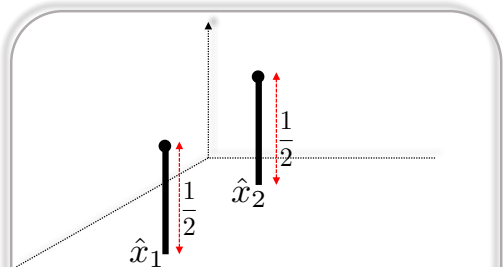


Main Framework

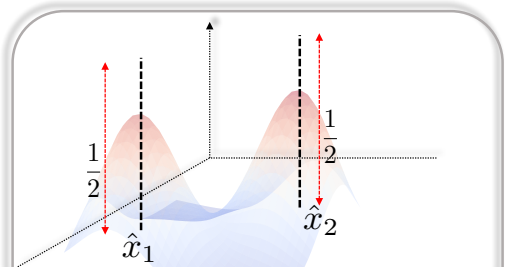
- Sinkhorn DRO:

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

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



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



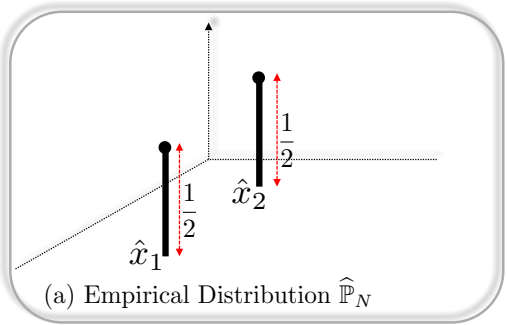
(d) Worst-case Distribution $\widetilde{\mathbb{P}}$ from $\mathbb{B}^{\text{Sinkhorn}}_{\rho,\varepsilon}(\widehat{\mathbb{P}}_N)$.

General DRO Models

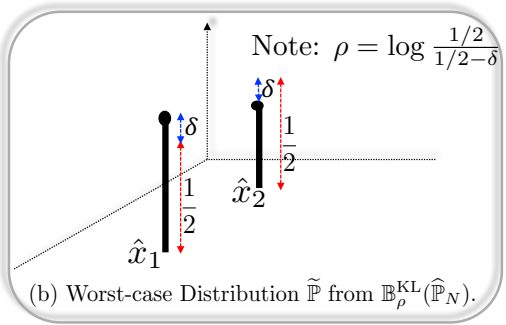
- KL-DRO:

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

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



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



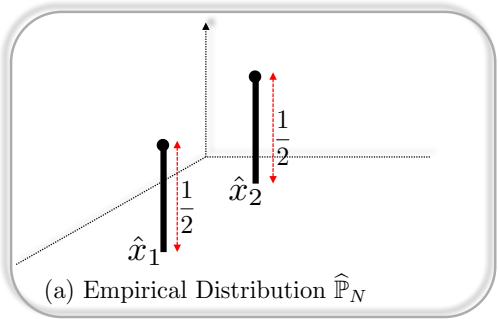
(b) Worst-case Distribution $\widetilde{\mathbb{P}}$ from $\mathbb{B}_{\rho}^{\mathrm{KL}}(\widehat{\mathbb{P}}_N)$.

General DRO Models

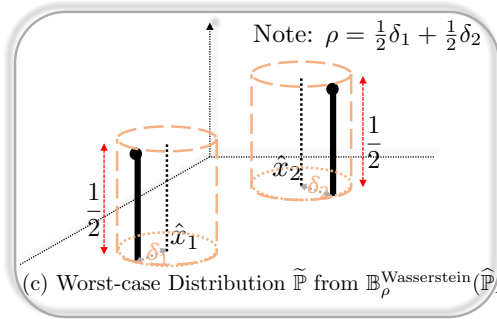
- Wasserstein-DRO:

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

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



(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:

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

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

  - Duality Formulation and Properties
  - Algorithm with Optimal Sample Complexity
  - Numerical Results

Tractable Formulation

Under mild conditions, the primal

$$V_{\mathrm{P}} = \sup_{\mathbb{P} \in \mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})} \mathbb{E}_{z\sim\mathbb{P}}[f(z)], \quad \text{where } \mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}}) = \big\{\mathbb{P} :\ W_\epsilon(\widehat{\mathbb{P}}, \mathbb{P}) \le \rho\big\}.$$

(Sinkhorn DRO)

has the **strong dual reformulation**:

$$V_{\mathrm{D}} = \inf_{\lambda \ge 0} \left\{ \lambda\overline{\rho} + \lambda\epsilon\, \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\big[e^{f(z)/(\lambda\epsilon)}\big]\Big] \right\},$$

where

$$\overline{\rho} = \rho + \epsilon\, \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\log \mathbb{E}_{z\sim\nu}\big[e^{-c(x,z)/\epsilon}\big]\Big],$$

$$\mathrm{d}\mathbb{Q}_{x,\epsilon}(z) = \frac{e^{-c(x,z)/\epsilon}}{\mathbb{E}_{u\sim\nu}\left[e^{-c(x,u)/\epsilon}\right]}\, \mathrm{d}\nu(z).$$

Interpretation of Worst-case Distribution

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

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

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

  where α_x is the normalizing constant.
- Closed-form expression on $\tilde{\mathbb{P}}$:

$$\frac{\mathrm{d}\tilde{\mathbb{P}}(z)}{\mathrm{d}\nu(z)} = \mathbb{E}_{x\sim\widehat{\mathbb{P}}} \Big[\alpha_x \cdot \exp\Big(\big(f(z) - \lambda^* c(x,z)\big)/(\lambda^*\epsilon) \Big) \Big].$$

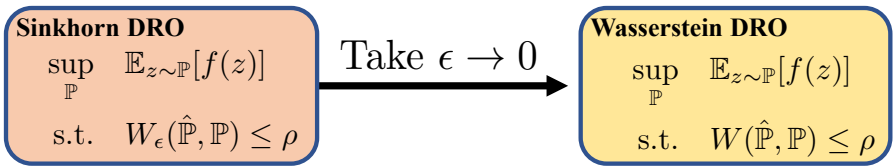
Worst-case distribution $\tilde{\mathbb{P}}$ support on whole space, while W-DRO is discrete.

Connection of Sinkhorn DRO with Wasserstein DRO

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

$$\lambda\rho + \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\text{ess-sup}_{\nu}\ \big(f(\cdot) - \lambda c(x,\cdot)\big)\Big].$$

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



Connection of Sinkhorn DRO with KL DRO

Upper bound of Sinkhorn DRO:

$$
\begin{aligned}
V_{\mathrm{D}} &\triangleq \inf_{\lambda>0} \ \lambda\overline{\rho} + \lambda\epsilon\, \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\Big] \\
&\leq \inf_{\lambda>0} \ \lambda\overline{\rho} + \lambda\epsilon \log \mathbb{E}_{y\sim\mathbb{P}^0}\left[e^{f(y)/(\lambda\epsilon)}\right]
\end{aligned}
$$

$\mathbb{P}^0$: kernel density estimate based on $\widehat{\mathbb{P}}$:

$$
\mathrm{d}\mathbb{P}^0(z) = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}[\,\mathrm{d}\mathbb{Q}_x(z)].
$$

Sinkhorn DRO

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

Take $\mathbb{P}^0$ as the KDE estimate of $\hat{\mathbb{P}}$

KL DRO

$$
\begin{aligned}
\sup_{\mathbb{P}} \quad & \mathbb{E}_{z\sim\mathbb{P}}[f(z)] \\
\text{s.t.} \quad & D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{P}^0) \leq \bar{\rho}/\epsilon
\end{aligned}
$$

Connection of Sinkhorn DRO with SAA

$$V_{\mathrm{P}} = \sup_{\gamma_x \in \mathcal{P}(\mathcal{Z}), x \in \mathrm{supp}(\widehat{\mathbb{P}})} \left\{ \mathbb{E}_{x\sim\widehat{\mathbb{P}}} \mathbb{E}_{z\sim\gamma_x}[f(z)] : \ \epsilon \mathbb{E}_{x\sim\widehat{\mathbb{P}}} \Big[D_{\mathsf{KL}}(\gamma_x \| \mathbb{Q}_{x,\epsilon}) \Big] \leq \overline{\rho} \right\}$$

When $\overline{\rho} = 0$, Sinkhorn DRO becomes SAA:

$$V_{\mathrm{P}} = \mathbb{E}_{z\sim\mathbb{P}^0}[f(z)]$$

$\mathbb{P}^0$: kernel density estimate based on $\widehat{\mathbb{P}}$:

$$\mathrm{d}\mathbb{P}^0(z) = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}[\,\mathrm{d}\mathbb{Q}_x(z)].$$

Sinkhorn DRO

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

Take $\bar{\rho} = 0$

SAA

$$\mathbb{E}_{z\sim\mathbb{P}^0}[f(z)]$$

Optimization Algorithm for Sinkhorn DRO

- Based on strong duality,

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

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

- Solve the Monte-Carlo approximated formulation [Shapiro et. al 2014]:

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

  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 $\mathbb{Q}_{\hat{x}_i,\epsilon}$.
- **Cons:** It requires $\tilde{O}(\delta^{-3})$ samples to obtain δ-optimal solution [Yifan et. al SIAMOP2020].

Optimization Algorithm for Sinkhorn DRO

- Based on strong duality,

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

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

- Solve the Monte-Carlo approximated formulation [Shapiro et. al 2014]:

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

  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 $\mathbb{Q}_{\hat{x}_i,\epsilon}$.

- **Cons:** It requires $\tilde{O}(\delta^{-3})$ samples to obtain δ-optimal solution [Yifan et. al SIAMOP2020].

Optimization Algorithm for Sinkhorn DRO: Biased Gradient Update

- Goal:

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

- Biased gradient update: for each iteration t,
  - Construct a gradient estimate 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: average (or randomly selected) over $\{\theta_t\}_{t=1}^T$.
- **Remark:** optimally pick gradient estimator to balance bias versus (2nd)moment (or variance by [Hu, Chen and He 2021]) trade-off.

Optimization Algorithm for Sinkhorn DRO: Biased Gradient Update

- Goal:

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

- Biased gradient update: for each iteration t,
  - Construct a gradient estimate 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: average (or randomly selected) over $\{\theta_t\}_{t=1}^T$.

- **Remark:** optimally pick gradient estimator to balance bias versus (2nd)moment (or variance by [Hu, Chen and He 2021]) trade-off.

Optimization Algorithm for Sinkhorn DRO: Biased Gradient Update

- Goal:

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

- Biased gradient update: for each iteration t,
  - Construct a gradient estimate 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: average (or randomly selected) over $\{\theta_t\}_{t=1}^T$.
- **Remark:** optimally pick gradient estimator to balance bias versus (2nd)moment (or variance by [Hu, Chen and He 2021]) trade-off.

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

Bias-(2nd)Moment-Cost Trade-off for SMD

- Consider convex optimization problem

$$\begin{array}{ll} \text{Minimize} & F(\theta) \\ \text{s.t.} & \theta \in \Theta \subseteq \mathbb{R}^d. \end{array}$$

- Stochastic Mirror Descent: iteratively,
  - Step 1: generate random vector $v(\theta_t)$ with

$$\mathbb{E}[v(\theta_t)] = \nabla \overline{F}(\theta_t), \quad \Delta_F := \sup_{\theta \in \Theta} |\overline{F}(\theta) - F(\theta)|, \quad \mathbb{E}\left[\|v(\theta_t)\|^2\right] \leq M^2.$$

  - Step 2: $\theta_{t+1} = \textbf{Proximal}_{\theta_t}\big(\gamma v(\theta_t)\big)$.
- Take $\widehat{\theta}_{1:T}$ as average over $\{\theta_t\}_{t=1}^T$, then

$$\mathbb{E}\big[F(\widehat{\theta}_{1:T}) - F(\theta^*)\big] \leq c \cdot \left(\Delta_F + \sqrt{\frac{M^2}{T}}\right).$$

Gradient Estimators

- Goal:

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

- Construct a sequence of approximation functions $\{F^\ell(\theta)\}_{\ell\geq 0}$ instead, where

$$F^\ell(\theta) = \mathbb{E}_{x^\ell\sim\widehat{\mathbb{P}}} \, \mathbb{E}_{\{z_j^\ell\}_{j\in[2^\ell]}|x^\ell} \left[\lambda\epsilon \log \left(\frac{1}{2^\ell} \sum_{j\in[2^\ell]} \exp\left(\frac{f_\theta(z_j^\ell)}{\lambda\epsilon} \right) \right) \right].$$

Remark: generating unbiased gradient estimate of $F^\ell(\theta)$ is easy!

Gradient Estimators

- Goal:

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

- Construct a sequence of approximation functions $\{F^\ell(\theta)\}_{\ell\geq 0}$ instead, where

$$F^\ell(\theta) = \mathbb{E}_{x^\ell\sim\widehat{\mathbb{P}}} \, \mathbb{E}_{\{z_j^\ell\}_{j\in[2^\ell]}|x^\ell} \left[\lambda\epsilon \log \left(\frac{1}{2^\ell} \sum_{j\in[2^\ell]} \exp\left(\frac{f_\theta(z_j^\ell)}{\lambda\epsilon} \right) \right) \right].$$

Remark: generating unbiased gradient estimate of $F^\ell(\theta)$ is easy!

Vanilla SGD Estimator

Sample $x^L \sim \widehat{\mathbb{P}}$ and next sample $\{z_j^L\}_{j\in[2^L]} \sim \mathbb{Q}_{x^L,\epsilon}$. Construct

$$v^L(\theta) = \nabla_\theta \left\{ \lambda\epsilon \log\left(\frac{1}{2^L} \sum_{j\in[2^L]} \exp\left(\frac{f_\theta(z_j^L)}{\lambda\epsilon} \right) \right) \right\}.$$

Pros:

- Low-bias $\Delta_F = \mathcal{O}(2^{-L})$.
- Bounded moment $M^2 = \mathcal{O}(1)$.

Cons:

- Generating a single gradient estimator has cost $\mathcal{O}(2^L)$.

Overall:

- Sample complexity to get δ-optimal solution is $\mathcal{O}(\delta^{-3})$.

Algorithm Improvement

- Directly computing $v^L(\theta)$ for large L seems expensive;
- Define $v^{-1}(\theta) \equiv 0$ and rewrite

$$\begin{aligned} v^L(\theta) &= \sum_{\ell=0}^{L} \left[v^\ell(\theta) - v^{\ell-1}(\theta) \right] \\ &= \sum_{\ell=0}^{L} p_\ell \cdot \frac{v^\ell(\theta) - v^{\ell-1}(\theta)}{p_\ell} = \mathbb{E}_{\ell \sim \{p_\ell\}_{\ell=0}^{L}} \left[\frac{v^\ell(\theta) - v^{\ell-1}(\theta)}{p_\ell} \right] \end{aligned}$$

- **Randomized Sampling Gradient Estimator**: sample ℓ from truncated geometric distribution $\{p_\ell\}_{\ell=0}^{L}$ with $p_\ell \propto 2^{-\ell}$. Construct

$$v^{\text{RT-MLMC}}(\theta) = \frac{1}{p_\ell} \cdot \left[v^\ell(\theta) - v^{\ell-1}(\theta) \right].$$

Bias, 2nd Moment, and Costs

- **Bias**: For same level L, the bias of RT-MLMC/Vanilla SGD are same.
- **2nd Moment**: $v^{\ell}(\theta) - v^{\ell-1}(\theta) \to 0$ for large ℓ:

$$\mathbb{E}\left[\|v^{\text{RT-MLMC}}(\theta_t)\|^2\right] = \mathcal{O}(L) = \tilde{\mathcal{O}}(1).$$

- **Sampling Cost**: Cost for generating RT-MLMC estimator reduces from $\mathcal{O}(2^L)$ to $\mathcal{O}(L)$!

The sample complexity of RT-MLMC is $\tilde{O}(\delta^{-2})$ with storage cost $\tilde{O}(1)$.

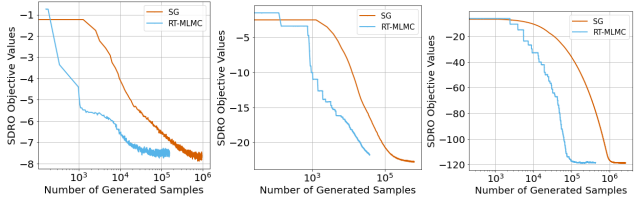
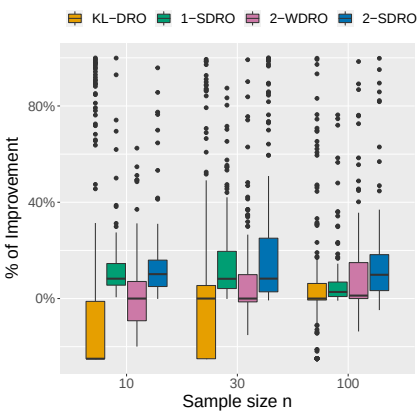
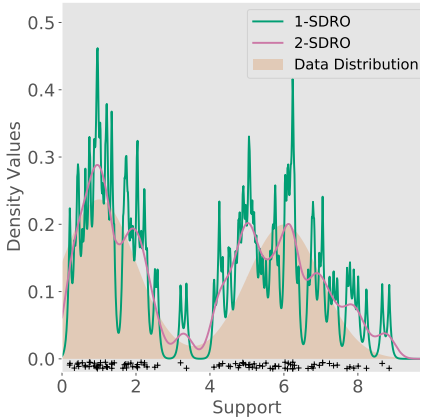


Figure: Plot for Portfolio optimizaiton with $d \in \{50, 100, 400\}$.

Newsvendor Model

$$\min_{\theta} \max_{\mathbb{P}\in\mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})} \mathbb{E}_{z\sim\mathbb{P}}\big[k\theta - u\min(\theta, z)\big].$$





Newsvendor Model

$$\min_{\theta} \max_{\mathbb{P}\in\mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})} \mathbb{E}_{z\sim\mathbb{P}}\big[k\theta - u\min(\theta, z)\big].$$

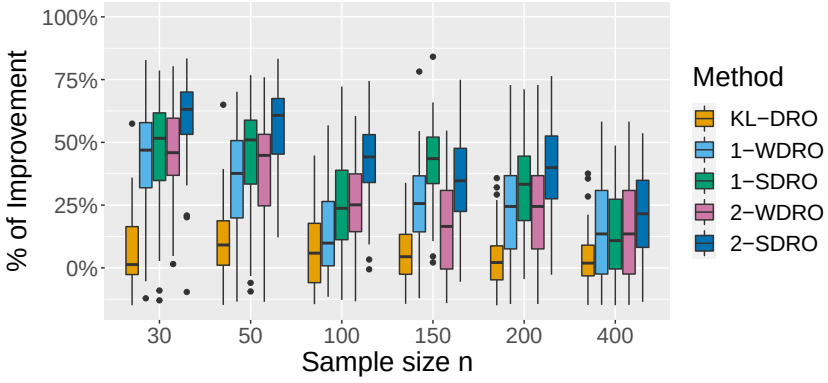
Average computational time (in seconds) per problem instance:

Model	Exponential			Gamma			Gaussian Mixture		
	$n=10$	$n=30$	$n=100$	$n=10$	$n=30$	$n=100$	$n=10$	$n=30$	$n=100$
SAA	0.017	0.017	0.018	0.019	0.019	0.019	0.023	0.024	0.024
KL-DRO	0.027	0.029	0.040	0.027	0.028	0.039	0.027	0.028	0.038
1-SDRO	0.110	0.124	0.161	0.105	0.119	0.162	0.105	0.119	0.157
2-WDRO	0.123	0.358	1.307	0.128	0.354	1.337	0.134	0.402	1.428
2-SDRO	0.061	0.069	0.106	0.100	0.121	0.161	0.095	0.115	0.154

Mean-risk Portfolio Optimization

$$\min_{\theta} \max_{\mathbb{P}\in\mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})} \quad \mathbb{E}_{z\sim\mathbb{P}}[-\theta^{\mathrm{T}} z] + \varrho \cdot \mathbb{P}\text{-CVaR}_{\alpha}(-\theta^{\mathrm{T}} z)$$

$$\text{s.t.} \quad \theta \in \Theta = \{\theta \in \mathbb{R}^{d}_{+} : \theta^{\mathrm{T}} 1 = 1\},$$



Mean-risk Portfolio Optimization

$$\min_{\theta} \max_{\mathbb{P}\in\mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})} \quad \mathbb{E}_{z\sim\mathbb{P}}[-\theta^{\mathrm{T}} z] + \varrho \cdot \mathbb{P}\text{-CVaR}_{\alpha}(-\theta^{\mathrm{T}} z)$$

$$\text{s.t.} \quad \theta \in \Theta = \{\theta \in \mathbb{R}^{d}_{+} : \theta^{\mathrm{T}} 1 = 1\},$$

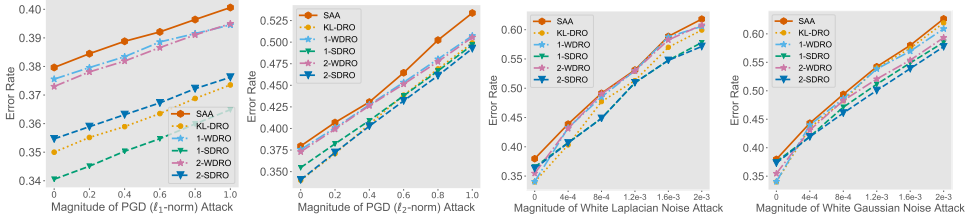
Average computational time (in seconds) per problem instance:

(n, d) Values	SAA	KL-DRO	1-WDRO	1-SDRO	2-WDRO	2-SDRO
$(30, 30)$	0.013	0.038	0.018	0.125	0.015	0.090
$(50, 30)$	0.014	0.042	0.020	0.163	0.016	0.110
$(100, 30)$	0.017	0.065	0.024	0.167	0.021	0.144
$(150, 30)$	0.019	0.084	0.029	0.158	0.027	0.152
$(200, 30)$	0.023	0.115	0.035	0.151	0.032	0.150
$(400, 30)$	0.045	0.136	0.061	0.167	0.056	0.142

Adversarial Multi-class Logistic Regression

$$\min_{B\in\mathbb{R}^{d\times C}} \max_{\mathbb{P}\in\mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})} \mathbb{E}_{(x,\boldsymbol{y})\sim\mathbb{P}}\big[h_B(x,\boldsymbol{y})\big], \quad h_B(x,\boldsymbol{y}) = -\boldsymbol{y}^{\mathrm{T}} B^{\mathrm{T}} x + \log\left(1^{\mathrm{T}} e^{B^{\mathrm{T}} x}\right).$$

- We solve WDRO formulations using gradient-descent-ascent heuristic in [Sinha et. al 2018].
- Error rate for tinyImageNet dataset (90000 training samples with dimension 512):



- Computation Time:

Dataset	SAA	KL-DRO	1-WDRO	1-SDRO	2-WDRO	2-SDRO
tinyImageNet	45.54	44.50	325.25	227.91	347.16	197.55

Extension (I): Entropy Regularization for ∞-Wasserstein DRO

- Original formulation:

$$\begin{aligned}
&\min_{\theta\in\Theta} \left\{ \sup_{\mathbb{P}} \mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)] : \ W_\infty(\mathbb{P},\widehat{\mathbb{P}}) \le \rho \right\} \\
=&\min_{\theta\in\Theta} \sup_{\mathbb{P},\gamma} \left\{ \sup_{\mathbb{P}} \mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)] : \begin{array}{l} \mathrm{Proj}_{1\#\gamma} = \widehat{\mathbb{P}}, \mathrm{Proj}_{2\#\gamma} = \mathbb{P}, \\ \mathrm{ess.sup}_\gamma d(\zeta_1,\zeta_2) \le \rho \end{array} \right\}
\end{aligned}$$

- Formulation with entropic regularization [Chiara et. al 2023]:

$$\begin{aligned}
&\min_{\theta\in\Theta} \sup_{\mathbb{P},\gamma} \left\{ \sup_{\mathbb{P}} \mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)] - \eta H(\gamma \mid \pi) : \begin{array}{l} \mathrm{Proj}_{1\#\gamma} = \widehat{\mathbb{P}}, \mathrm{Proj}_{2\#\gamma} = \mathbb{P}, \\ \mathrm{ess.sup}_\gamma d(\zeta_1,\zeta_2) \le \rho \end{array} \right\} \\
=&\min_{\theta\in\Theta} \mathbb{E}_{x\sim\widehat{\mathbb{P}}} \left[\eta \log \mathbb{E}_{z\sim\mathbb{Q}_x} \left[\exp\left(\frac{f_\theta(z)}{\eta} \right) \right] \right], \quad \text{where } \frac{\mathrm{d}\mathbb{Q}_x}{\mathrm{d}\nu_x}(z) = \mathsf{vol}(\nu_x)^{-1}.
\end{aligned}$$

Extension (II): Bilevel Optimization for Sinkhorn DRO

- Reformulation:

$$\begin{aligned}&\min_{\theta} \ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lambda\epsilon\log\mathbb{E}_{z\sim\mathbb{Q}_x}\left[e^{f_\theta(z)/(\lambda\epsilon)}\right]\right]\\ =&\min_{\theta} \ -\lambda\epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\max_{u\leq 0}\left\{u\mathbb{E}_{z\sim Q_{x,\epsilon}}\left[e^{f_\theta(z)/(\lambda\epsilon)}\right]+1+\log(-u)\right\}.\end{aligned}$$

- Contextual bilevel optimization[1]:

$$\begin{aligned}&\min_{x\in\mathbb{R}^{d_x}} \quad F(x):=\mathbb{E}_{\xi\sim\mathbb{P}_\xi,\eta\sim\mathbb{P}_{\eta|\xi}}[f(x,y^*(x;\xi);\eta,\xi)] && \text{(upper level)}\\ &\text{where} \quad y^*(x;\xi):=\operatorname{argmin}_{y\in\mathbb{R}^{d_y}}\mathbb{E}_{\eta\sim\mathbb{P}_{\eta|\xi}}[g(x,y;\eta,\xi)] \quad \forall\ \xi \text{ and } x. && \text{(lower-level)}\end{aligned}$$

Yifan Hu, MS132, Thursday, June 1, 11:45-12:10

Contextual Stochastic Bilevel Optimization and Its Applications to Wasserstein DRO with Side Information

[1] Yifan Hu, Jie Wang, Yao Xie, Andreas Krause, Daniel Kuhn. "Contextual Stochastic Bilevel Optimization". In: (2023). To be released soon.

Sinkhorn Distributionally Robust Optimization

- Winner of 2022 Informs Best Poster Award

Submitted to Operations Research

Online Available: https://arxiv.org/abs/2109.11926

(originally posted in Sep 2021, Updated on May 2023)

Code: https://github.com/WalterBabyRudin/SDRO_code

Website: https://walterbabyrudin.github.io/

Contact: jwang3163@gatech.edu