Sinkhorn Distributionally Robust Optimization

Jie Wang

The Chinese University of Hong Kong, Shenzhen

Joint work with Rui Gao (UT Austin) and Yao Xie (Georgia Tech)

George Nicholson Student Paper Competition





人工智能學院
School of Artificial Intelligence

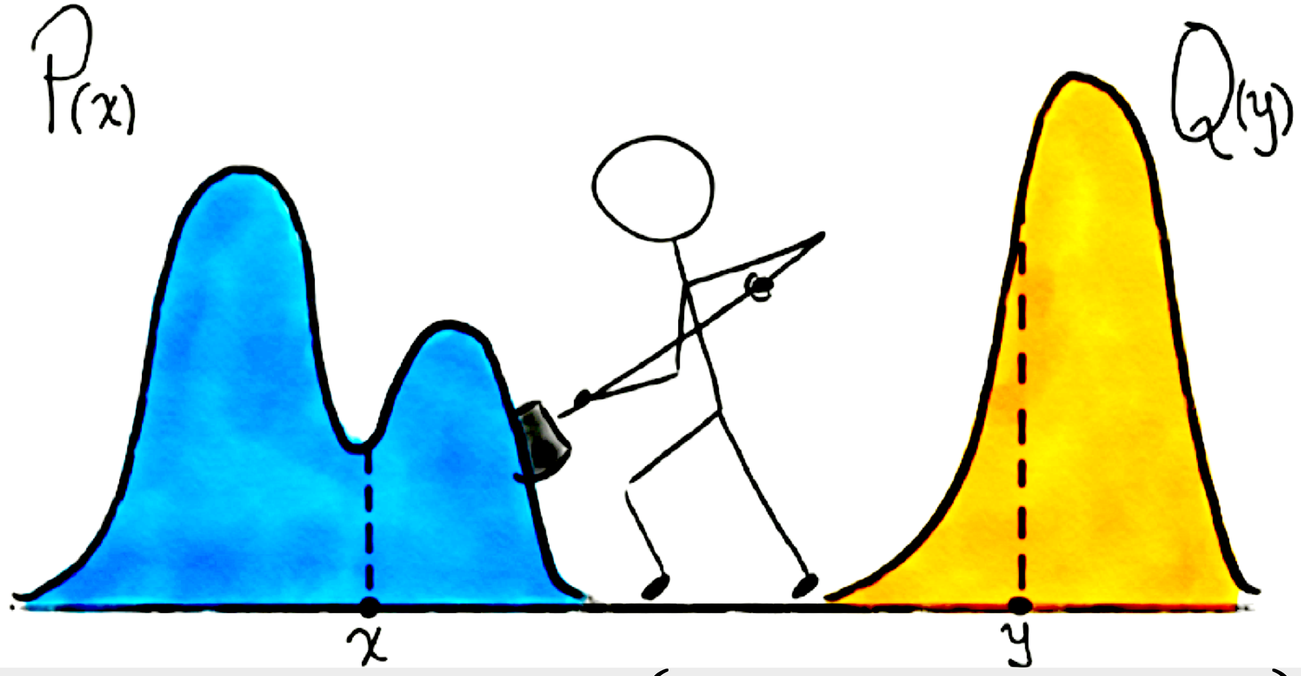


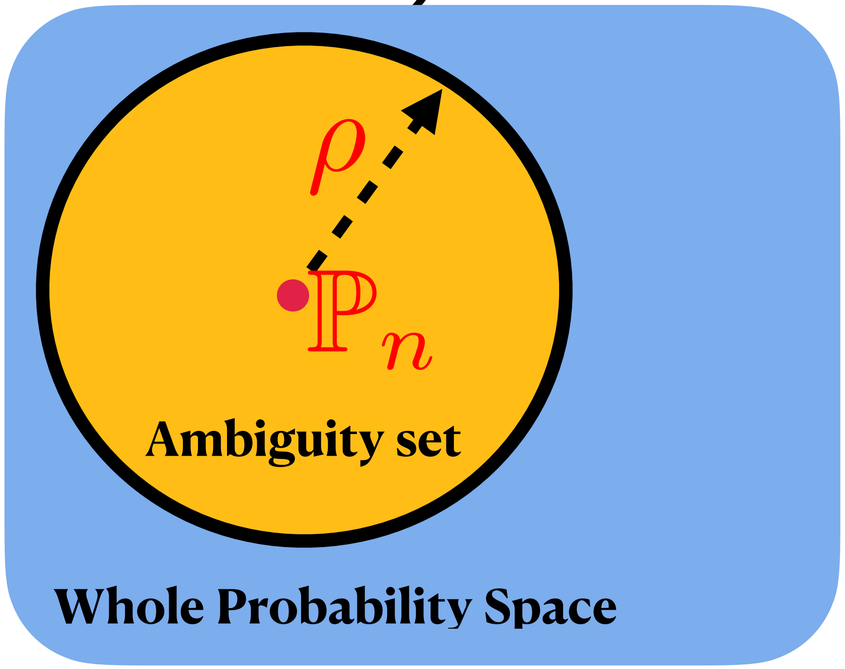
SCHOOL OF
DATA SCIENCE
數據科學學院

1. Introduction

Wasserstein Distributionally Robust Optimization

$$\min_{\theta} \left\{ \sup_{\mathbb{P}:\ \mathcal{W}(\mathbb{P},\mathbb{P}_n)\leq\rho} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] \right\}$$





$$\mathcal{W}(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{ \mathbb{E}_{(x,y)\sim\gamma}\left[\|x-y\|^2\right] \right\}$$

- data-driven, non-parametric, free of distributional assumptions

Tractability of Wasserstein DRO

$$\min_{\theta} \left\{ \sup_{\mathbb{P}:\ \mathcal{W}(\mathbb{P},\mathbb{P}_n)\le\rho} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] \right\}$$

$$= \min_{\theta,\lambda\ge 0} \left\{ \lambda\rho + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\underbrace{\sup_{z} \left\{ \ell(z;\theta) - \lambda\|x-z\|^2 \right\}}_{\text{Moreau-Yoshida regularization}} \right] \right\}$$ **(Strong Dual Reformulation)**

1. Probability support is *discrete* and *finite* [Pflug G et. al 2008, ...]
2. Loss $\ell(z;\theta)$ is *piecewise concave / generalized linear model* [Esfahani PM et. al 2018, Shafieezade et al 2015, ...]
3. $z \mapsto \ell(z;\theta) - \lambda^*\|x-z\|^2$ is *strongly concave* [Sinha et. al 2018, ...]

Tractability of Wasserstein DRO

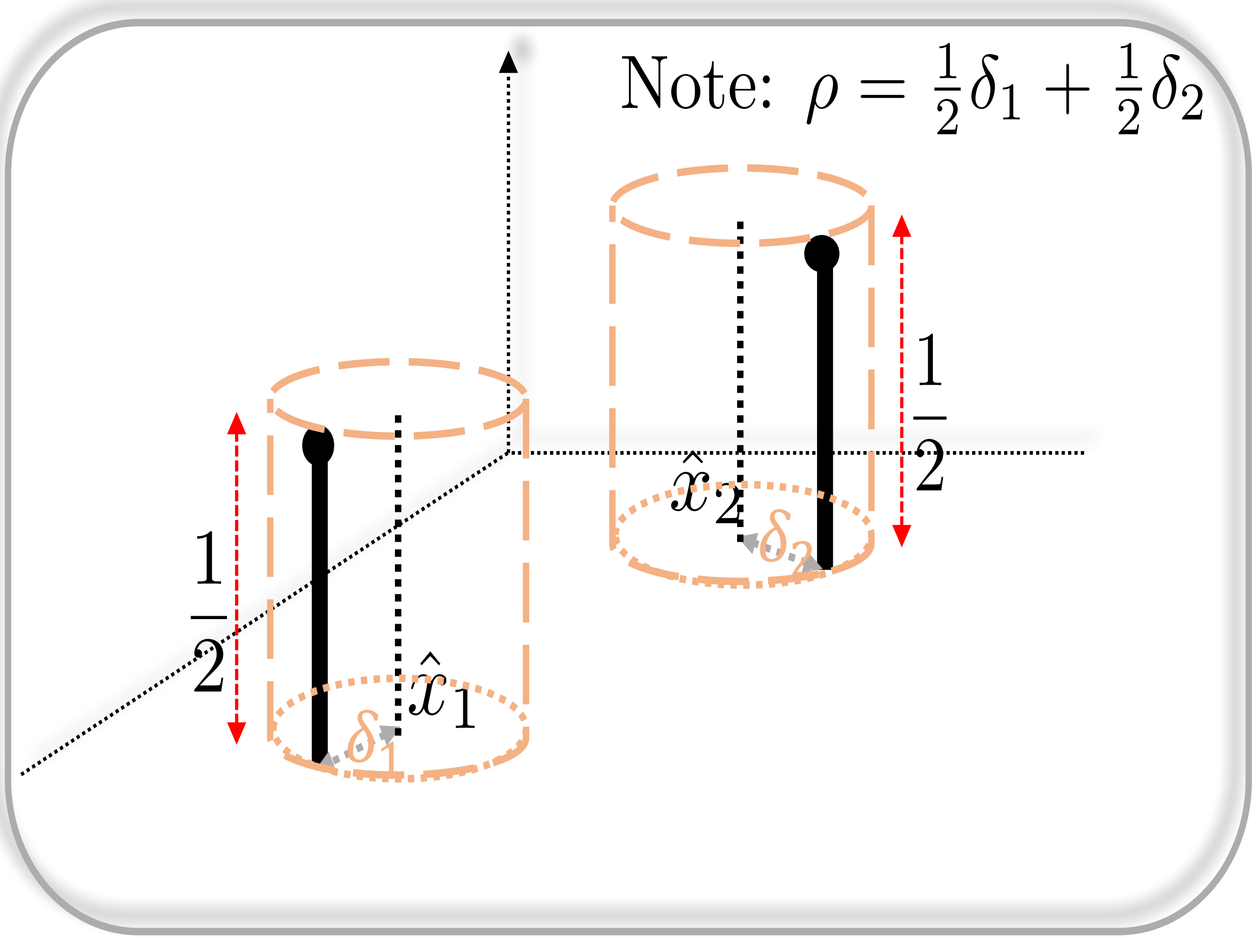
$$\min_{\theta} \left\{ \sup_{\mathbb{P}:\ \mathcal{W}(\mathbb{P},\mathbb{P}_n)\le\rho} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] \right\}$$

$$= \min_{\theta,\lambda\ge 0} \left\{ \lambda\rho + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\underbrace{\sup_{z} \left\{ \ell(z;\theta) - \lambda\|x-z\|^2 \right\}}_{\text{Moreau-Yoshida regularization}} \right] \right\}$$ **(Strong Dual Reformulation)**

1. Probability support is *discrete* and *finite* [Pflug G et. al 2008, ...]
2. Loss $\ell(z;\theta)$ is *piecewise concave / generalized linear model* [Esfahani PM et. al 2018, Shafieezade et al 2015, ...]
3. $z \mapsto \ell(z;\theta) - \lambda^*\|x-z\|^2$ is *strongly concave* [Sinha et. al 2018, ...]

Cons: Wasserstein DRO is not necessarily tractable for general applications

Worst-case Distribution of Wasserstein DRO



- Hedges against discrete distributions; [Blanchet et. al 2019, Gao et. al 2023]
- For certain applications (news vendor, non-parametric contextual learning, etc), share the same support as empirical distribution [Zhang et. al 2024, Yang et. al 2026]

Central Question

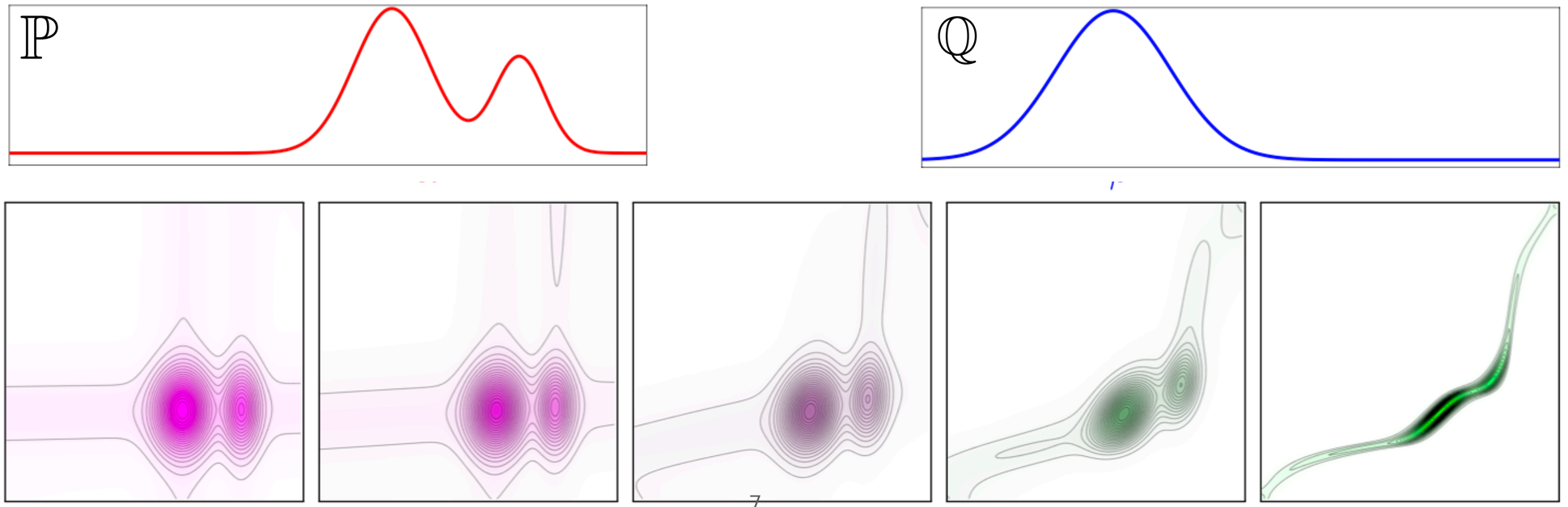
Is it possible to derive a data-driven DRO model such that

1. Its **tractability** applies to general loss function;
2. Its **worst-case distribution** is absolutely continuous over the sample space.

Sinkhorn Discrepancy

$$\mathcal{W}_\epsilon(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{ \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|^2] + \epsilon\mathbb{E}_{(x,y)\sim\gamma}\left[\log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\gamma(x)\mathrm{d}y}\right)\right]\right\}$$

- It does not satisfy the definition of "distance"
- Entropic regularization encourages moving each $x \in \operatorname{supp}\mathbb{P}$ to whole space



Sinkhorn Discrepancy

$$\mathcal{W}_\epsilon(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{ \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|^2] + \epsilon\mathbb{E}_{(x,y)\sim\gamma}\left[\log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\gamma(x)\mathrm{d}y}\right)\right]\right\}$$

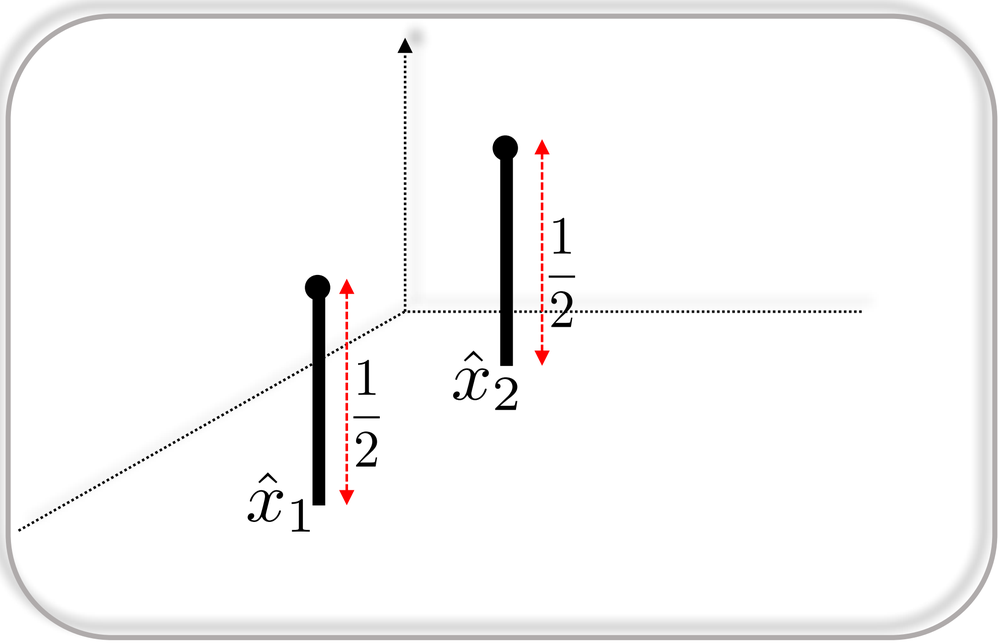
Historical Review:

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

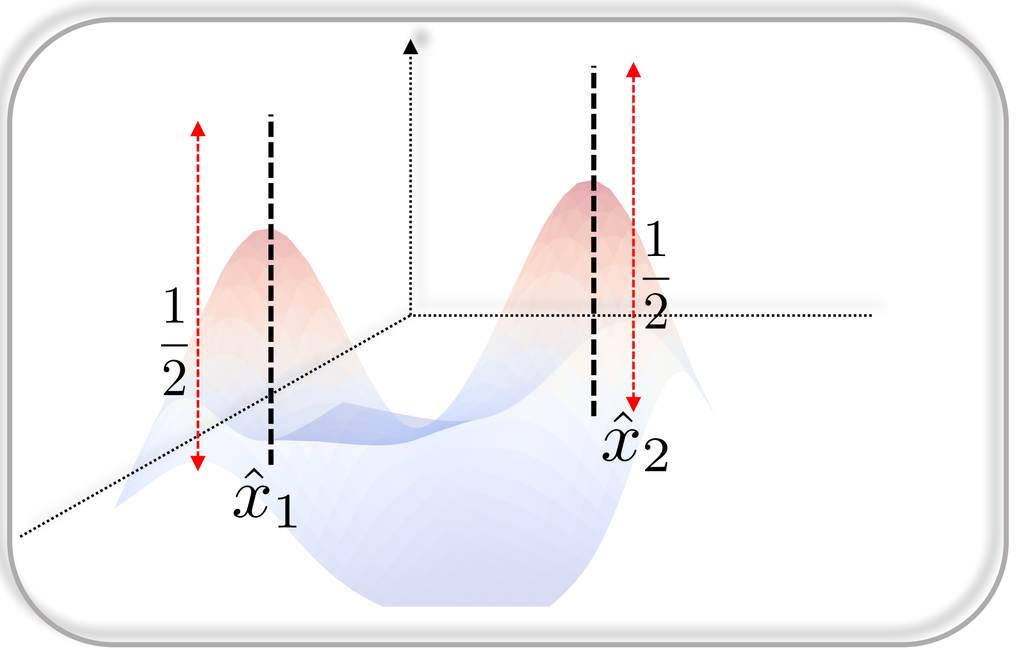
Main Framework

Sinkhorn DRO: $\min_{\theta} \left\{ \sup_{\mathbb{P}:\ \mathcal{W}_\epsilon(\mathbb{P}_n, \mathbb{P}) \le \rho} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] \right\}$

$$\mathcal{W}_\epsilon(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{ \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|^2] + \epsilon \mathbb{E}_{(x,y)\sim\gamma}\left[\log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\gamma(x)\mathrm{d}y}\right)\right] \right\}$$

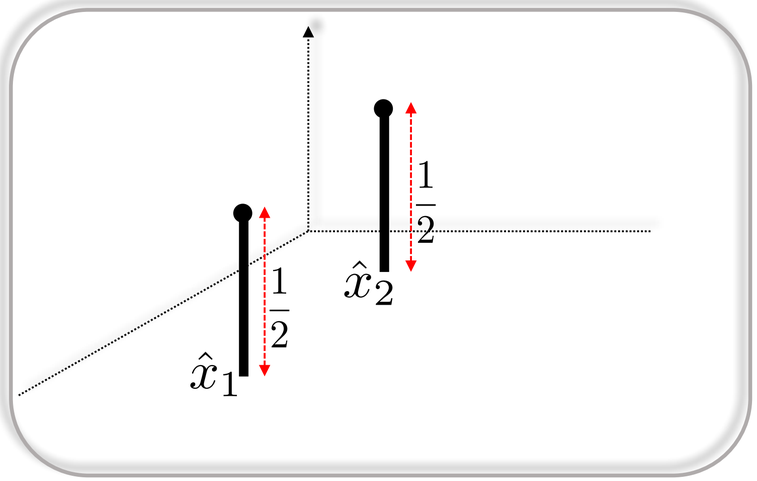


Empirical Distribution

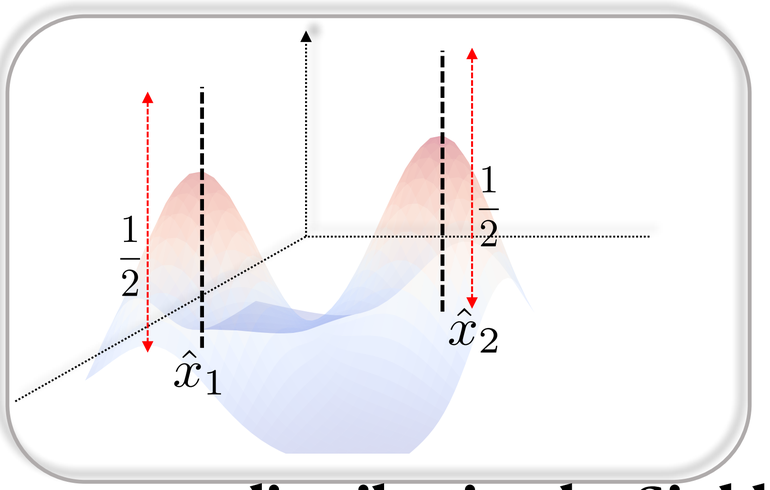


Worst-case distribution by Sinkhorn DRO

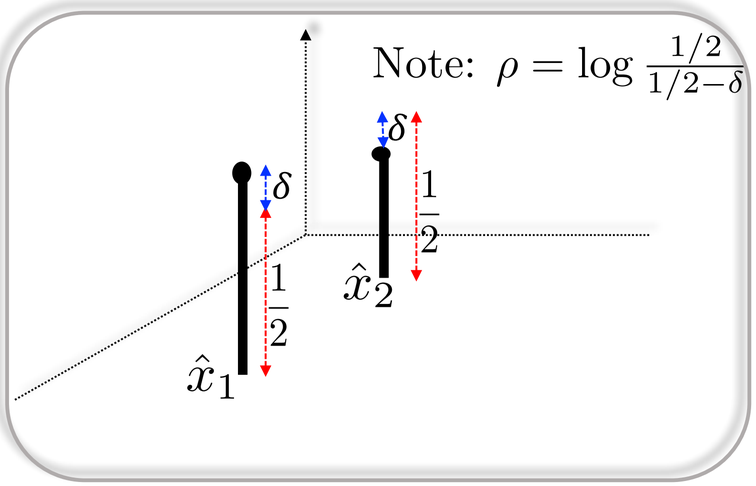
Comparison



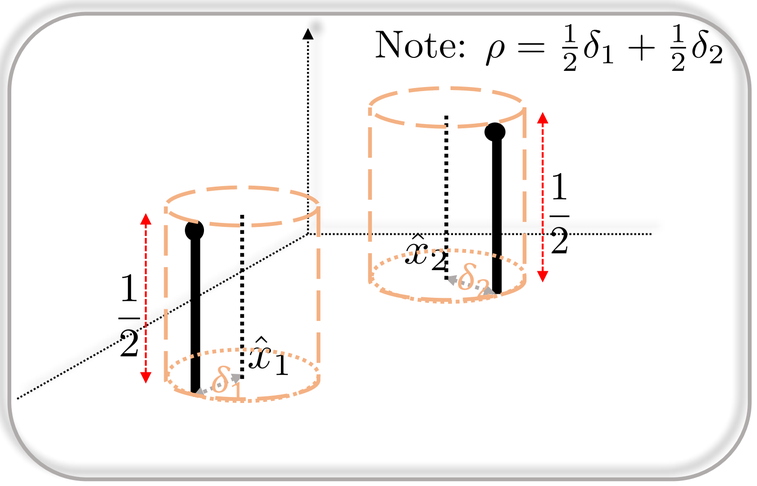
Empirical Distribution



Worst-case distribution by Sinkhorn DRO



Worst-case distribution by KL DRO



Worst-case distribution by Wasserstein DRO

2. Strong Duality and Related Properties

Strong Dual Reformulation

Under mild conditions, $V_{\text{Primal}}=V_{\text{dual}}$:

$$V_{\text{Primal}} = \sup_{\mathbb{P}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] : \ \mathcal{W}_\epsilon(\mathbb{P}_n, \mathbb{P}) \le \rho \right\}$$

$$V_{\text{Dual}} = \inf_{\lambda\ge 0} \left\{ \lambda\bar{\rho} + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbf{N}(x,\epsilon\mathbf{I})}[e^{\ell(z;\theta)/(\lambda\epsilon)}]\right] \right\}$$

- $\bar{\rho} = \rho + \epsilon\mathbb{E}_{x\sim\mathbb{P}_n}[\log(\int e^{-\|x-z\|^2/\epsilon}\mathrm{d}z)]$
- V_{dual} : **One-dimensional convex** minimization, **conditional stochastic optimization**

Recovery of Worst-case Distribution

$$\mathbb{P}^* = \arg\max_{\mathbb{P}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] : \ \mathcal{W}_\epsilon(\mathbb{P}_n, \mathbb{P}) \leq \rho \right\}$$

$$\frac{\mathrm{d}\mathbb{P}^*(z)}{\mathrm{d}z} = \mathbb{E}_{x\sim\mathbb{P}_n}\left[\alpha_x \cdot \exp\left(\frac{\ell(z;\theta)}{\lambda^*\epsilon} - \frac{\|x-z\|^2}{\epsilon}\right)\right]$$

Normalizing Constant

Density contributed by x

- Worst-case distribution supported on whole sample space, while W-DRO is discrete

Comparison with Wasserstein DRO

Strong duality for **Sinkhorn DRO**:

$$V_{\mathrm{Primal}} = \sup_{\mathbb{P}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] : \ \mathcal{W}_\epsilon(\mathbb{P}_n, \mathbb{P}) \leq \rho \right\}$$

$$V_{\mathrm{Dual}} = \inf_{\lambda\geq 0} \left\{ \lambda\bar{\rho} + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbf{N}(x,\epsilon\mathbf{I})}[e^{\ell(z;\theta)/(\lambda\epsilon)}]\right] \right\}$$

Comparison with Wasserstein DRO

Strong duality for **Sinkhorn DRO**:

$$V_{\mathrm{Primal}} = \sup_{\mathbb{P}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] : \ \mathcal{W}_\epsilon(\mathbb{P}_n, \mathbb{P}) \le \rho \right\}$$

$$V_{\mathrm{Dual}} = \inf_{\lambda\ge 0} \left\{ \lambda\bar{\rho} + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbf{N}(x,\epsilon\mathbf{I})}[e^{\ell(z;\theta)/(\lambda\epsilon)}] \right] \right\}$$

Strong duality for **Wasserstein DRO** ($\epsilon = 0$):

$$V_{\mathrm{Primal}} = \sup_{\mathbb{P}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] : \ \mathcal{W}(\mathbb{P}_n, \mathbb{P}) \le \rho \right\}$$

$$V_{\mathrm{Dual}} = \inf_{\lambda\ge 0} \left\{ \lambda\rho + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\sup_z \left\{ \ell(z;\theta) - \lambda\|x - z\|^2 \right\} \right] \right\}$$

Comparison with Wasserstein DRO

Strong duality for **Sinkhorn DRO**:

$$V_{\text{Primal}} = \sup_{\mathbb{P}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] : \ \mathcal{W}_\epsilon(\mathbb{P}_n, \mathbb{P}) \le \rho \right\}$$

$$V_{\text{Dual}} = \inf_{\lambda\ge 0} \left\{ \lambda\bar{\rho} + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbf{N}(x,\epsilon\mathbf{I})}[e^{\ell(z;\theta)/(\lambda\epsilon)}] \right] \right\}$$

Approximate "sup" using log-sum-exp

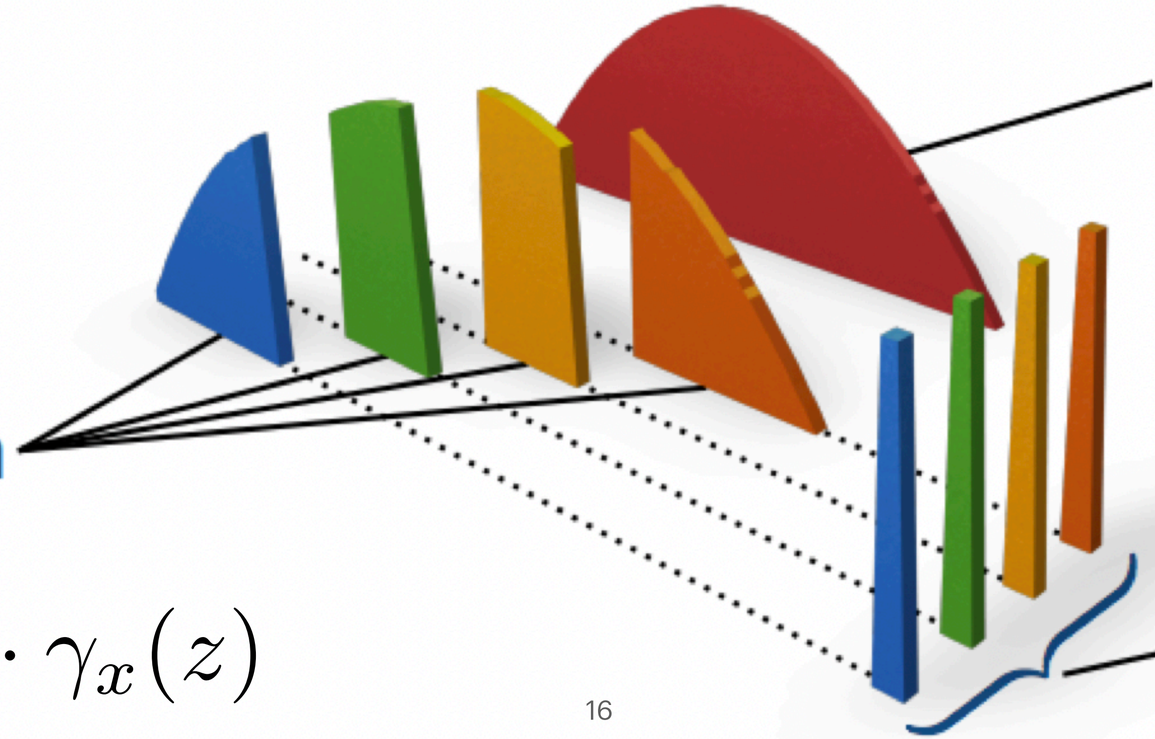
Strong duality for **Wasserstein DRO** ($\epsilon = 0$):

$$V_{\text{Primal}} = \sup_{\mathbb{P}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] : \ \mathcal{W}(\mathbb{P}_n, \mathbb{P}) \le \rho \right\}$$

$$V_{\text{Dual}} = \inf_{\lambda\ge 0} \left\{ \lambda\rho + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\sup_z \left\{ \ell(z;\theta) - \lambda\|x - z\|^2 \right\} \right] \right\}$$

Comparison with KL-divergence DRO

$$V_{\text{Primal}} = \begin{cases} \sup\limits_{\gamma_x, \forall x} \mathbb{E}_{x\sim\mathbb{P}_n} \mathbb{E}_{z\sim\gamma_x}[\ell(z;\theta)] \\ \text{s.t.} \quad \mathbb{E}_{x\sim\mathbb{P}_n}\Big[\mathrm{KL}\big(\gamma_x \| \mathbf{N}(x, \epsilon\mathbf{I})\big)\Big] \leq \bar{\rho}/\epsilon \end{cases}$$



$\mathbb{P}$ = Worst-case Distribution

$$\mathbb{P}(z) = \frac{1}{n}\sum_{i=1}^{n} \gamma_{\hat{x}_i}(z)$$

γ = transportation plan

$$\gamma(x, z) = \mathbb{P}_n(x) \cdot \gamma_x(z)$$

$\mathbb{P}_n$= empirical distribution

Comparison with KL-divergence DRO

$$V_{\text{Primal}} = \begin{cases} \sup\limits_{\gamma_x, \forall x} \mathbb{E}_{x\sim\mathbb{P}_n} \mathbb{E}_{z\sim\gamma_x}[\ell(z;\theta)] \\ \text{s.t.} \quad \mathbb{E}_{x\sim\mathbb{P}_n}\Big[\text{KL}\big(\gamma_x \| \mathbf{N}(x, \epsilon\mathbf{I})\big)\Big] \leq \bar{\rho}/\epsilon \end{cases}$$

1. When $\bar{\rho} = 0$, Sinkhorn DRO becomes SAA with kernel density estimation:

$$V_{\text{Primal}} = \mathbb{E}_{z\sim\mathbb{P}^0}[\ell(z;\theta)], \qquad \mathbb{P}^0 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{N}(x, \epsilon\mathbf{I})$$

2. When $\bar{\rho} > 0$, Sinkhorn DRO robustifies $\mathbb{P}^0$ in terms of KL-divergence.

3. Optimization Algorithms

Monte Carlo Sampling

- **Ideal formulation:**

$$\min_{\theta,\lambda\geq 0} \left\{ \lambda\bar{\rho} + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbf{N}(x,\epsilon\mathbf{I})}\left[e^{\ell(z;\theta)/(\lambda\epsilon)}\right]\right] \right\}$$

Monte Carlo Sampling

- **Ideal formulation:**

$$\min_{\theta,\lambda\geq 0} \left\{ \lambda\bar{\rho} + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbf{N}(x,\epsilon\mathbf{I})}\left[e^{\ell(z;\theta)/(\lambda\epsilon)}\right]\right] \right\}$$

"As long as you can sample from $\mathbb{P}_n$ and $\mathbf{N}(x, \epsilon\mathbf{I})$, the problem is solved".

- A. Shapiro

Monte Carlo Sampling

- **Ideal formulation:**

$$\min_{\theta,\lambda\geq 0} \left\{ \lambda\bar{\rho} + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbf{N}(x,\epsilon\mathbf{I})}\left[e^{\ell(z;\theta)/(\lambda\epsilon)}\right]\right] \right\}$$

"As long as you can sample from $\mathbb{P}_n$ and $\mathbf{N}(x, \epsilon\mathbf{I})$, the problem is solved".

- A. Shapiro

For each $\hat{x}_i$ in $\mathbb{P}_n$,

sample m i.i.d. samples $\{z_{i,j}\}_{j=1}^{m}$ from $\mathbf{N}(\hat{x}_i, \epsilon\mathbf{I})$



$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \lambda\epsilon \log\left(\frac{1}{m}\sum_{j=1}^{m} e^{\ell(z_{i,j};\theta)/(\lambda\epsilon)}\right)$$

Monte Carlo Sampling

- **Ideal formulation:**

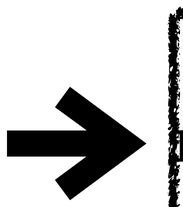
$$\min_{\theta,\lambda\geq 0} \left\{ \lambda\bar{\rho} + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbf{N}(x,\epsilon\mathbf{I})}\left[e^{\ell(z;\theta)/(\lambda\epsilon)}\right]\right] \right\}$$

"As long as you can sample from $\mathbb{P}_n$ and $\mathbf{N}(x, \epsilon\mathbf{I})$, the problem is solved".

- A. Shapiro

For each $\hat{x}_i$ in $\mathbb{P}_n$,

sample m i.i.d. samples $\{z_{i,j}\}_{j=1}^m$ from $\mathbf{N}(\hat{x}_i, \epsilon\mathbf{I})$



$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \lambda\epsilon \log\left(\frac{1}{m}\sum_{j=1}^{m} e^{\ell(z_{i,j};\theta)/(\lambda\epsilon)}\right)$$

Cons: Sample complexity is sub-optimal, $O(\delta^{-3})$

Stochastic Algorithm for Sinkhorn DRO

- **Goal:**

$$\min_{\theta} \left\{ \mathbb{E}_{x\sim\mathbb{P}_n} \left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbf{N}(x,\epsilon\mathbf{I})} \left[e^{\ell(z;\theta)/(\lambda\epsilon)} \right] \right] \right\} \triangleq F(\theta)$$

- **Approximation Problem:**

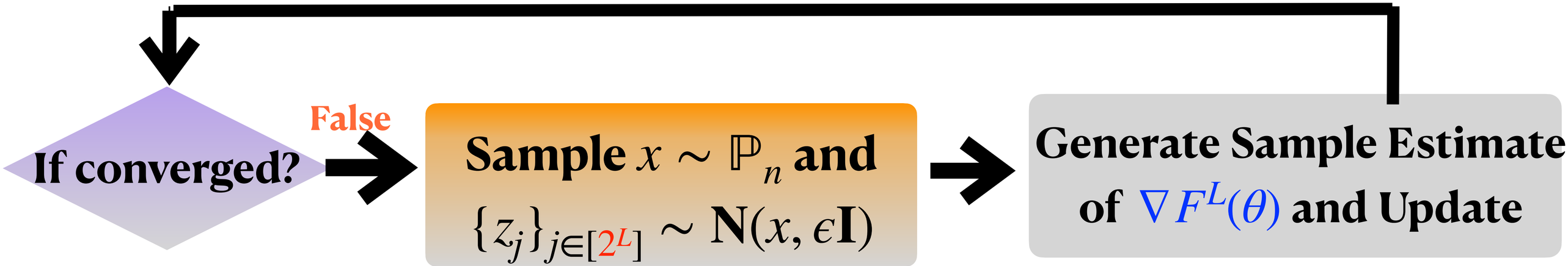
$$\min_{\theta} \left\{ \mathbb{E}_{\substack{x\sim\mathbb{P}_n, \\ \{z_j\}_{j\in[2^l]}\sim\mathbf{N}(x,\epsilon\mathbf{I})}} \left[\lambda\epsilon \log \left(\frac{1}{2^l} \sum_{j\in[2^l]} e^{\ell(z_j;\theta)/(\lambda\epsilon)} \right) \right] \right\} \triangleq F^l(\theta)$$

Stochastic Algorithm for Sinkhorn DRO

- **Approximation Problem:**

$$\min_{\theta} \left\{ \mathbb{E}_{\substack{x\sim\mathbb{P}_n, \\ \{z_j\}_{j\in[2^l]}\sim\mathbf{N}(x,\epsilon\mathbf{I})}} \left[\lambda\epsilon \log\left(\frac{1}{2^l} \sum_{j\in[2^l]} e^{\ell(z_j;\theta)/(\lambda\epsilon)} \right) \right] \right\} \triangleq F^l(\theta)$$

L-SGD: Fix a large $\ell \equiv L$

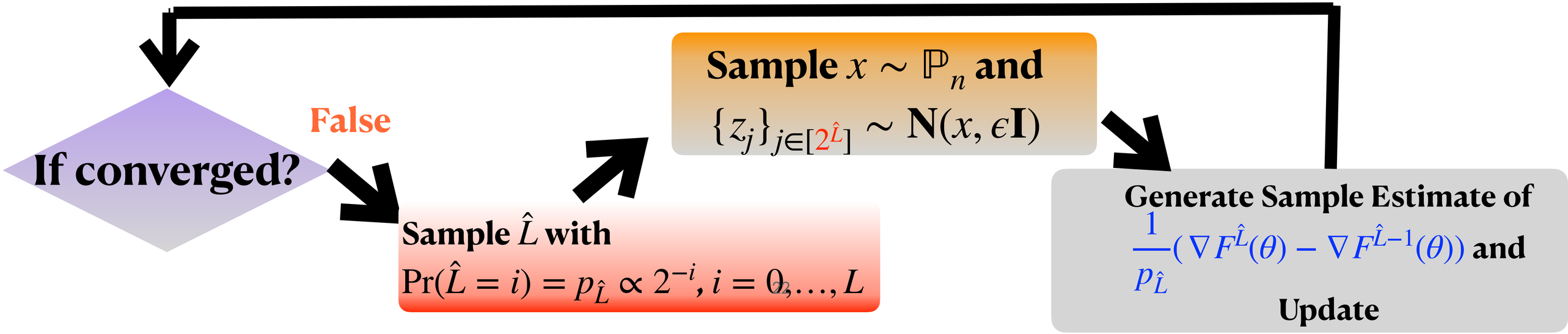


Stochastic Algorithm for Sinkhorn DRO

- **Approximation Problem:**

$$\min_{\theta} \left\{ \mathbb{E}_{\substack{x\sim\mathbb{P}_n, \\ \{z_j\}_{j\in[2^l]}\sim\mathbf{N}(x,\epsilon\mathbf{I})}} \left[\lambda\epsilon \log\left(\frac{1}{2^l} \sum_{j\in[2^l]} e^{\ell(z_j;\theta)/(\lambda\epsilon)} \right) \right] \right\} \triangleq F^l(\theta)$$

SGD with Random Sampling Estimator: Adaptively Choose l



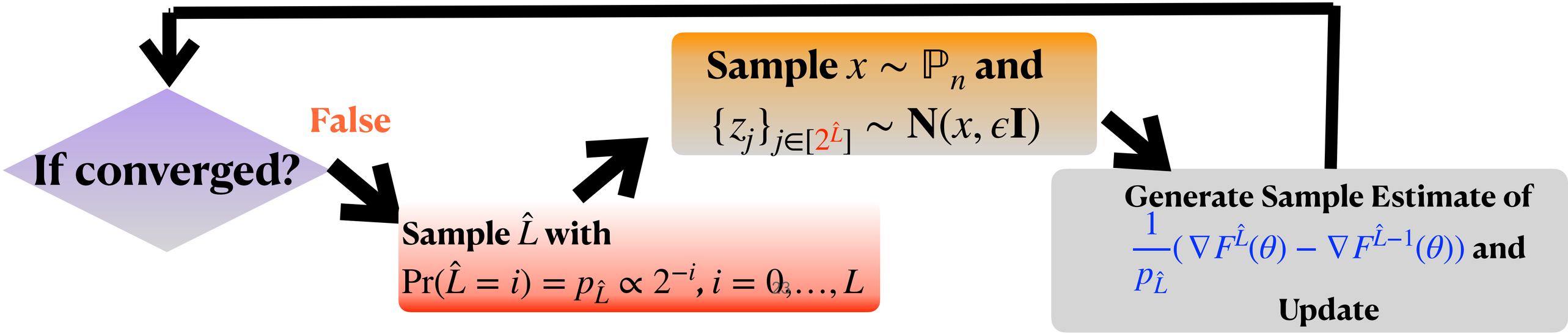
Stochastic Algorithm for Sinkhorn DRO

1. This estimator is **unbiased** gradient estimator of F^L

$$\mathbb{E}_{\hat{L}}\left[\frac{1}{p_{\hat{L}}}(\nabla F^{\hat{L}}(\theta)-\nabla F^{\hat{L}-1}(\theta))\right]=\sum_{\hat{L}=1}^{L}p_{\hat{L}}\cdot\left[\frac{1}{p_{\hat{L}}}(\nabla F^{\hat{L}}(\theta)-\nabla F^{\hat{L}-1}(\theta))\right]$$

$$=\sum_{\hat{L}=1}^{L}\left[\nabla F^{\hat{L}}(\theta)-\nabla F^{\hat{L}-1}(\theta)\right]=\nabla F^{\hat{L}}$$

SGD with Random Sampling Estimator: Adaptively Choose l



Stochastic Algorithm for Sinkhorn DRO

1. This estimator is **unbiased** gradient estimator of F^L
2. This estimator has **significantly lower cost**

Stochastic Algorithm for Sinkhorn DRO

1. This estimator is **unbiased** gradient estimator of F^L
2. This estimator has **significantly lower cost**

Conventional SGD:

- Fix large $l \equiv L$, e.g., $L = 10$
- Each iteration pay $2^L \approx 2^{10}$ cost to get small-biased gradient estimators

Stochastic Algorithm for Sinkhorn DRO

1. This estimator is **unbiased** gradient estimator of F^L
2. This estimator has **significantly lower cost**

Conventional SGD:

- Fix large $l \equiv L$, e.g., $L = 10$
- Each iteration pay $2^L \approx 2^{10}$ cost to get small-biased gradient estimators

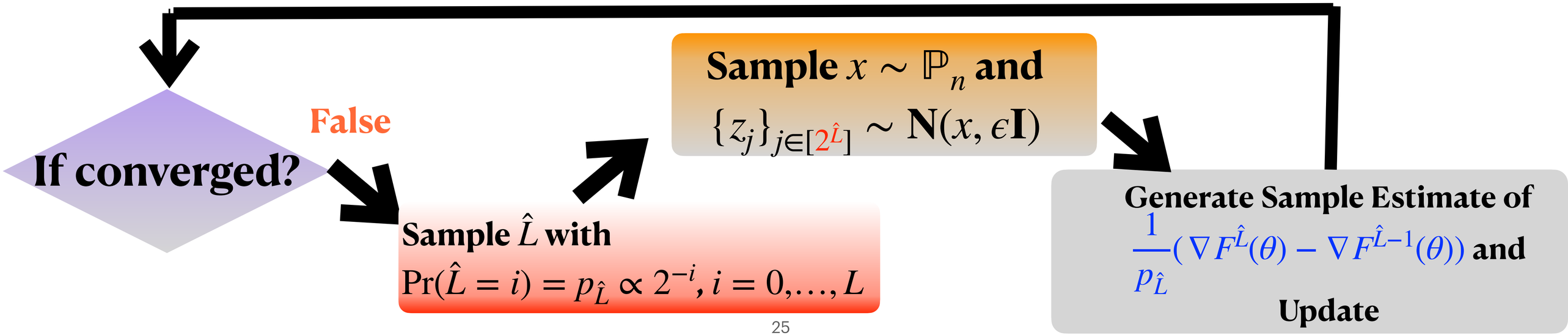
Random Sampling SGD

- Iteratively sample $\widehat{L}$ according to $\mathbb{P}(\widehat{L} = l) \propto 2^{-l}, l = 1, \ldots, L$
- Each iteration pay $2^{\widehat{L}} \approx 10$ cost to get small-biased gradient estimators

Stochastic Algorithm for Sinkhorn DRO

1. This estimator is **unbiased** gradient estimator of F^L
2. This estimator has **significantly lower cost**
3. This estimator has **sufficiently small variance**, due to control variates variance reduction [Nelson, 1990]

SGD with Random Sampling Estimator: Adaptively Choose l



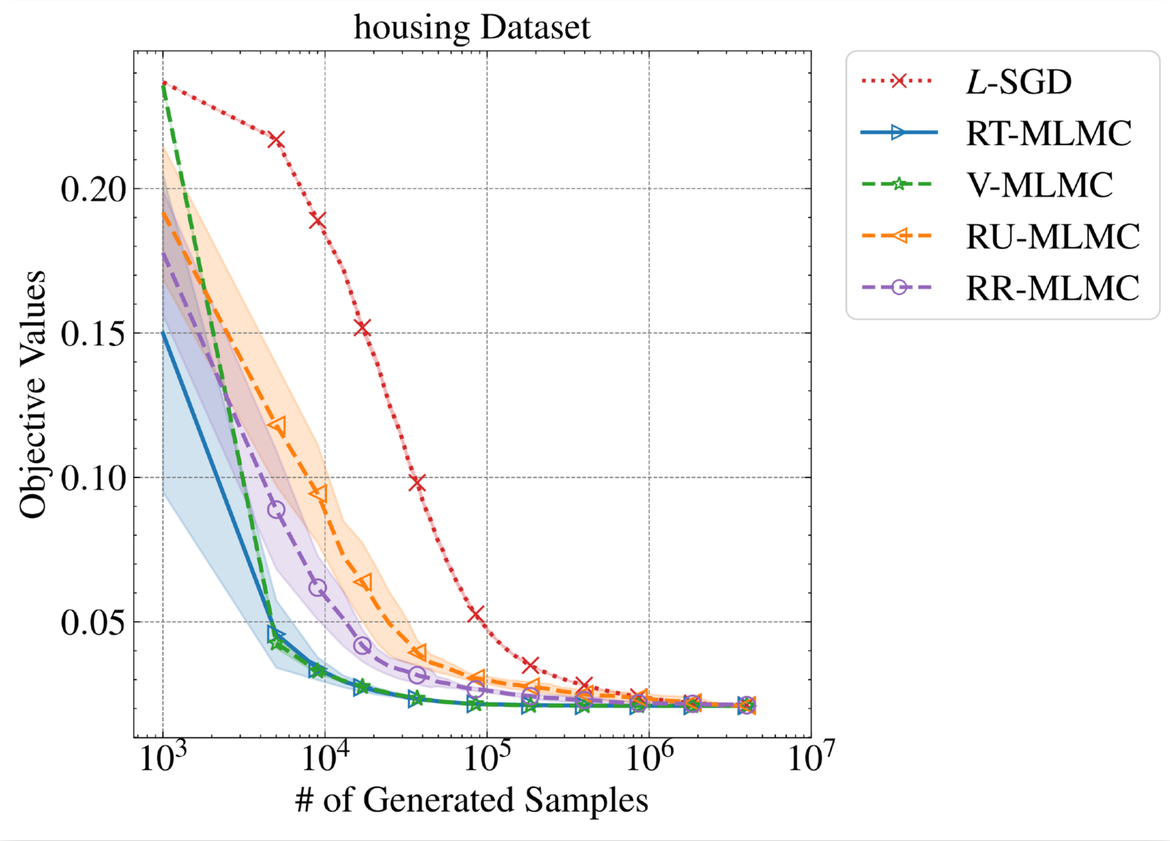
Complexity for Solving Sinkhorn DRO

$$\min_{\theta,\lambda\geq 0}\left\{\lambda\bar{\rho}+\mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon\log\mathbb{E}_{z\sim\mathbf{N}(x,\epsilon\mathbf{I})}\left[e^{\ell(z;\theta)/(\lambda\epsilon)}\right]\right]\right\}$$

Algorithm	**Naive Gradient Estimator**		**Random Sampling Estimator**	
Loss $\ell(z,\cdot)$	Convex	Nonconvex Smooth	Convex	Nonconvex Smooth
Complexity	$\tilde{O}(\delta^{-3})$	$\tilde{O}(\delta^{-6})$	$\tilde{O}(\delta^{-2})$	$\tilde{O}(\delta^{-4})$

Complexity for Solving Sinkhorn DRO

$$\min_{\theta,\lambda\geq 0}\left\{\lambda\bar{\rho}+\mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon\log\mathbb{E}_{z\sim\mathbf{N}(x,\epsilon\mathbf{I})}[e^{\ell(z;\theta)/(\lambda\epsilon)}]\right]\right\}$$

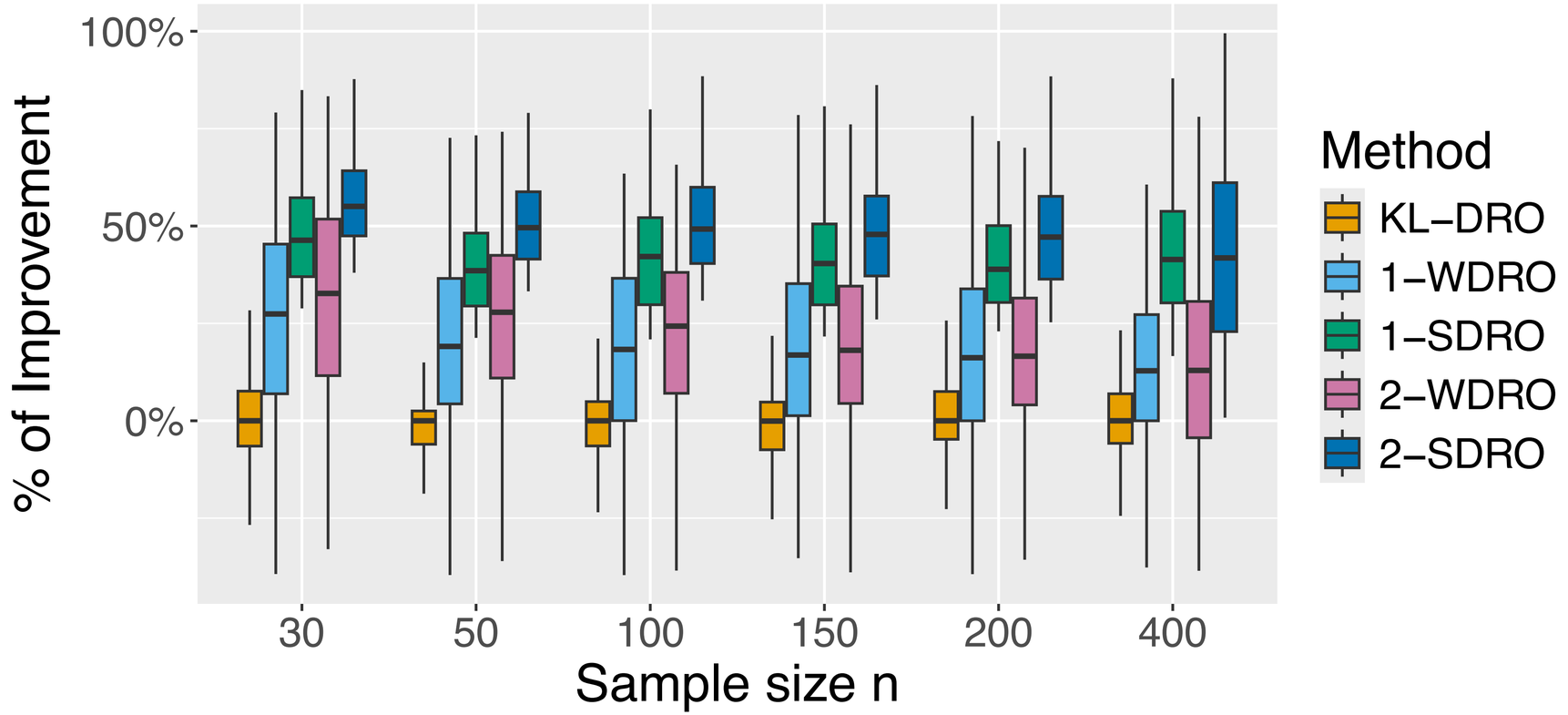


Random Sampling Estimator	
Convex	Nonconvex Smooth
$\tilde{O}(\delta^{-2})$	$\tilde{O}(\delta^{-4})$

4. Numerical Study and Discussion

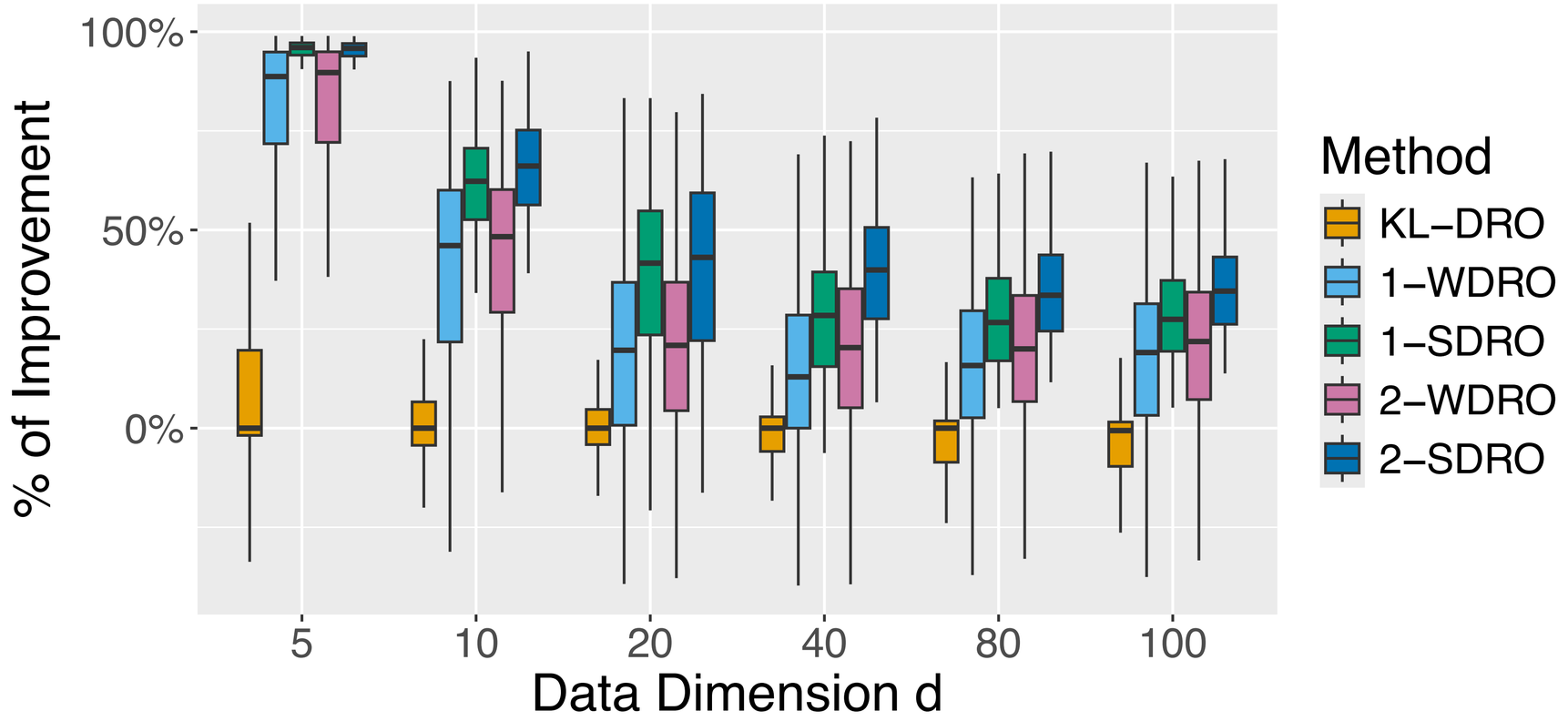
Numerical Results (Varying n)

$$\min_{x\in\mathbb{R}^d_+,\,\sum_i x_i=1} \mathbb{E}_{\mathbb{P}_{\text{True}}}[-\langle x,\xi\rangle] + \varrho\cdot \mathbb{P}_{\text{True}}\text{-CVaR}_\alpha(-\langle x,\xi\rangle)$$



Numerical Results (Varying d)

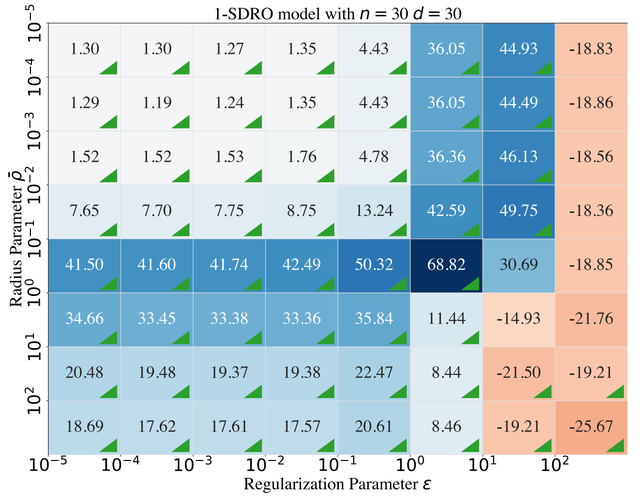
$$\min_{x\in\mathbb{R}^d_+,\,\sum_i x_i=1} \mathbb{E}_{\mathbb{P}_{\text{True}}}[-\langle x,\xi\rangle] + \varrho\cdot \mathbb{P}_{\text{True}}\text{-CVaR}_\alpha(-\langle x,\xi\rangle)$$

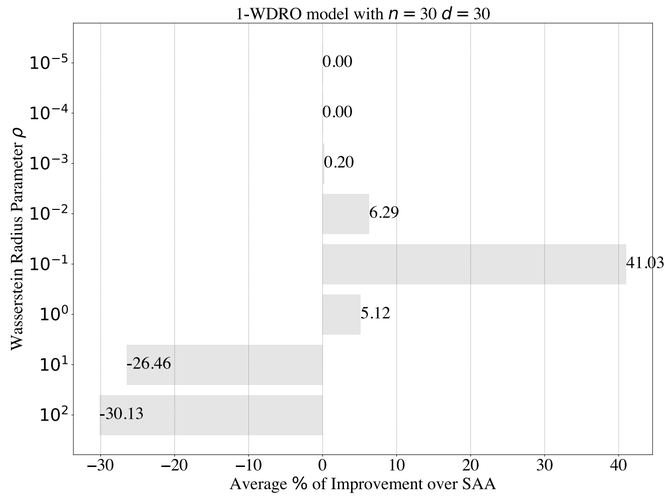


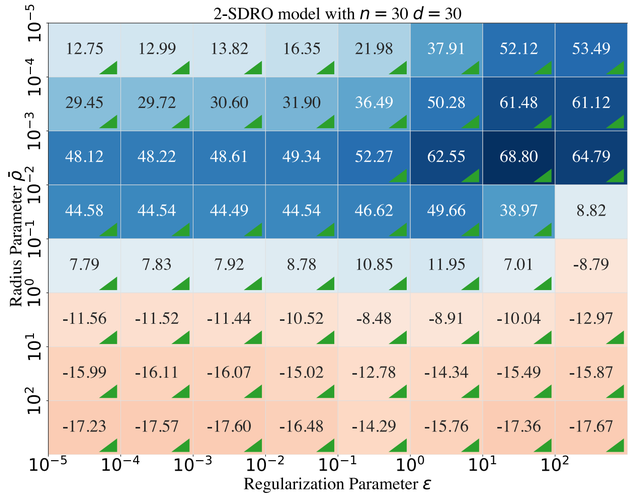
Numerical Results

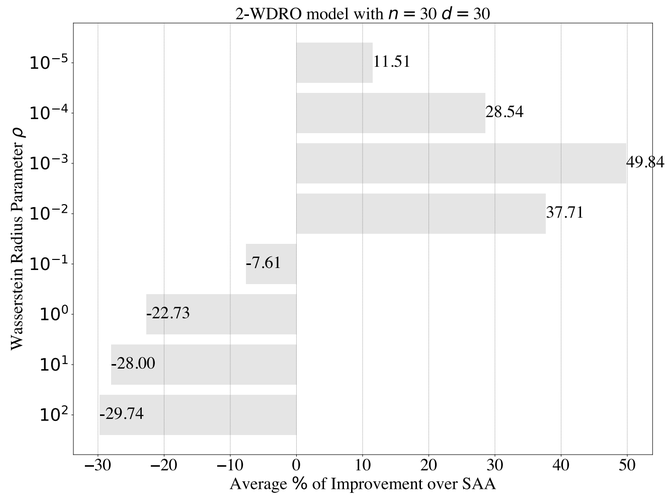
$$\min_{x\in\mathbb{R}^d_+,\,\sum_i x_i=1} \mathbb{E}_{\mathbb{P}_{\text{True}}}[-\langle x,\xi\rangle] + \varrho\cdot \mathbb{P}_{\text{True}}\text{-CVaR}_\alpha(-\langle x,\xi\rangle)$$

- **Ablation study** on robust level $\bar{\rho}$ and entropic regularization ϵ
- Entries at the same row use the same $\bar{\rho}$
- **Green** triangle indicates SDRO outperforms WDRO under the same $\bar{\rho}$





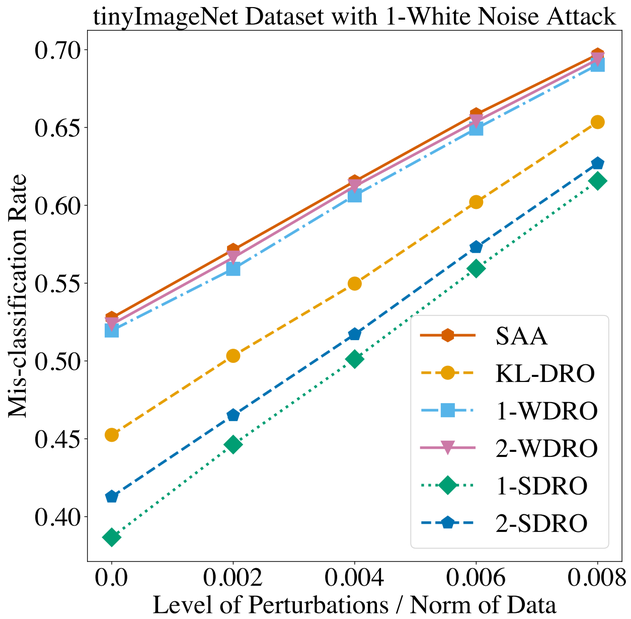


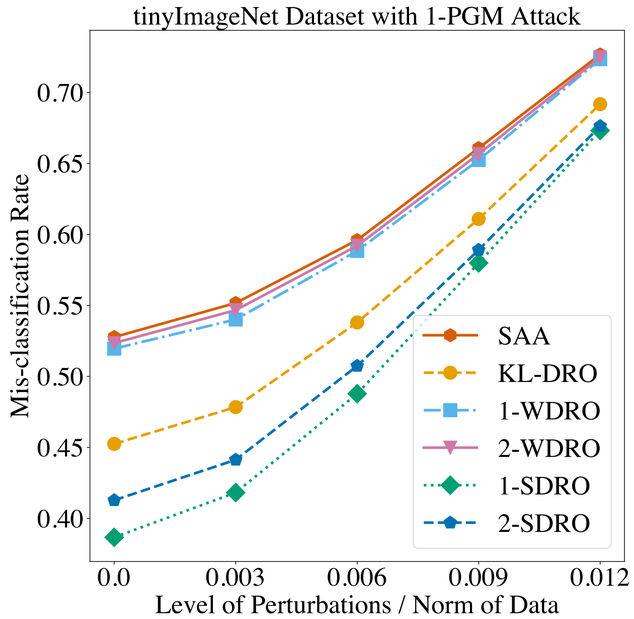


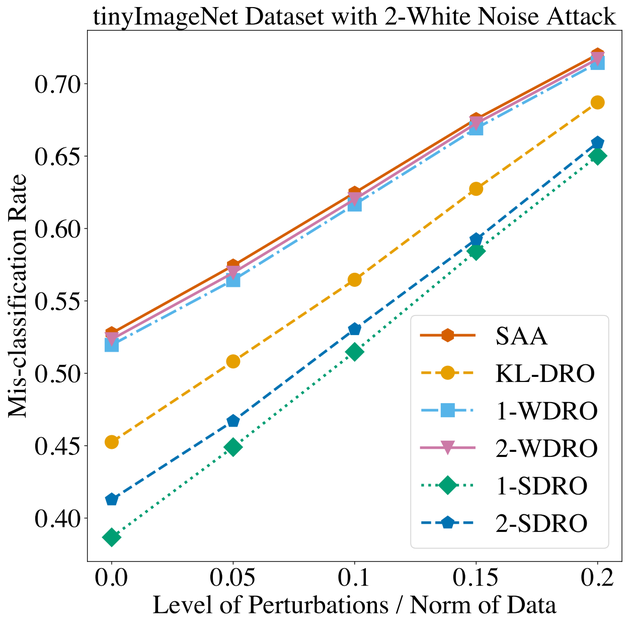
Numerical Results

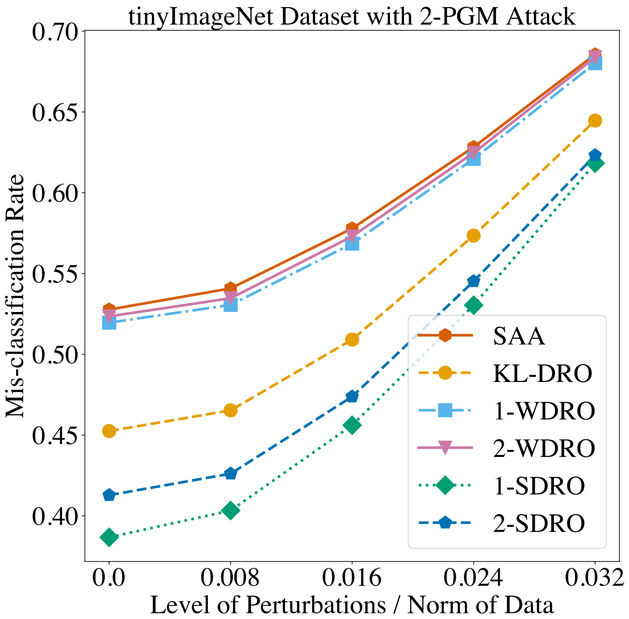
$$\min_{B\in\mathbb{R}^{d\times C}} \ \mathbb{E}_{(x,\mathbf{y})\sim\mathbb{P}_{\text{true}}}\left[h_B(x,\mathbf{y})\right], \quad h_B(x,\mathbf{y}) = -\mathbf{y}^\top B^\top x + \log\left(1^\top e^{B^\top x}\right).$$

- Error rate for tinyImageNet dataset (90000 training samples with dimension 512):





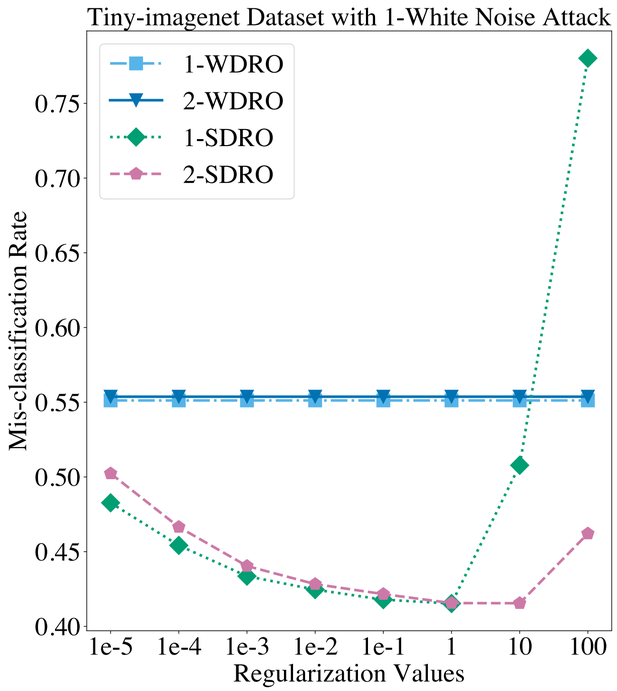


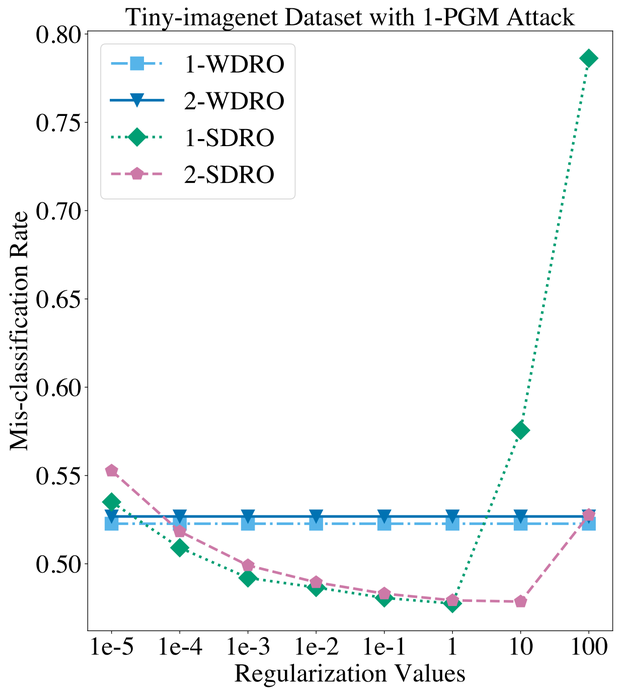


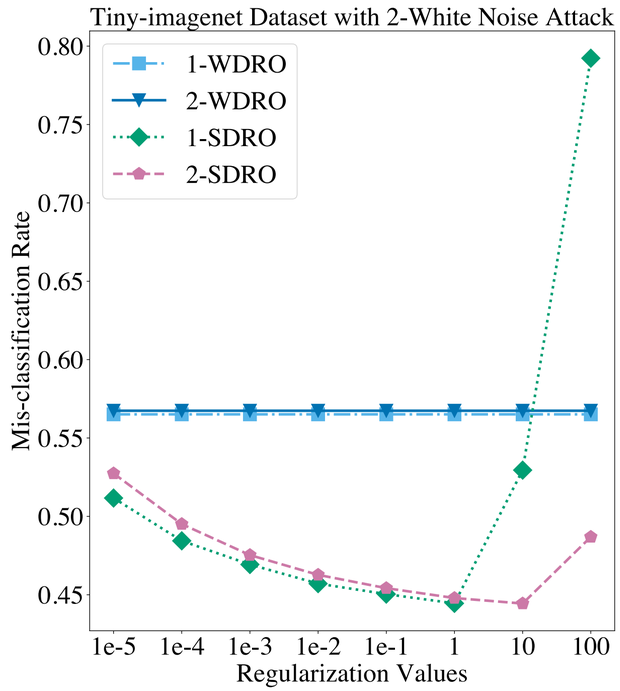
Numerical Results

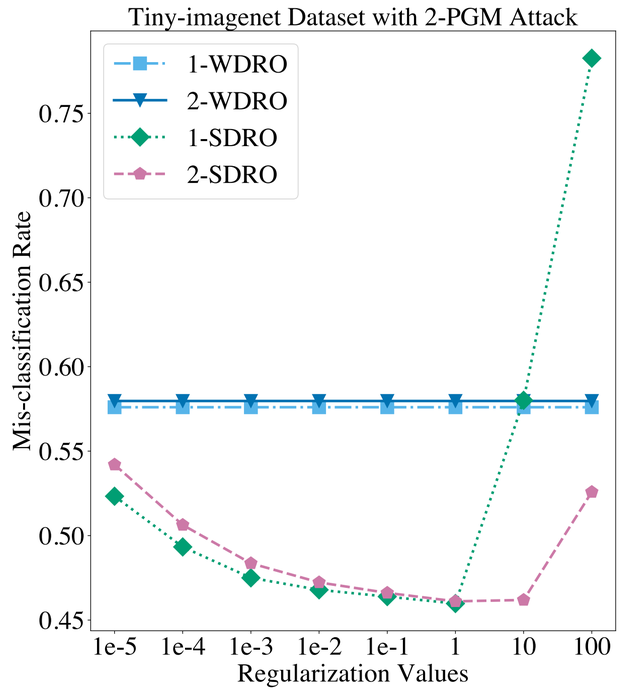
$$\min_{B\in\mathbb{R}^{d\times C}} \mathbb{E}_{(x,\mathbf{y})\sim\mathbb{P}_{\text{true}}}\left[h_B(x,\mathbf{y})\right], \quad h_B(x,\mathbf{y}) = -\mathbf{y}^\top B^\top x + \log\left(1^\top e^{B^\top x}\right).$$

- Ablation Study:

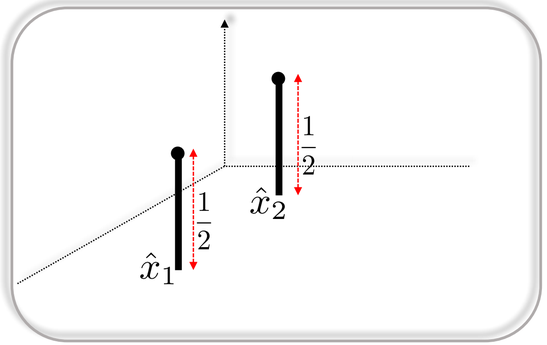


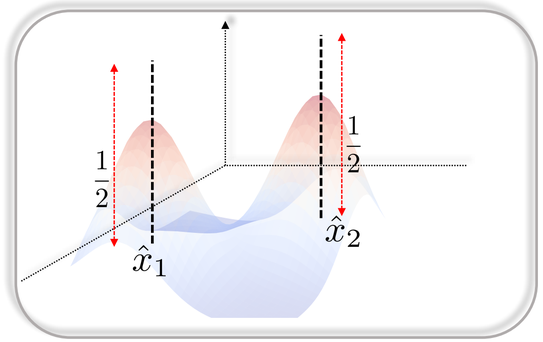






Conclusion





- Sinkhorn DRO is a variant of Wasserstein DRO:

  1. Absolutely continuous worst-case distribution thanks to entropic regularization;

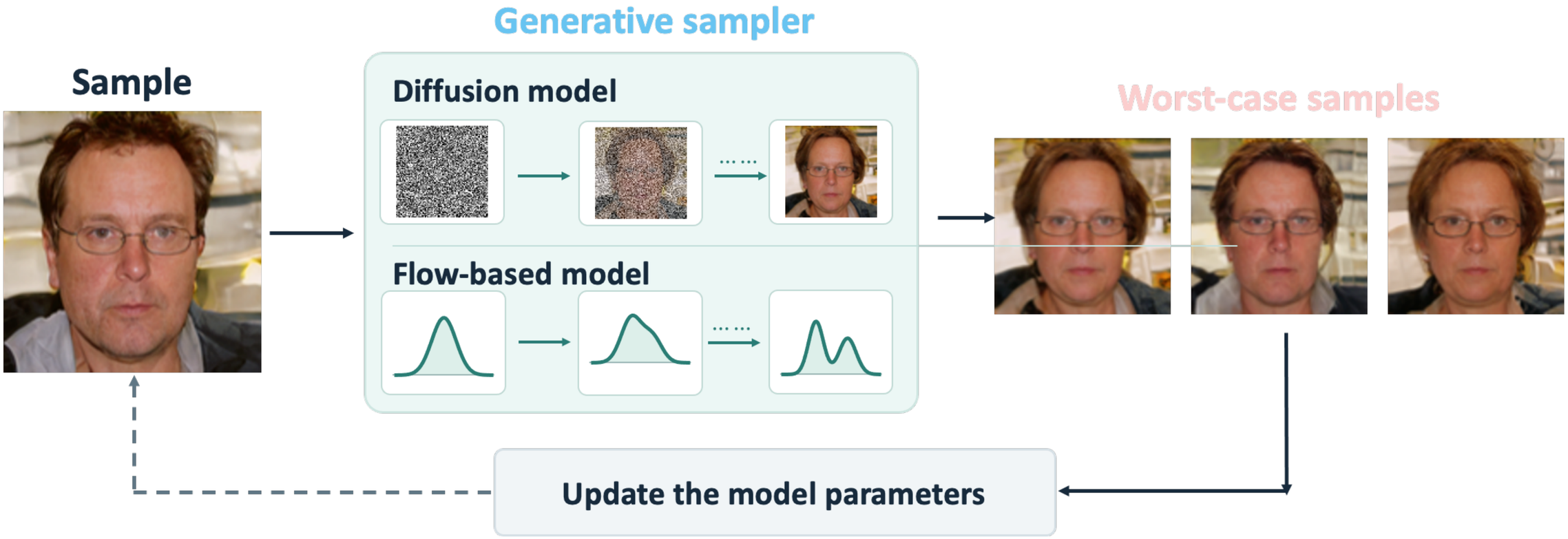
  2. Scalable computation by first-order method;

  3. Connection with optimal transport and KL-divergence.

	Random Sampling Estimator	
Loss $\ell(z, \cdot)$	Convex	Nonconvex Smooth
Complexity	$\tilde{O}(\delta^{-2})$	$\tilde{O}(\delta^{-4})$

Related Work

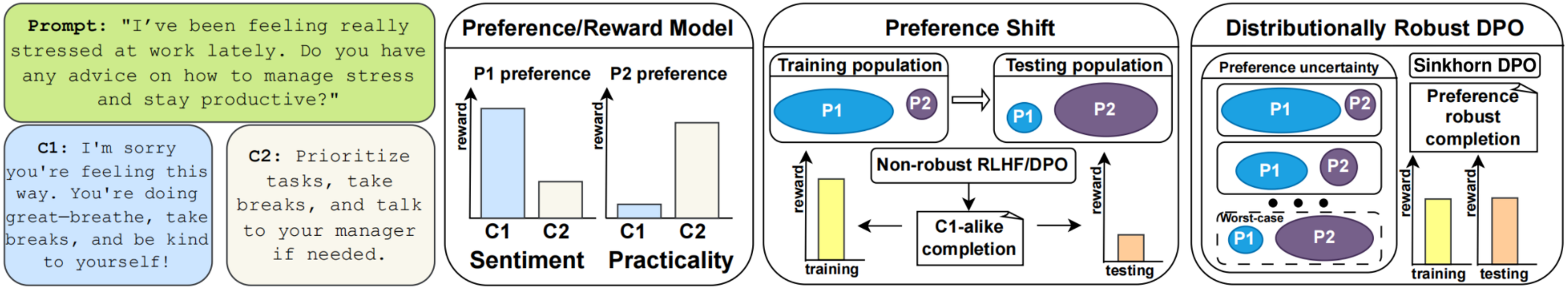
- **Generative Sampling for Sinkhorn DRO**



- Zusen Xu, and Jia-Jie Zhu. Gradient flow sampler-based distributionally robust optimization, ICML 2026
- Jie Wang. Iterative sampling methods for Sinkhorn distributionally robust optimization. arXiv preprint 2025
- Fenglin Zhang, Teyan Liu, and Jie Wang. "Generative Neural Networks for Sinkhorn Distributionally Robust Hypothesis Testing." arXiv preprint 2026

Related Work

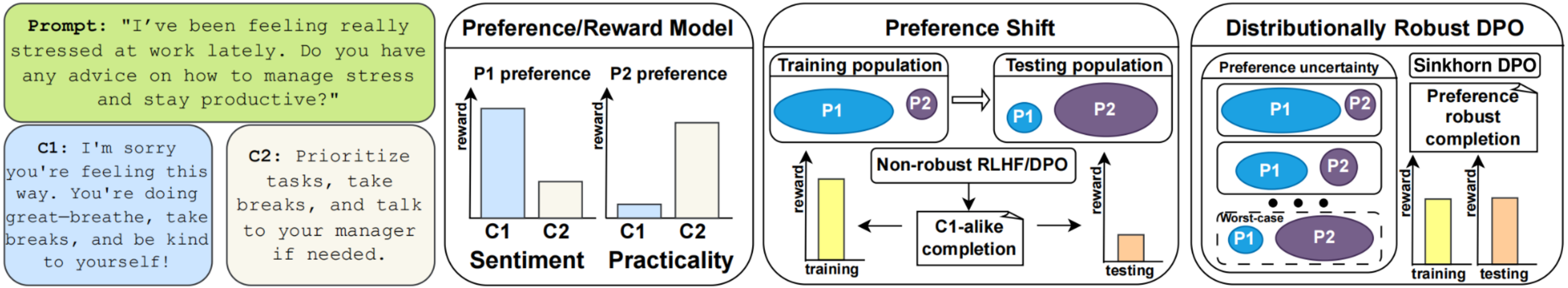
- **Generative Sampling for Sinkhorn DRO**
- **Sinkhorn DRO for broader applications (AI, energy, control, ...)**



- ICLR 2027 Anomoous Submission, Sinkhorn-DPO: Distributionally Robust Preference Alignment via Entropic Optimal Transport
- Sun, Haixiang, and Andrew L. Liu. Beyond Empirical Support: Structured Outlier Generation via Sinkhorn Optimal Transport. NeurIPS 2026
- Cescon, Riccardo, Andrea Martin, and Giancarlo Ferrari-Trecate. Sinkhorn Ambiguity Sets for Distributionally Robust Control: Convexity, Weak Compactness, and Tractability. arXiv preprint 2026

Related Work

- **Generative Sampling for Sinkhorn DRO**
- **Sinkhorn DRO for broader applications (AI, energy, control, ...)**



- **Future direction: explore the interaction of Sinkhorn DRO and AI**

- ICLR 2027 Anomoous Submission, Sinkhorn-DPO: Distributionally Robust Preference Alignment via Entropic Optimal Transport
- Sun, Haixiang, and Andrew L. Liu. Beyond Empirical Support: Structured Outlier Generation via Sinkhorn Optimal Transport. NeurIPS 2026
- Cescon, Riccardo, Andrea Martin, and Giancarlo Ferrari-Trecate. Sinkhorn Ambiguity Sets for Distributionally Robust Control: Convexity, Weak Compactness, and Tractability. arXiv preprint 2026

Related References

❋ **Wang J, Gao R, Xie Y (2026) Sinkhorn distributionally robust optimization. Operations Research, 2026, 74(3): 1581-1603.**

❋ Zusen Xu, and Jia-Jie Zhu. Gradient flow sampler-based distributionally robust optimization, ICML 2026

❋ Jie Wang. Iterative sampling methods for Sinkhorn distributionally robust optimization. arXiv preprint 2025

❋ Fenglin Zhang, Teyan Liu, and Jie Wang. "Generative Neural Networks for Sinkhorn Distributionally Robust Hypothesis Testing." arXiv preprint 2026

❋ ICLR 2027 Anomoous Submission, Sinkhorn-DPO: Distributionally Robust Preference Alignment via Entropic Optimal Transport

❋ Sun, Haixiang, and Andrew L. Liu. Beyond Empirical Support: Structured Outlier Generation via Sinkhorn Optimal Transport. NeurIPS 2026

❋ Cescon, Riccardo, Andrea Martin, and Giancarlo Ferrari-Trecate. Sinkhorn Ambiguity Sets for Distributionally Robust Control: Convexity, Weak Compactness, and Tractability. arXiv preprint 2026