Sinkhorn Distributionally Robust Optimization

Jie Wang

Georgia Institute of Technology

Analytics for X 2024 Conference

Collaborators

Rui Gao

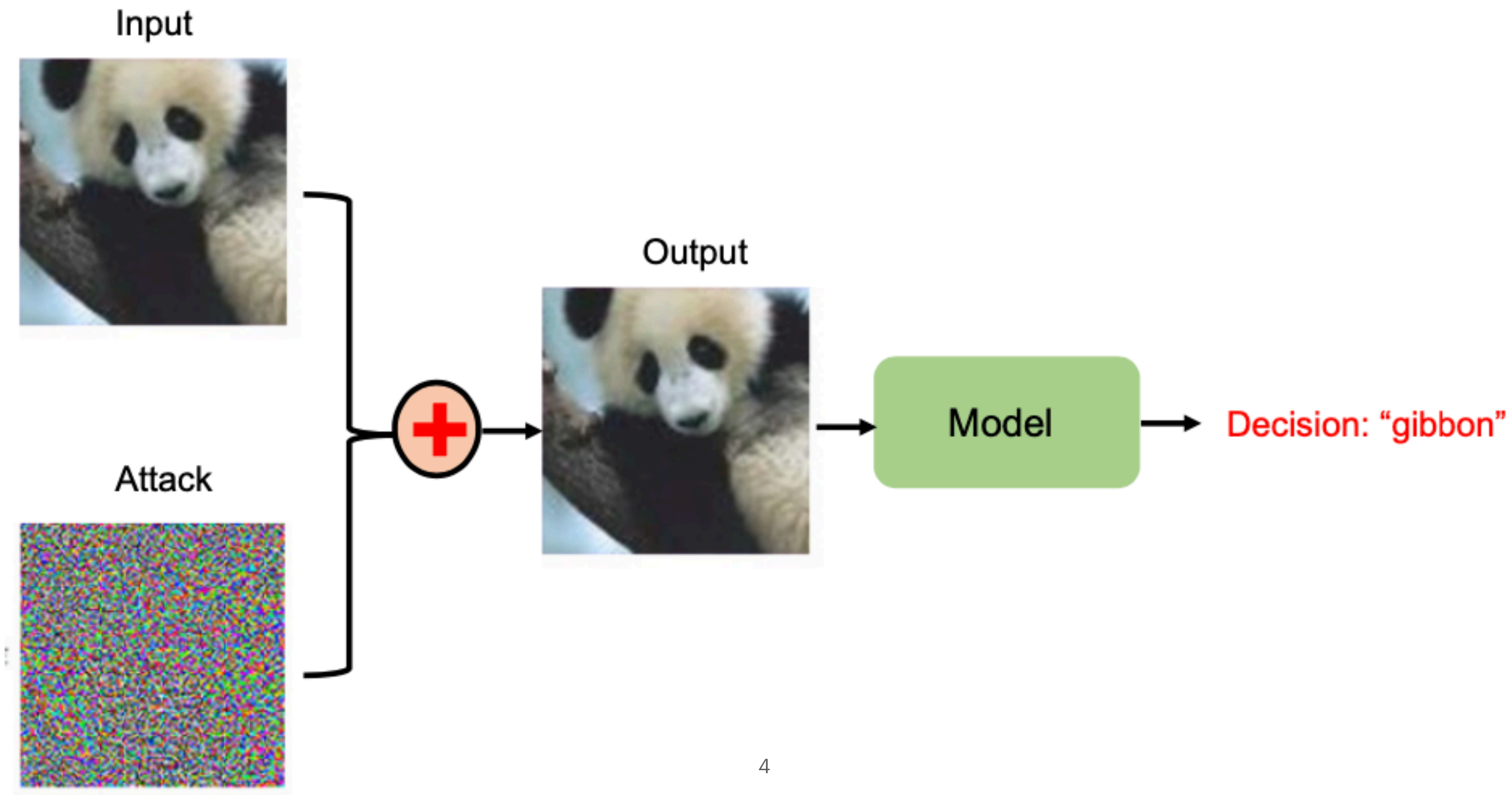
The University of Texas at Austin

Yao Xie

Georgia Institute of Technology

1. Introduction

Machine Learning [Goodfellow et al. 2015]



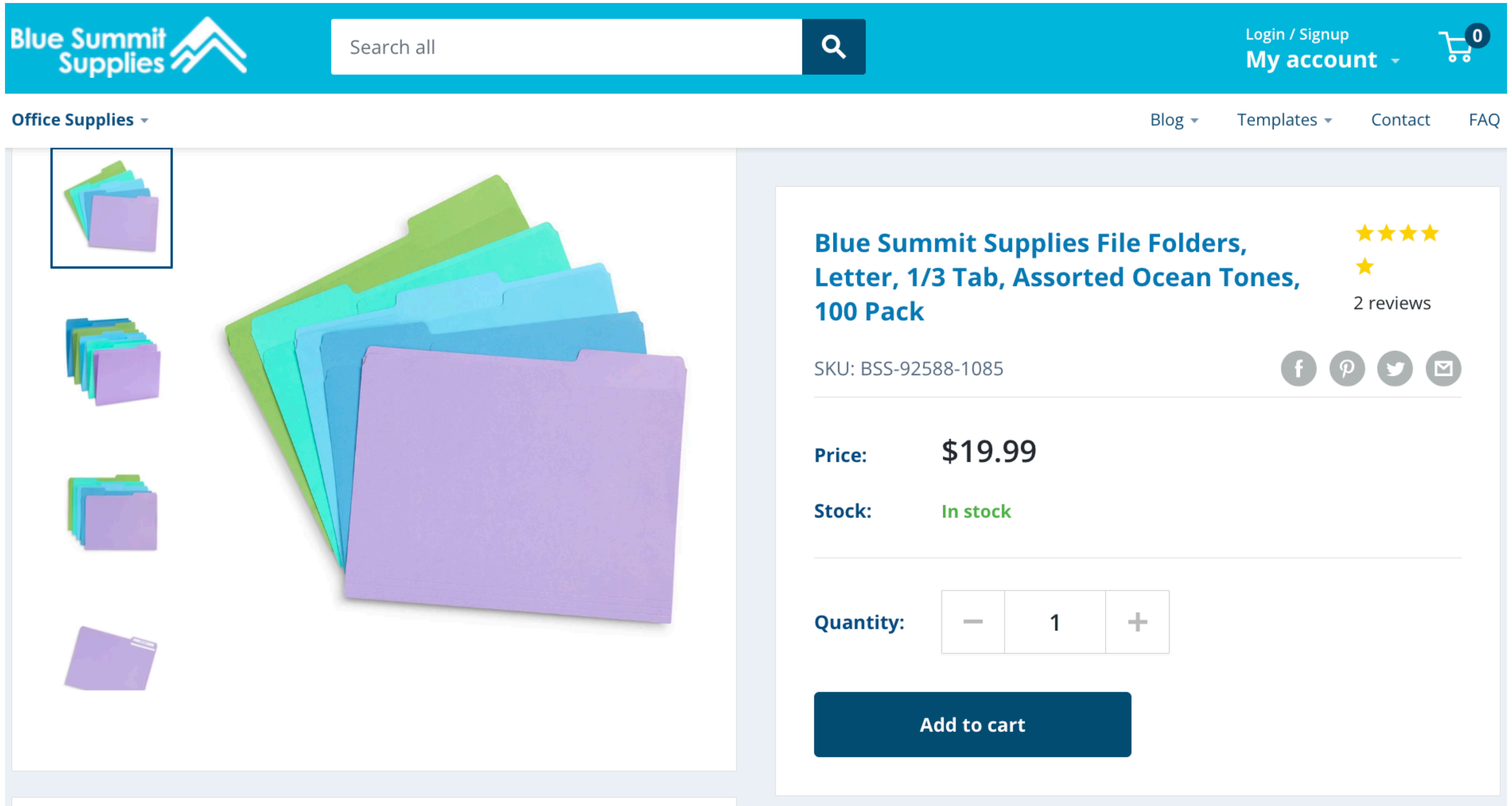
Machine Learning [Beery et al. ECCV2018]

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Strategic Pricing in eCommerce

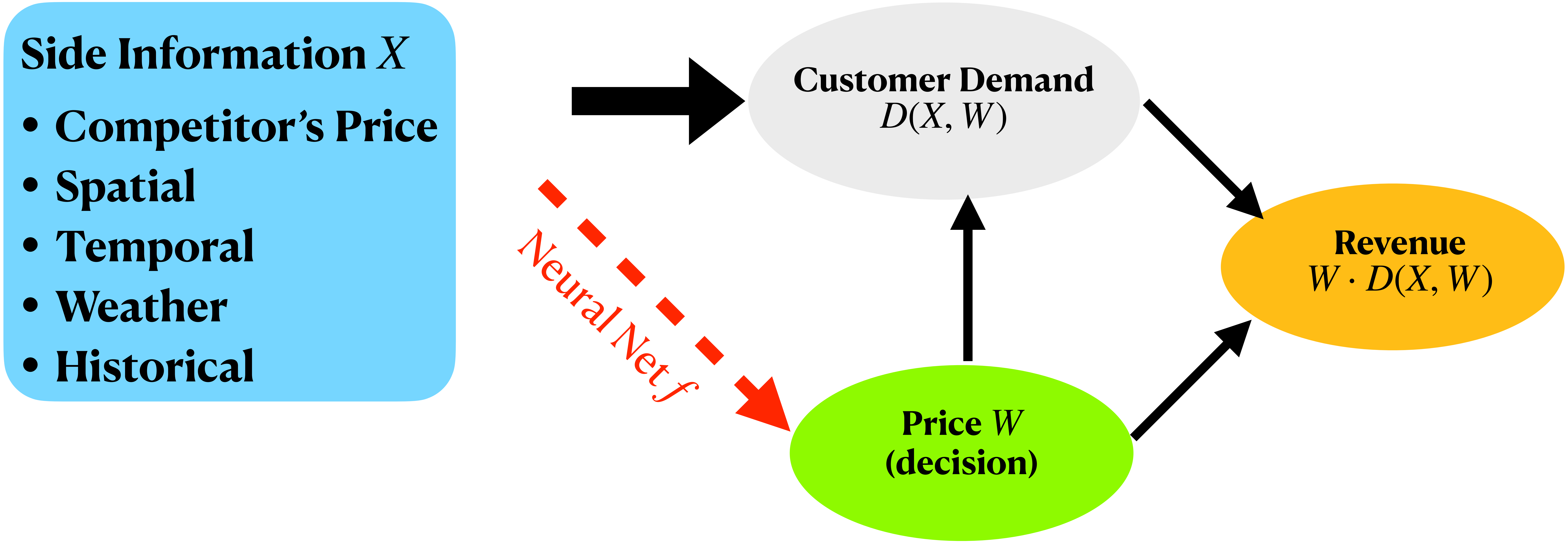


- Historical data over 2.5 years
- Pricing for real-world eCommerce market (Blue Summit Supplies)
- Online, data-driven decision making
- Criteria: Live Testing Performance

INFORMS (2023) INFORMS 2023 BSS Data Challenge Competition, https://sites.google.com/view/dmdaworkshop2023/data-challenge

Wang J (2023) Reliable Offline Pricing in eCommerce Decision-Making: A Distributionally Robust Viewpoint, Finalist of Competition

Strategic Pricing in eCommerce



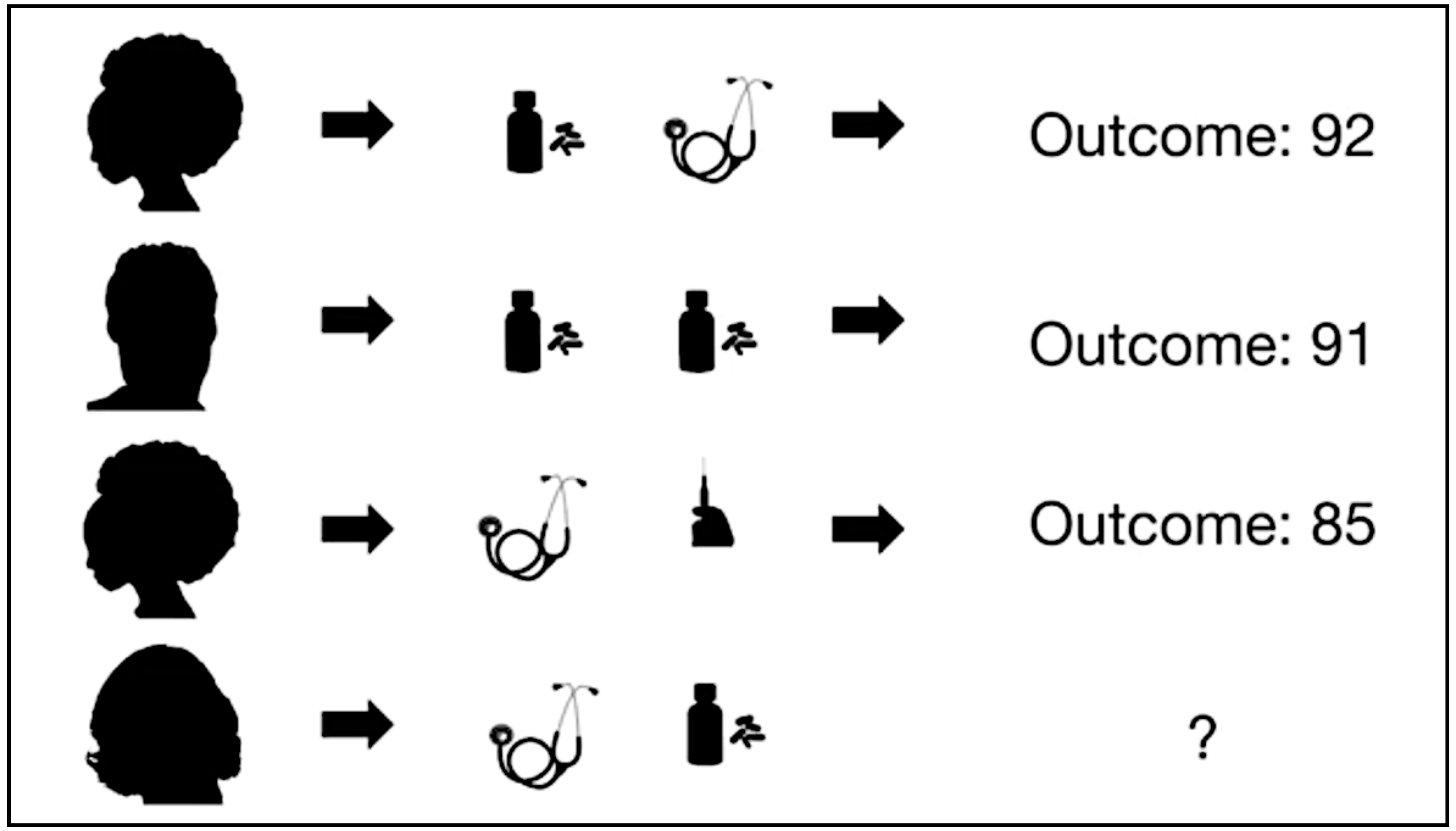
Cons: Distribution Shift on (side information, customer demand, price)!

INFORMS (2023) INFORMS 2023 BSS Data Challenge Competition, https://sites.google.com/view/dmdaworkshop2023/data-challenge

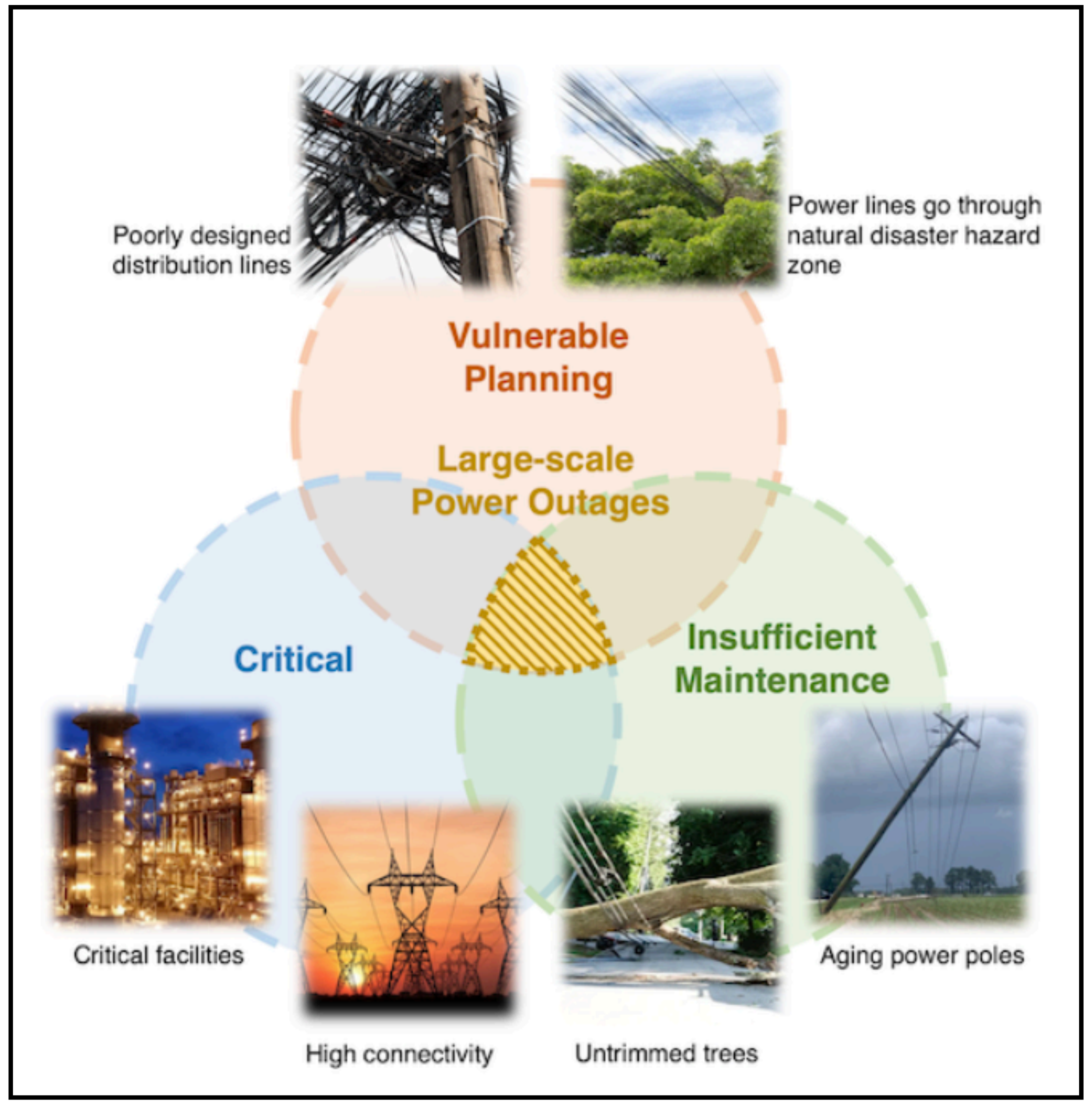
Wang J (2023) Reliable Offline Pricing in eCommerce Decision-Making: A Distributionally Robust Viewpoint, Finalist of Competition

Off-Policy Evaluation

Self-Driving



Healthcare



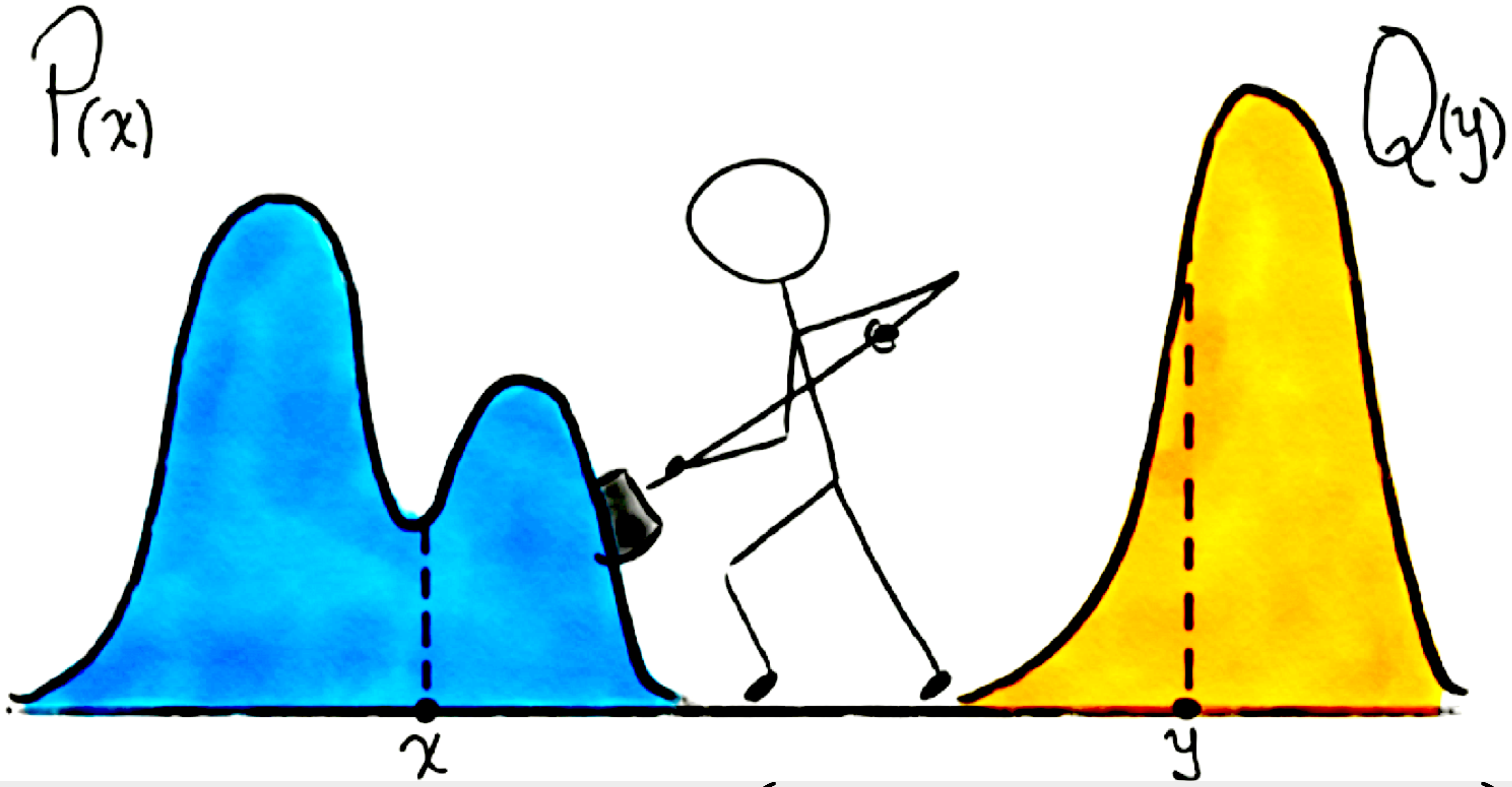
Power System: Resiliency

Worst-case scenarios for “System Stress-Test”

Wang J, Gao R, Zha H. Reliable off-policy evaluation for reinforcement learning. Operations Research, 2024, 72(2): 699-716

Wasserstein Distributionally Robust Optimization

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



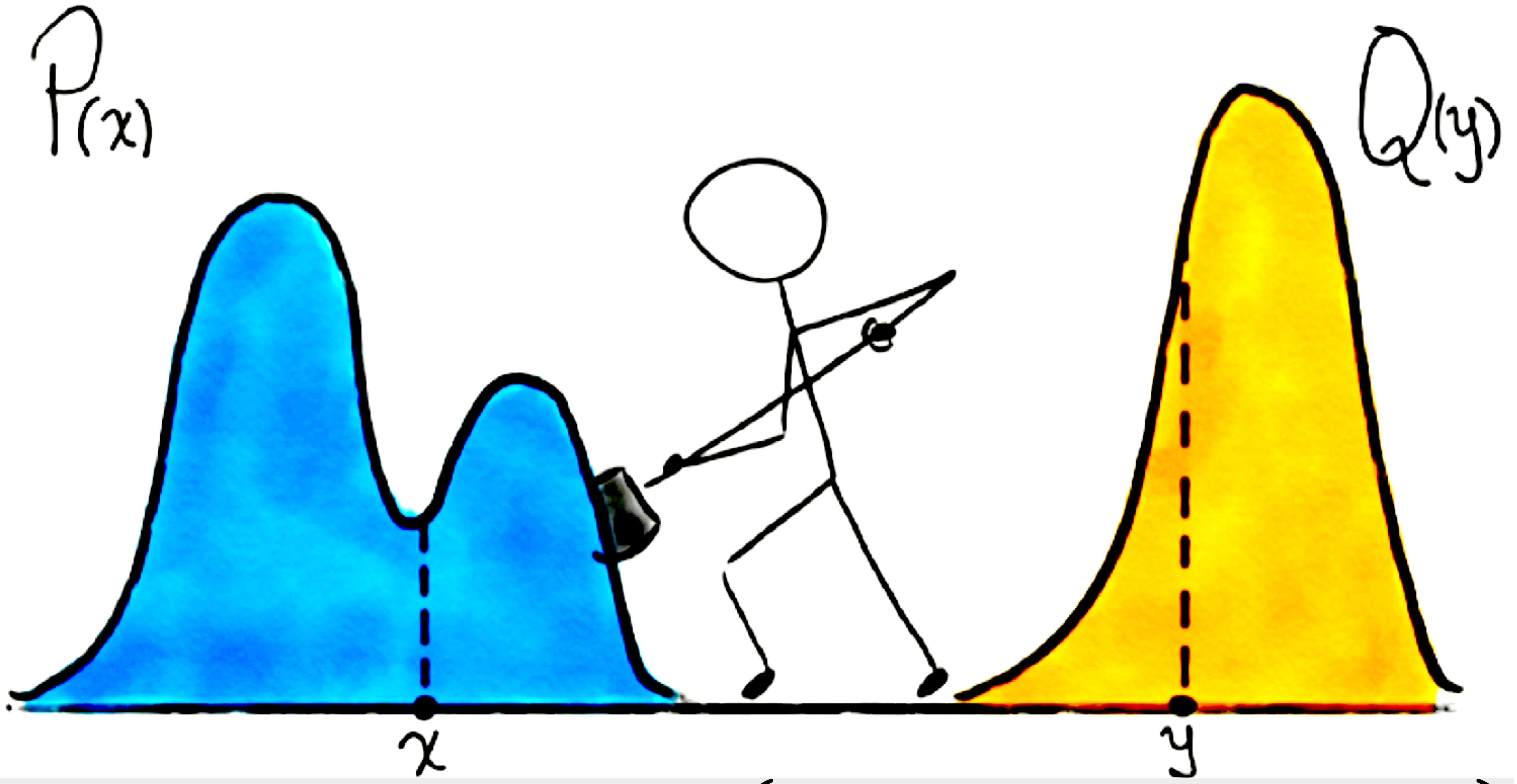
$$\mathcal{W}(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{ \mathbb{E}_{(x,y)\sim\gamma} \left[\|x-y\|^2 \right] \right\}$$

Whole Probability Space

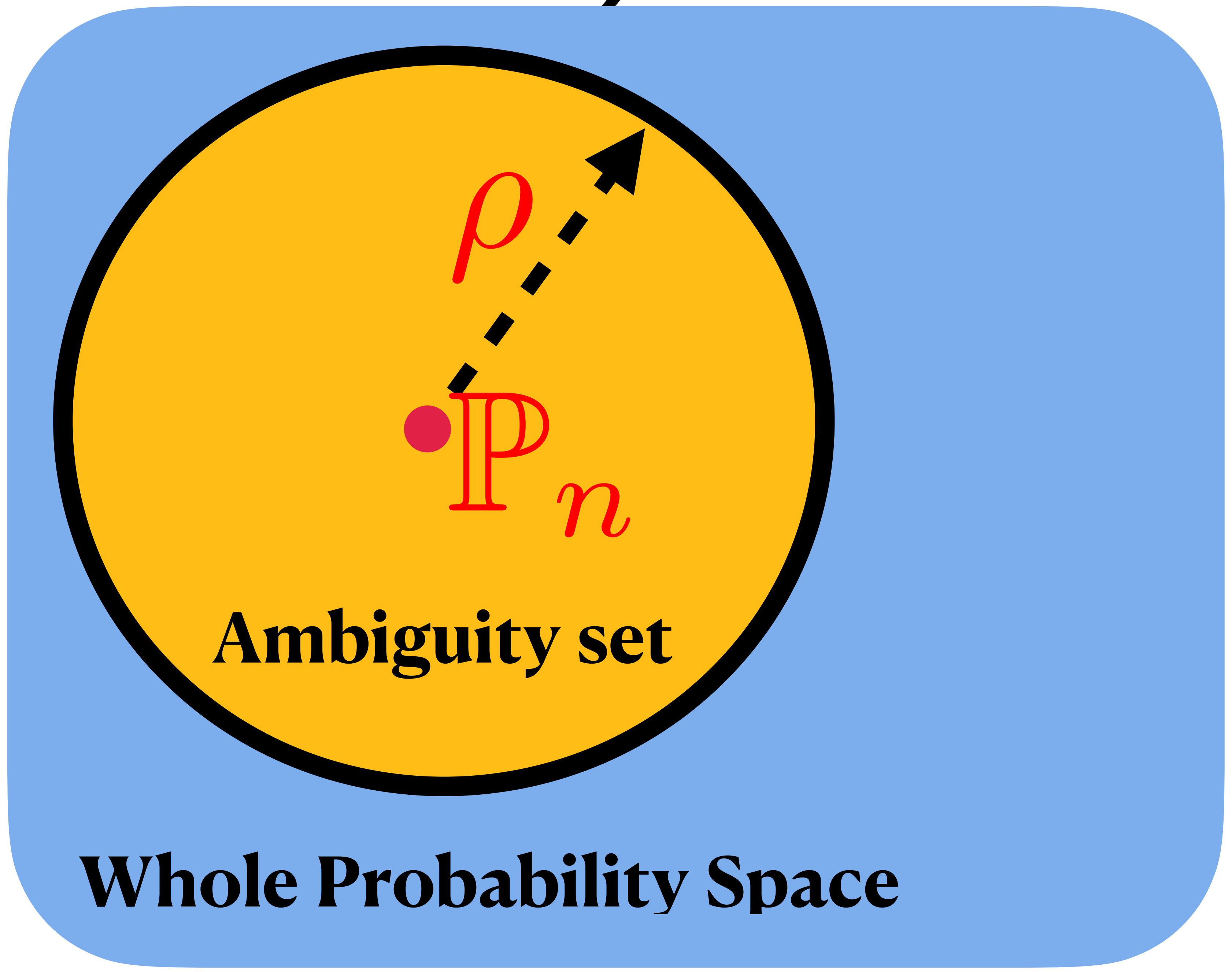
- data-driven, non-parametric, free of distributional assumptions

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

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

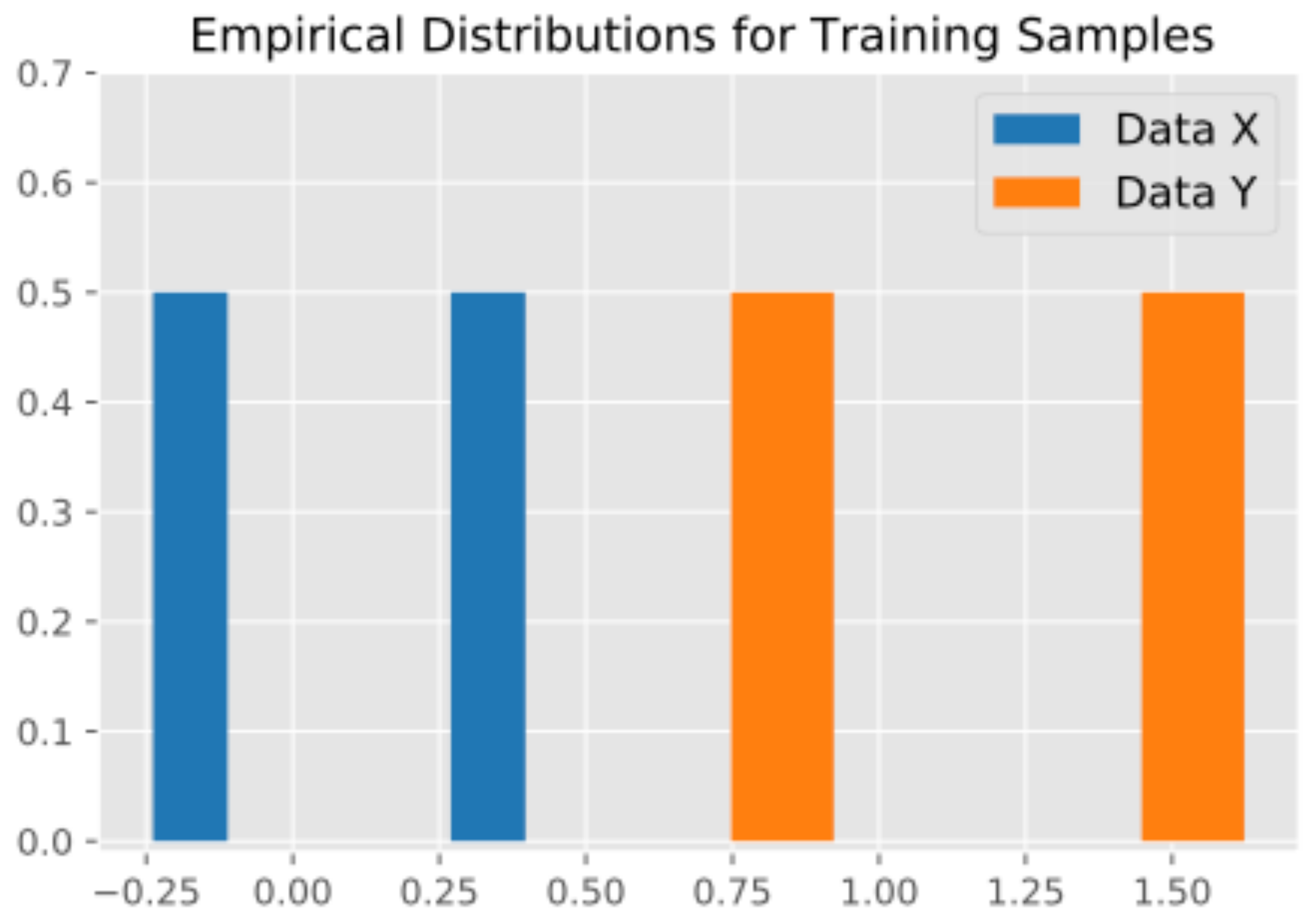
$$= \min_{\theta,\lambda\geq 0} \left\{ \lambda\rho + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\underbrace{\sup_{z} \left\{ \ell(z;\theta) - \lambda\|x-z\|^2 \right\}}_{\text{Moreau-Yoshida regularization}} \right] \right\}$$ **(Strong Dual Reformulation)**

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, ...]

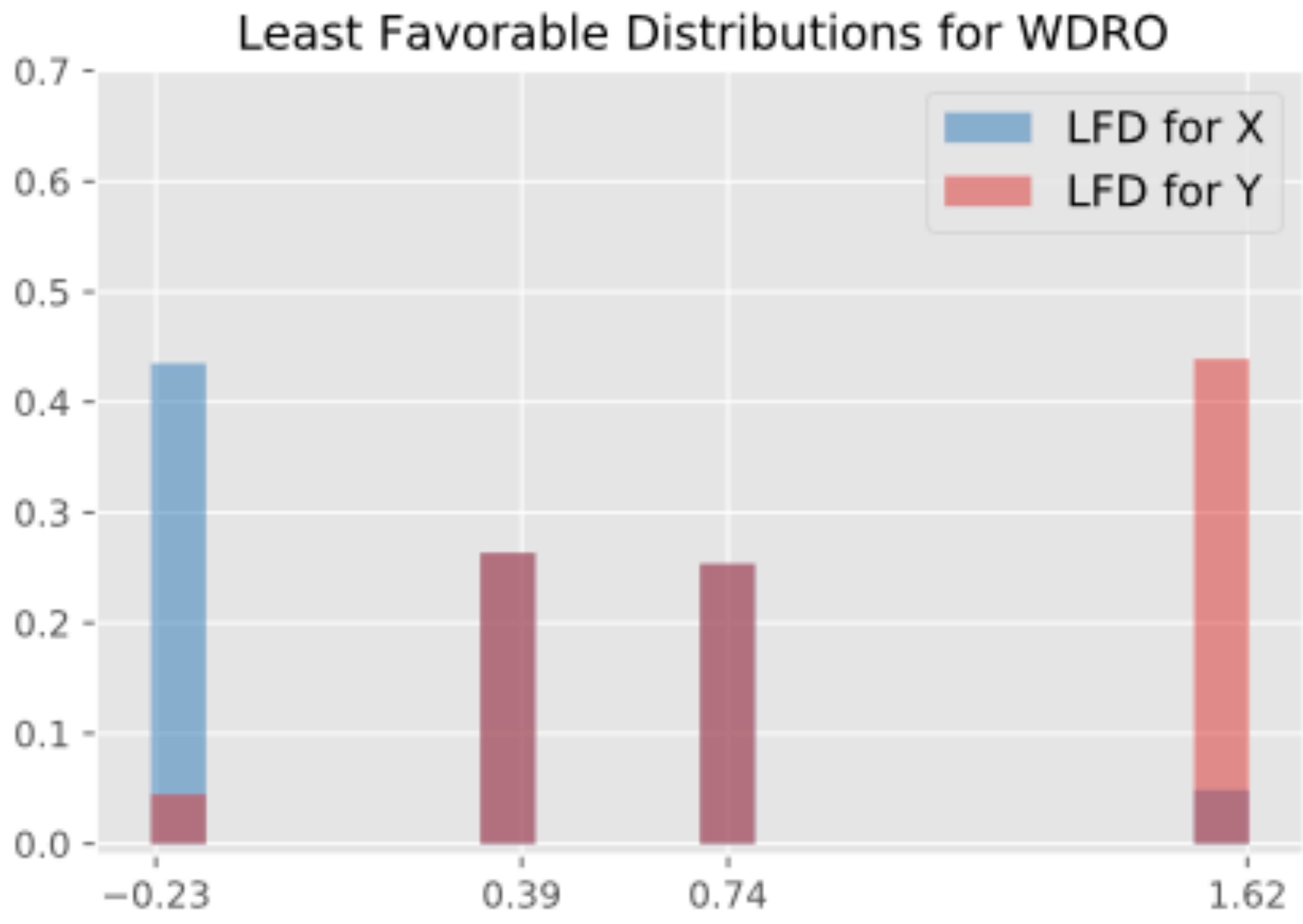
Cons: Wasserstein DRO is not necessarily tractable for general applications

Worst-case Distribution of Wasserstein DRO

- The worst-case distribution (LFD) $\mathbb{P}^*$ for WDRO is discrete
- In general, difficult to compute the LFD, and not directly generalizable beyond training samples
- Desired: **Continuous LFD, generalize to the unseen**



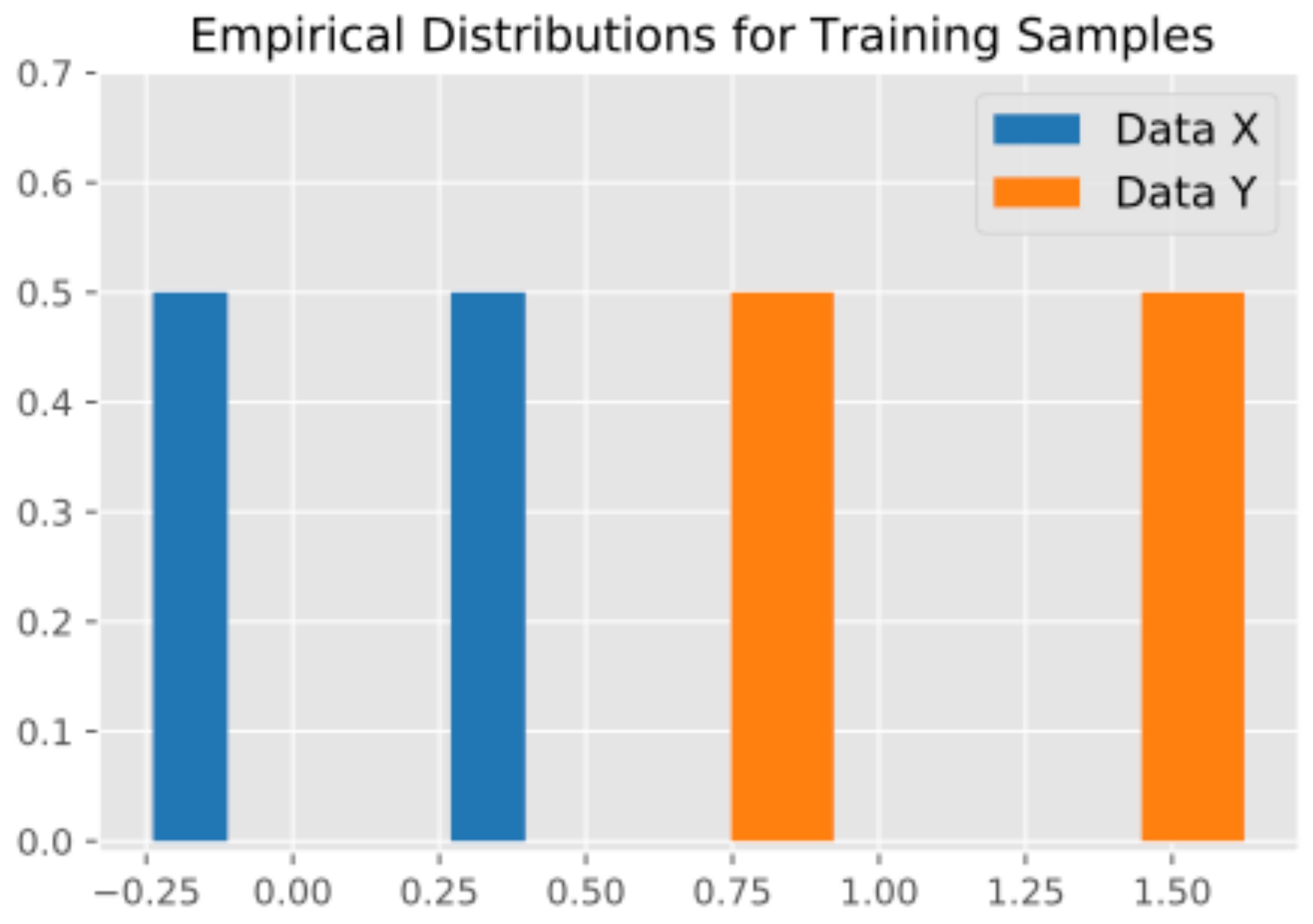
(a) Histogram of Training Samples



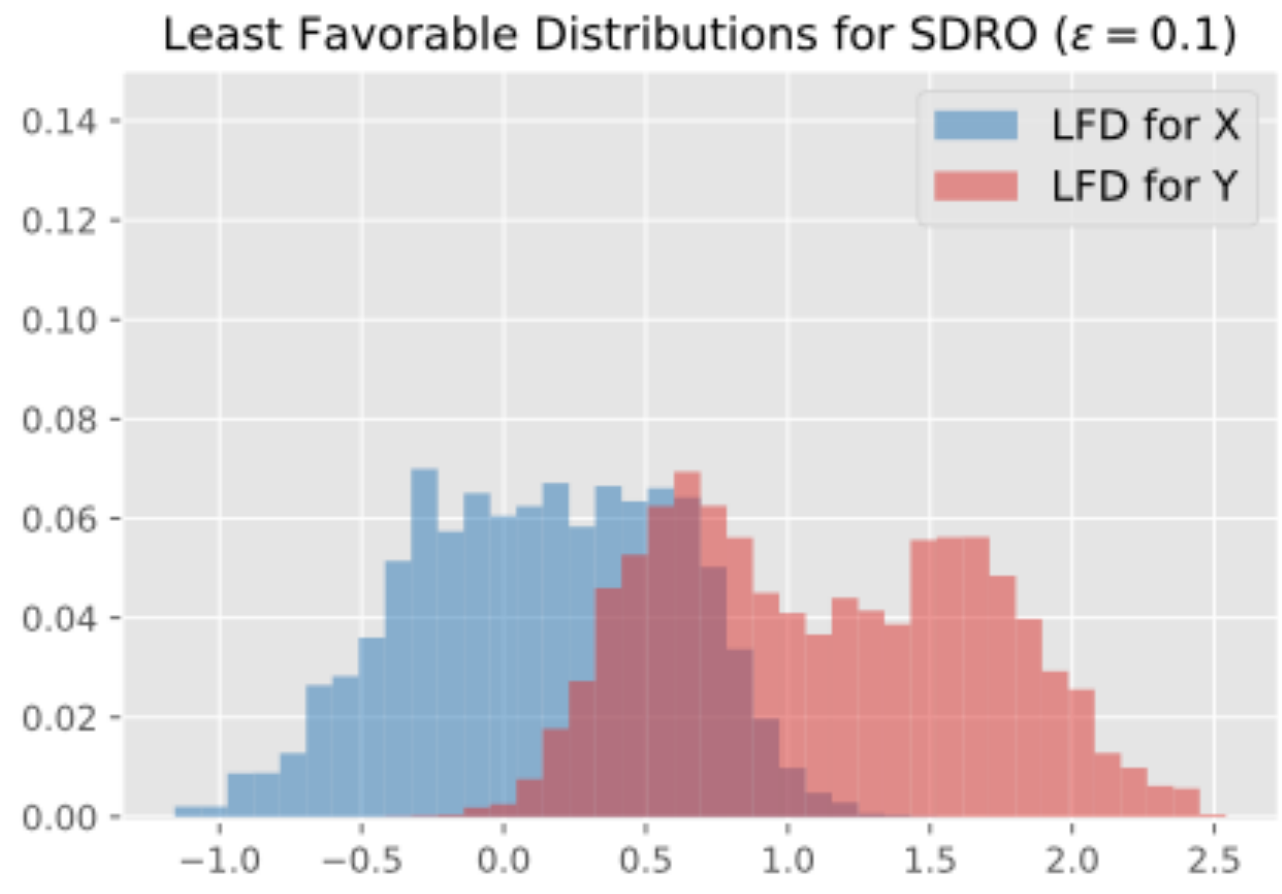
(b) LFD from WDRO

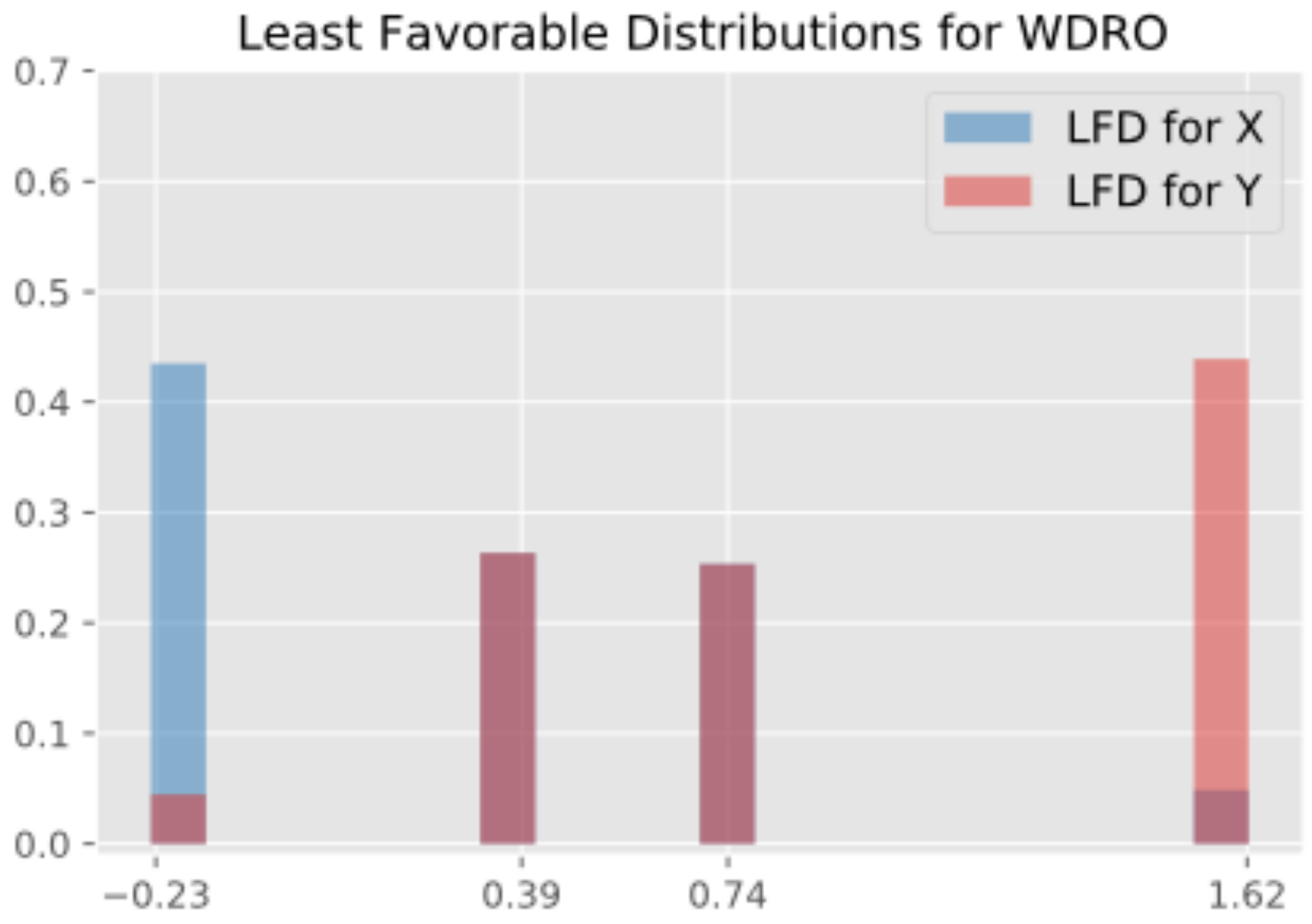
Worst-case Distribution of Wasserstein DRO

- The worst-case distribution (LFD) $\mathbb{P}^*$ for WDRO is discrete
- In general, difficult to compute the LFD, and not directly generalizable beyond training samples
- Desired: **Continuous LFD, generalize to the unseen**



(a) Histogram of Training Samples

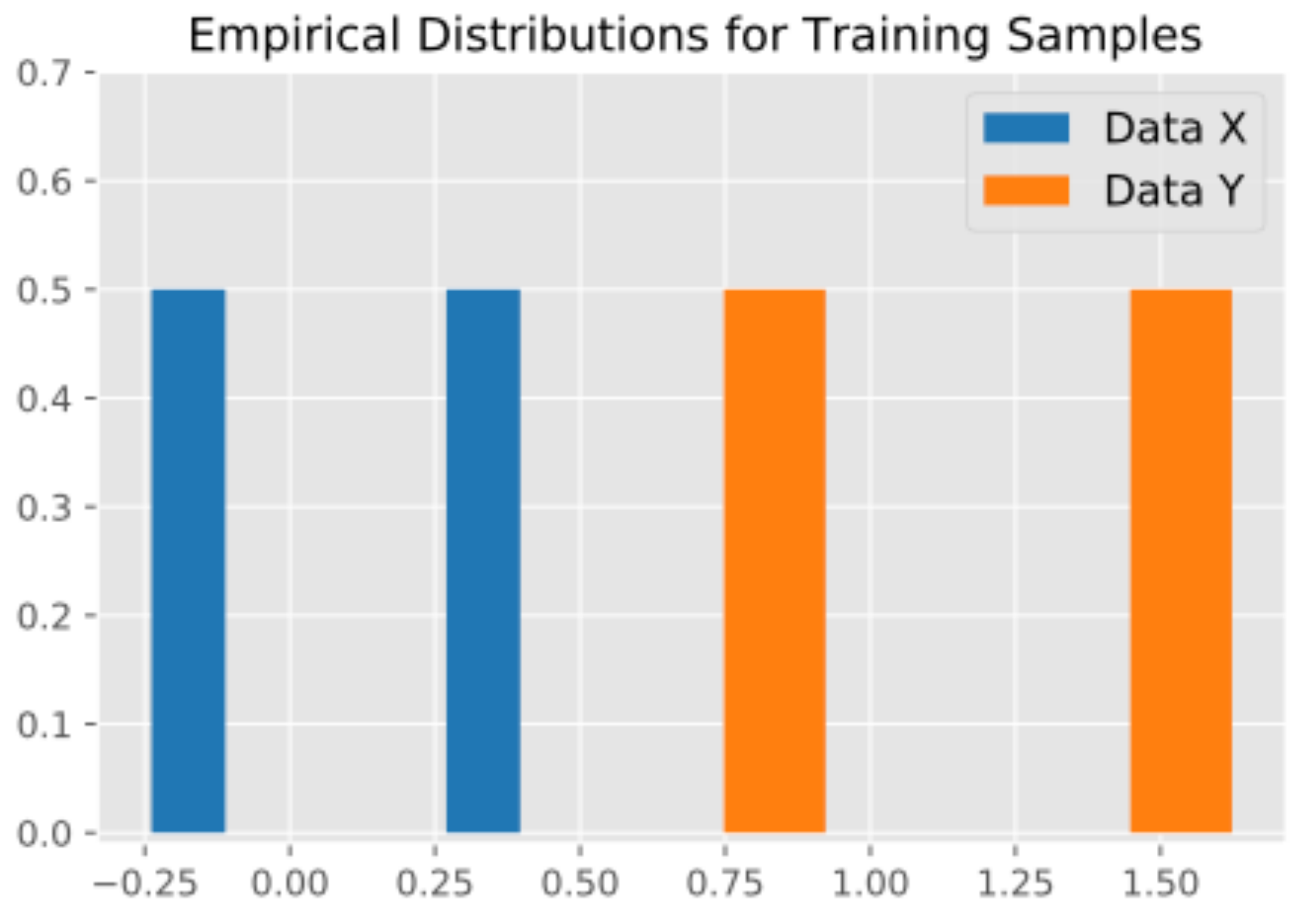




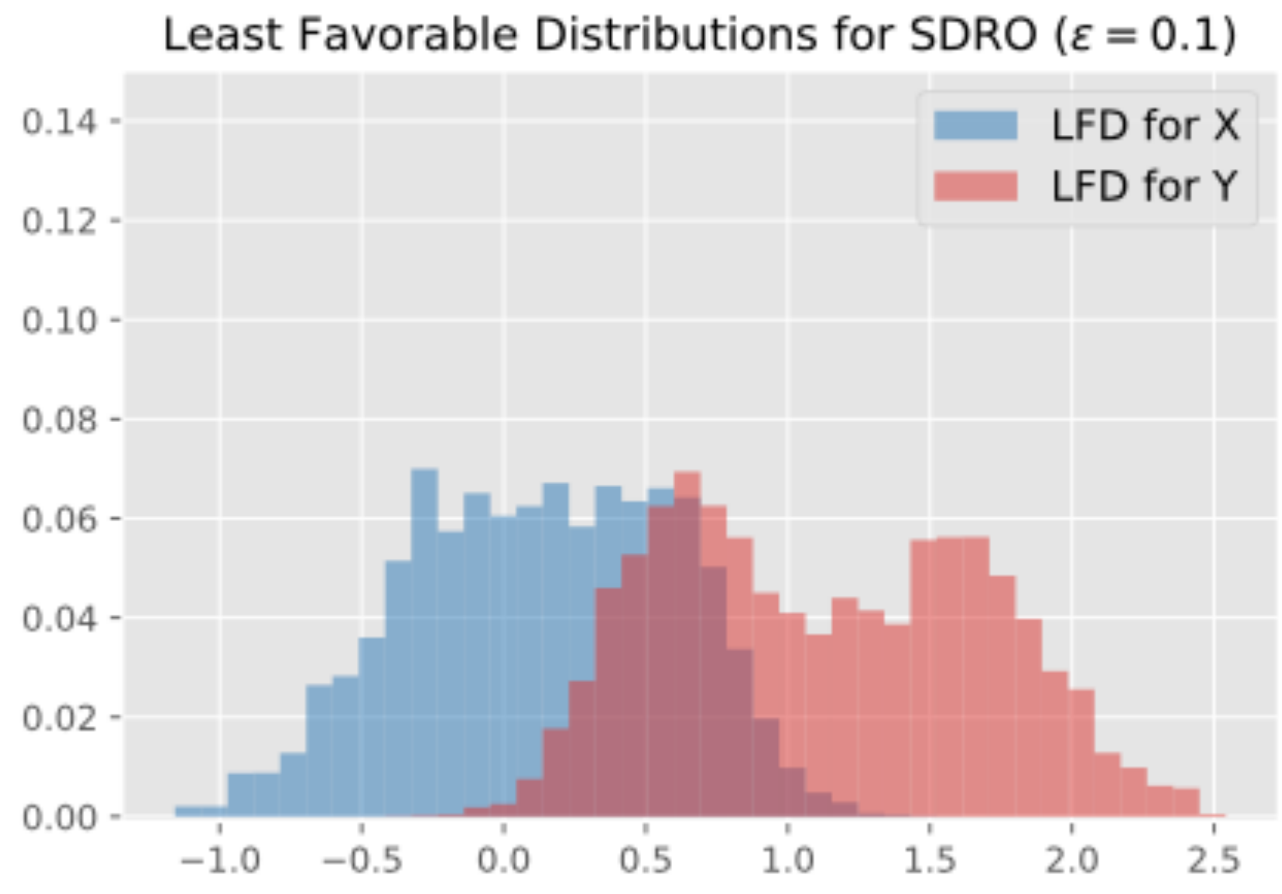
(b) LFD from WDRO

Worst-case Distribution of Wasserstein DRO

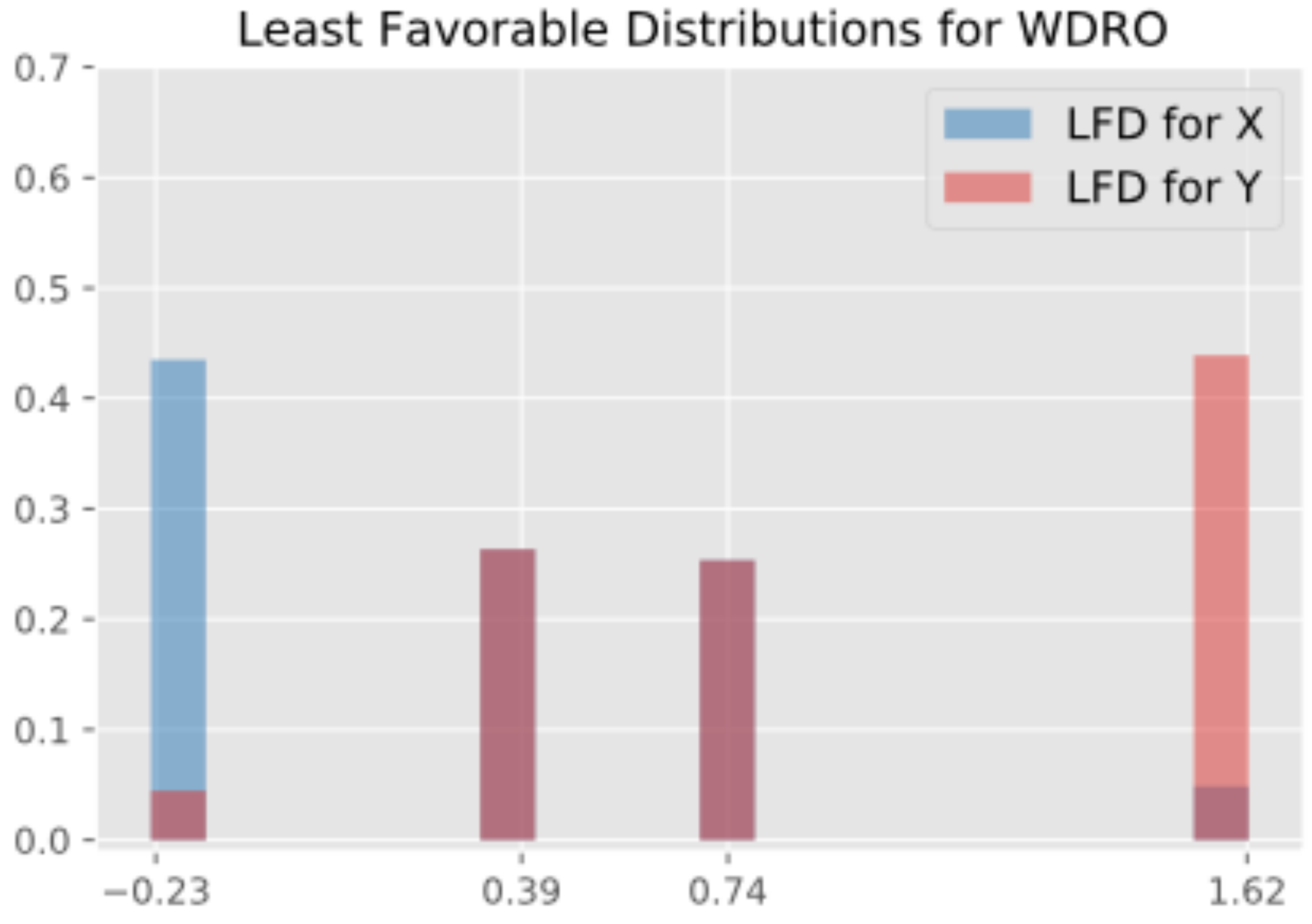
- The worst-case distribution (LFD) $\mathbb{P}^*$ for WDRO is discrete
- In general, difficult to compute the LFD, and not directly generalizable beyond training samples
- Desired: **Continuous LFD, generalize to the unseen**



(a) Histogram of Training Samples



LFD from Sinkhorn DRO (Proposed)

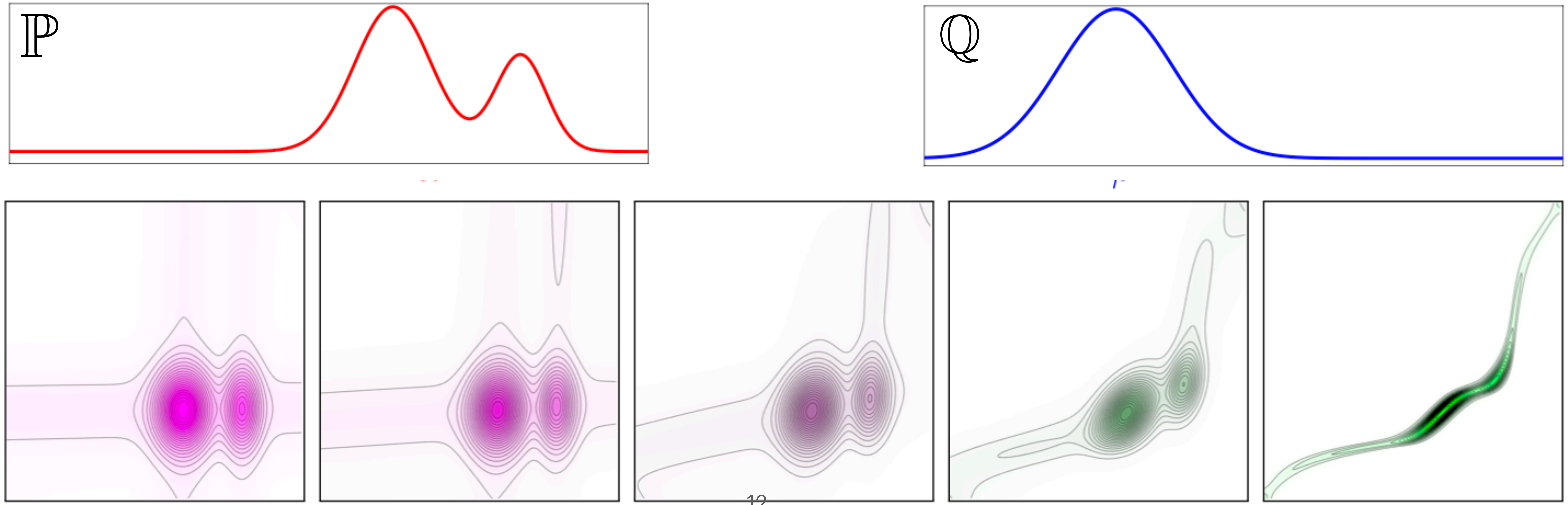


(b) LFD from WDRO

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



$\epsilon = 10$ $\epsilon = 1$ $\epsilon = 0.5$ $\epsilon = 0.1$ $\epsilon = 0.01$

Sinkhorn Discrepancy

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

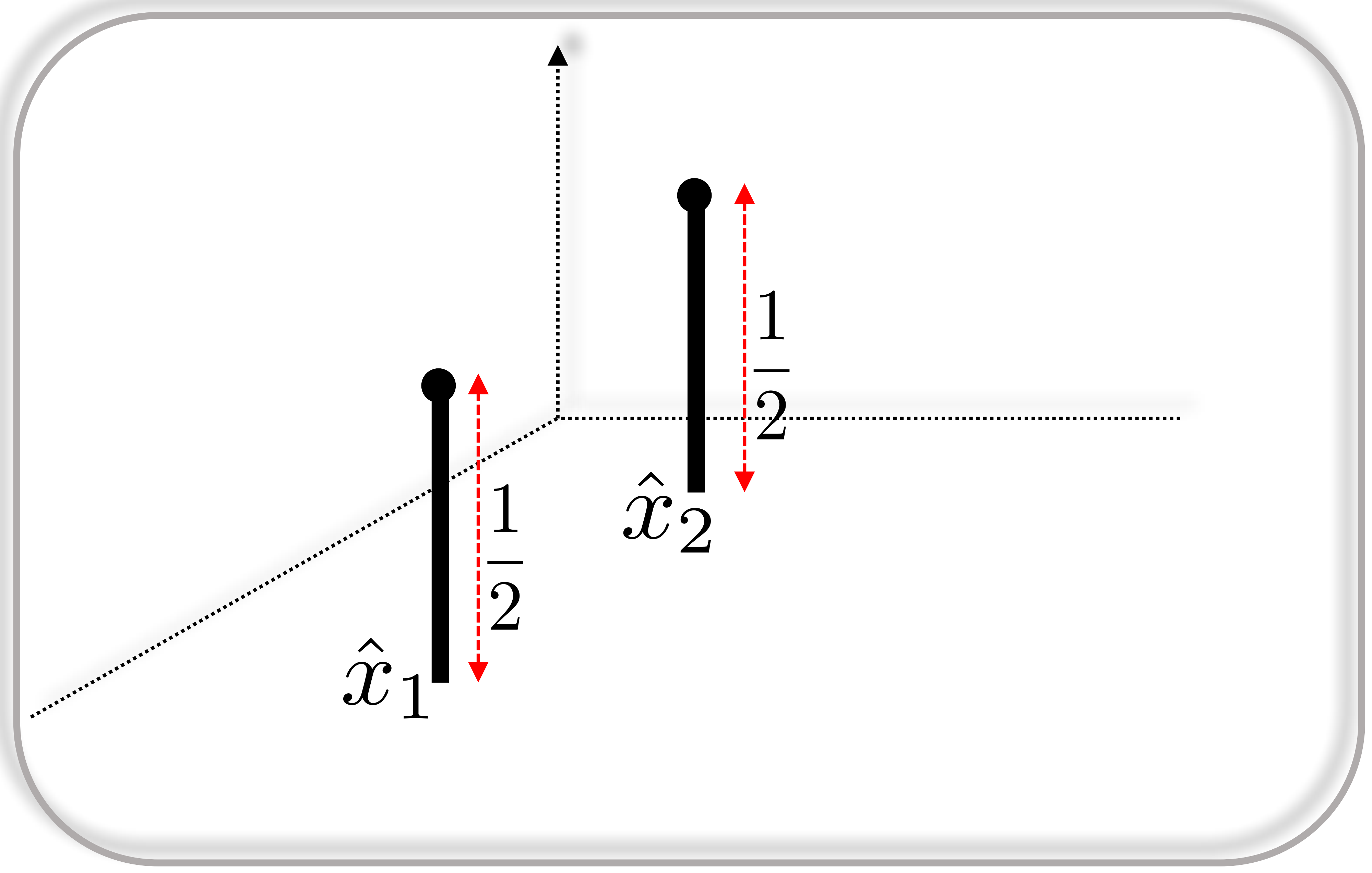
Historical Review:

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

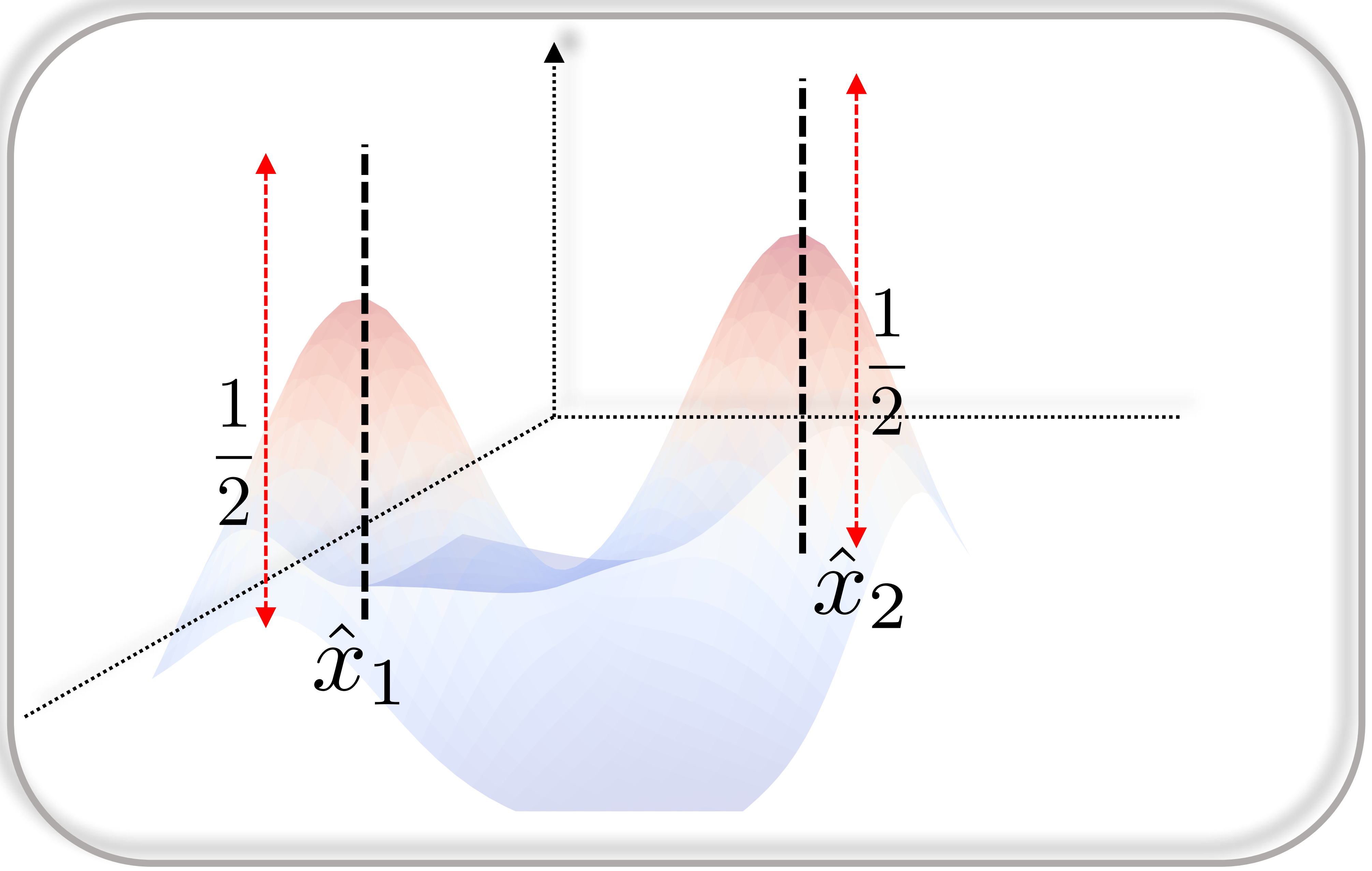
Main Framework

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

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

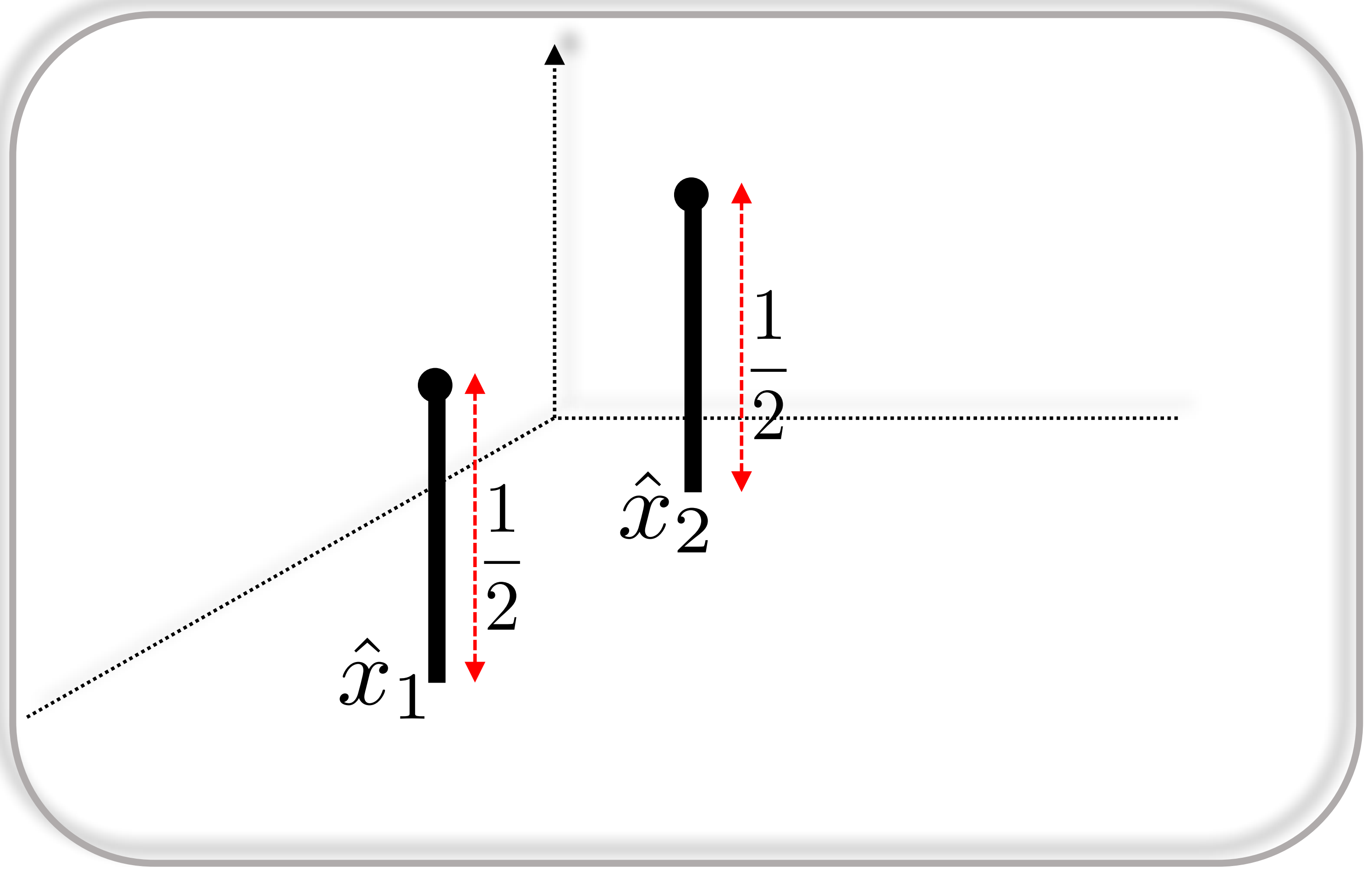


Empirical Distribution

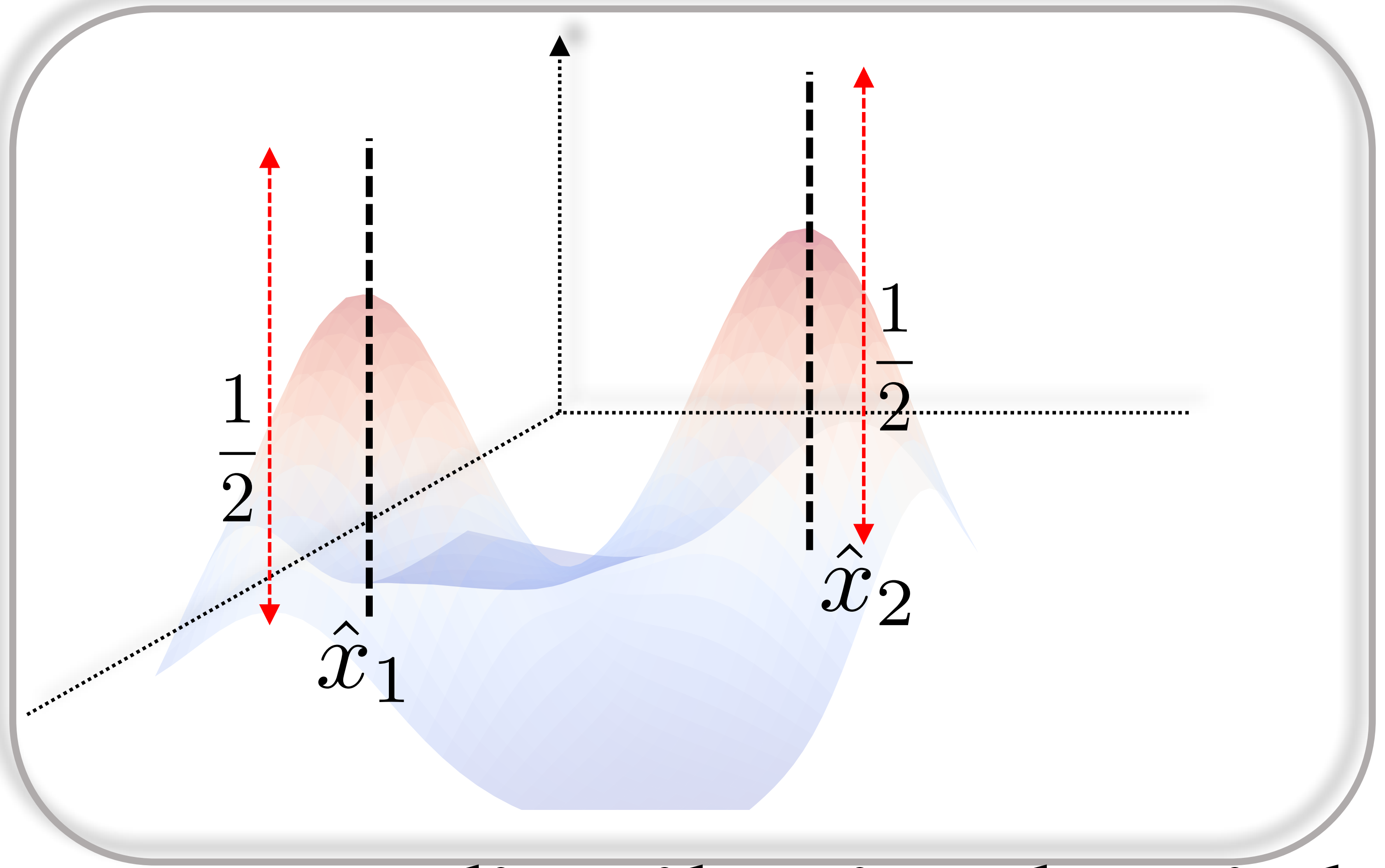


Worst-case distribution by Sinkhorn DRO

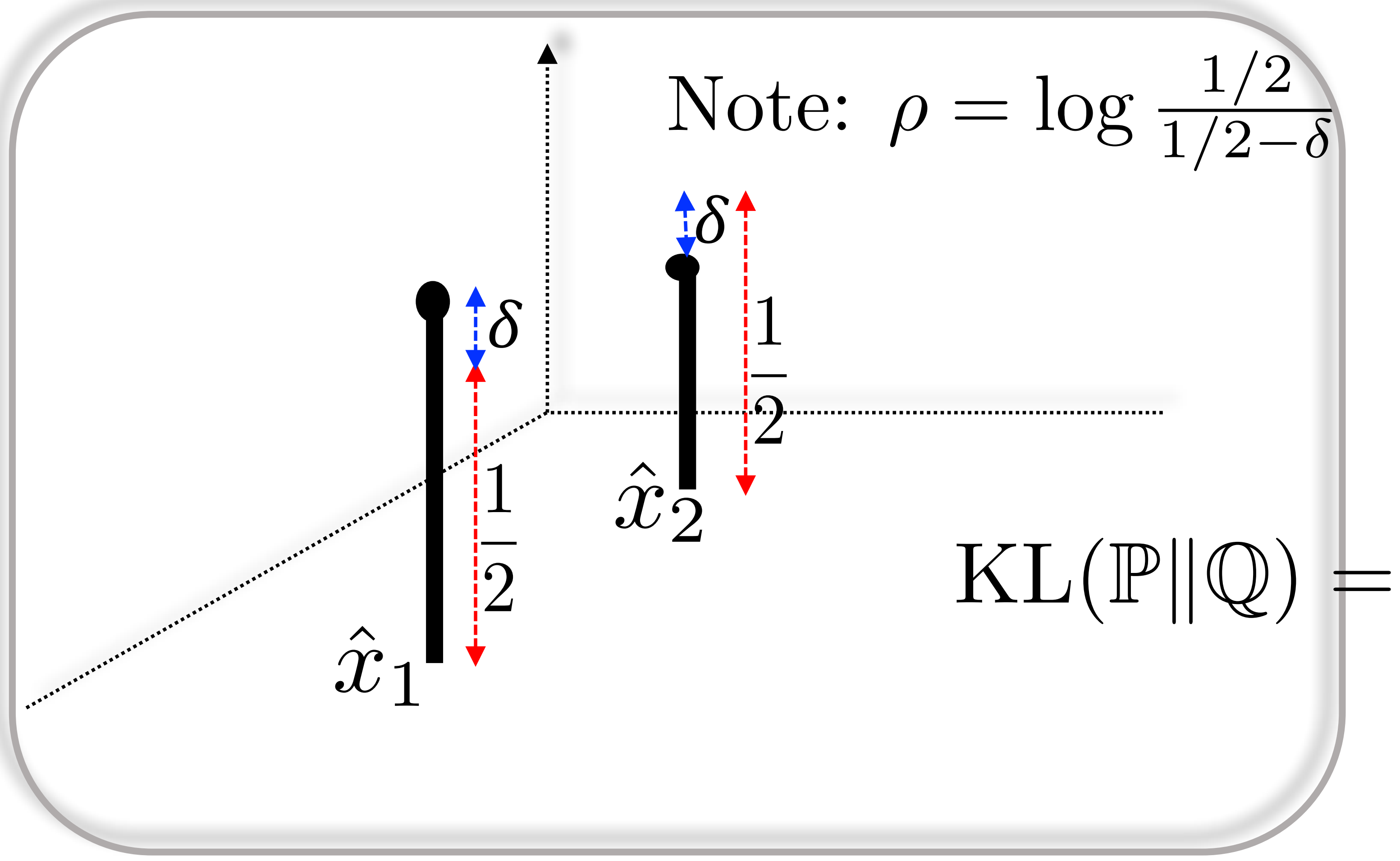
Comparison



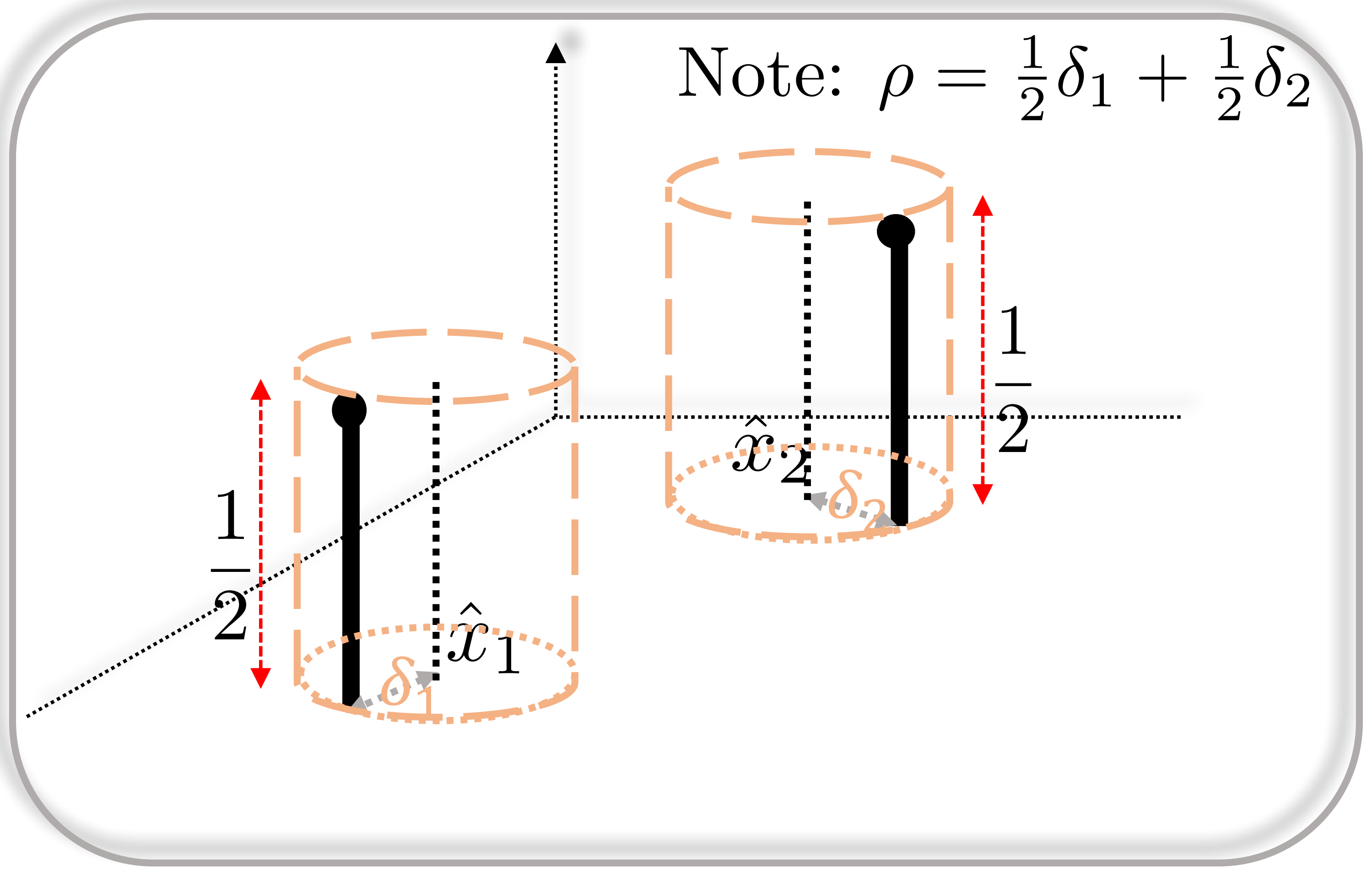
Empirical Distribution



Worst-case distribution by Sinkhorn DRO



$$\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) = \int \log\left(\frac{\mathrm{d}\mathbb{P}(x)}{\mathrm{d}\mathbb{Q}(x)}\right) \mathrm{d}\mathbb{P}(x)$$



Worst-case distribution by KL DRO

Worst-case distribution by Wasserstein DRO

2. Strong Duality and Related Properties

Strong Dual Reformulation

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

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

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

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

Recovery of Worst-case Distribution

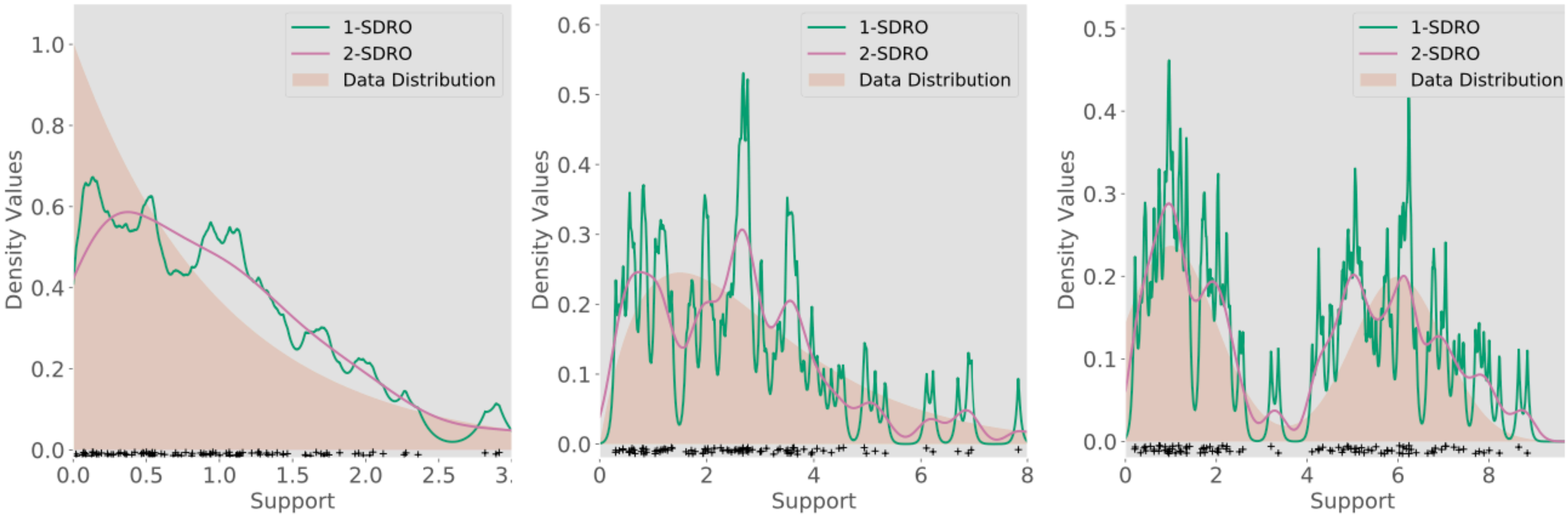
$$\mathbb{P}^* = \arg\max_{\mathbb{P}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] : \ \mathcal{W}_\epsilon(\mathbb{P}_n, \mathbb{P}) \le \rho \right\}$$

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

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

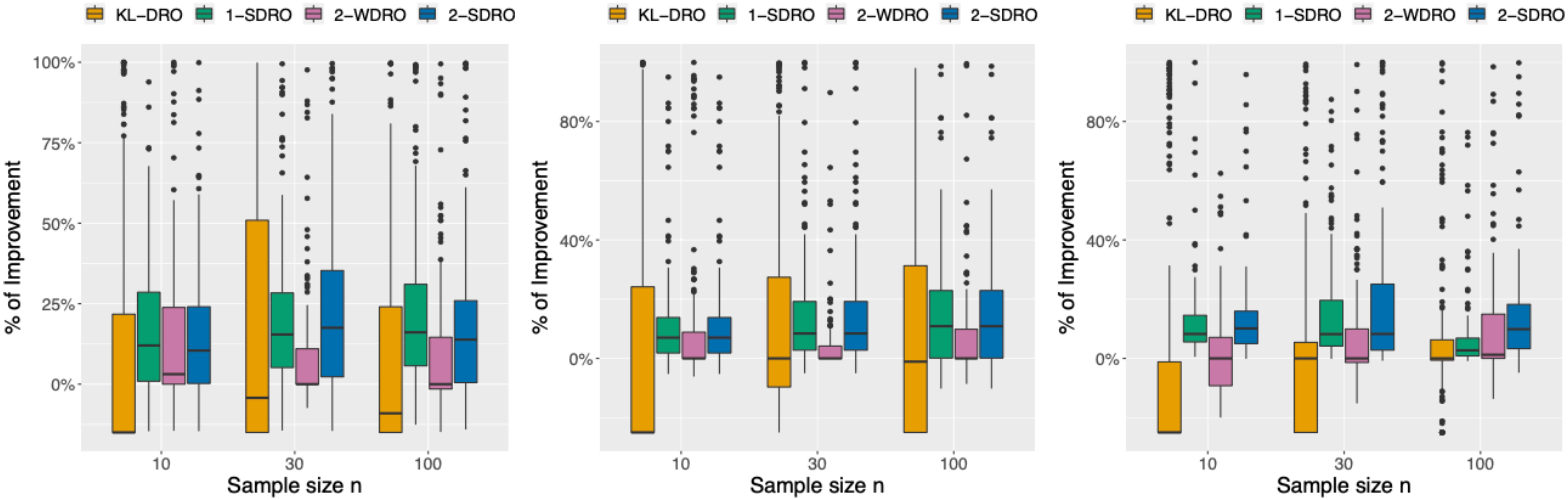
Toy Example: Newsvendor

$$\min_{\beta} \ \mathbb{E}_{z\sim\mathbb{P}_{\text{true}}}[k\beta - u\min\{\beta, z\}], \qquad k = 5, u = 7$$



Toy Example: Newsvendor

$$\min_{\beta} \; \mathbb{E}_{z\sim\mathbb{P}_{\text{true}}}[k\beta - u\min\{\beta, z\}], \qquad k = 5, u = 7$$



Comparison with Wasserstein DRO

Strong duality for **Sinkhorn DRO**:

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

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

Comparison with Wasserstein DRO

Strong duality for **Sinkhorn DRO**:

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

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

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 Wasserstein DRO

Strong duality for **Sinkhorn DRO**:

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

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

Approximate "sup" using log-sum-exp

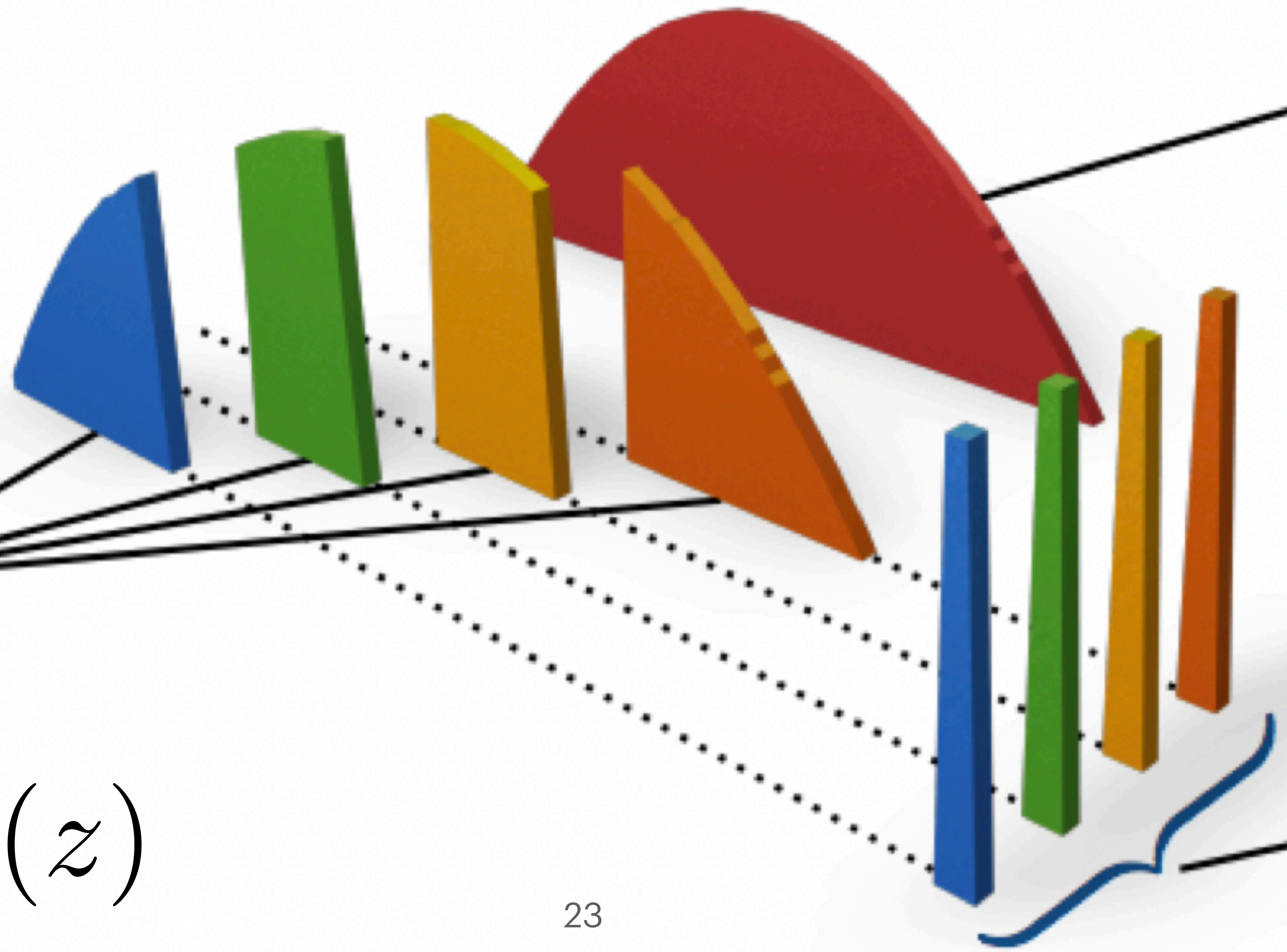
Strong duality for **Wasserstein DRO** ($\epsilon = 0$):

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

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

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

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

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

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

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:**

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

"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 $\mathbf{N}(\hat{x}_i, \epsilon\mathbf{I})$

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

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 $\mathbf{N}(\hat{x}_i, \epsilon\mathbf{I})$

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

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:**

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

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

Stochastic Algorithm for Sinkhorn DRO

- **Approximation Problem:**

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

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

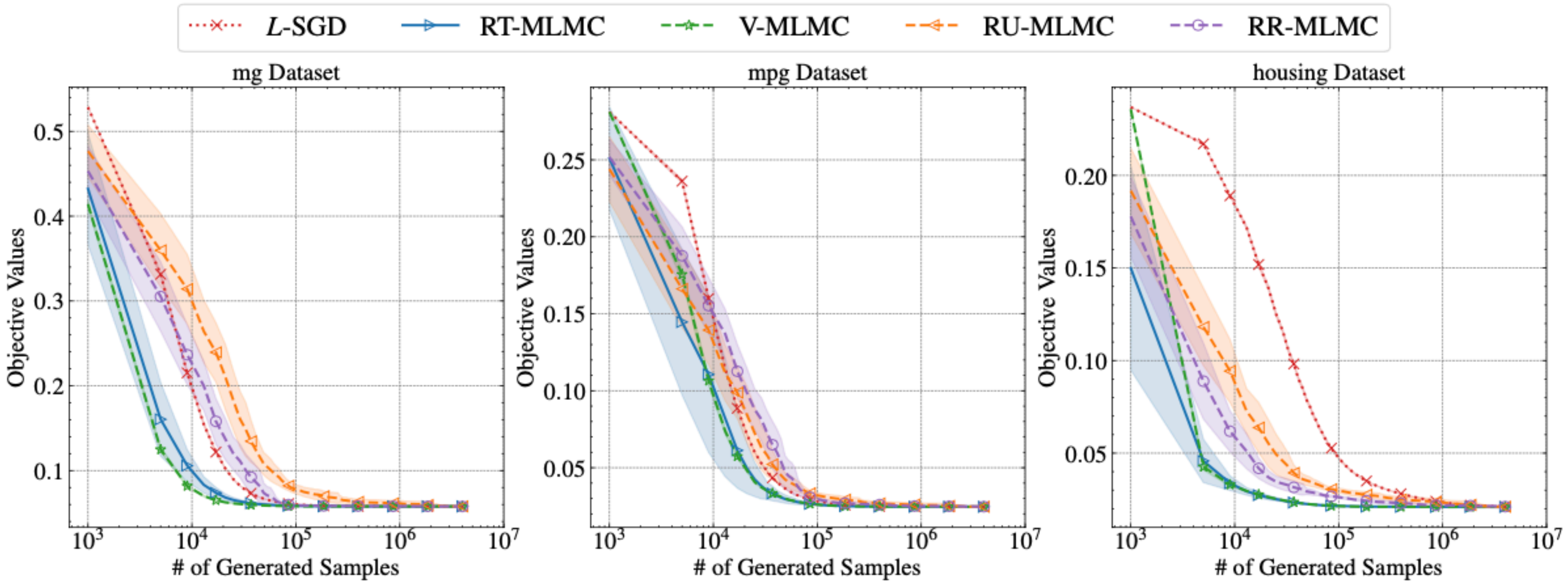
1. This estimator is **unbiased** gradient estimator of F^L
2. This estimator has **significantly lower cost**
3. This estimator has **sufficiently small variance**, due to control variates variance reduction [Nelson, 1990]

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	$\tilde{O}(\delta^{-3})$	$\tilde{O}(\delta^{-6})$	$\tilde{O}(\delta^{-2})$	$\tilde{O}(\delta^{-4})$



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

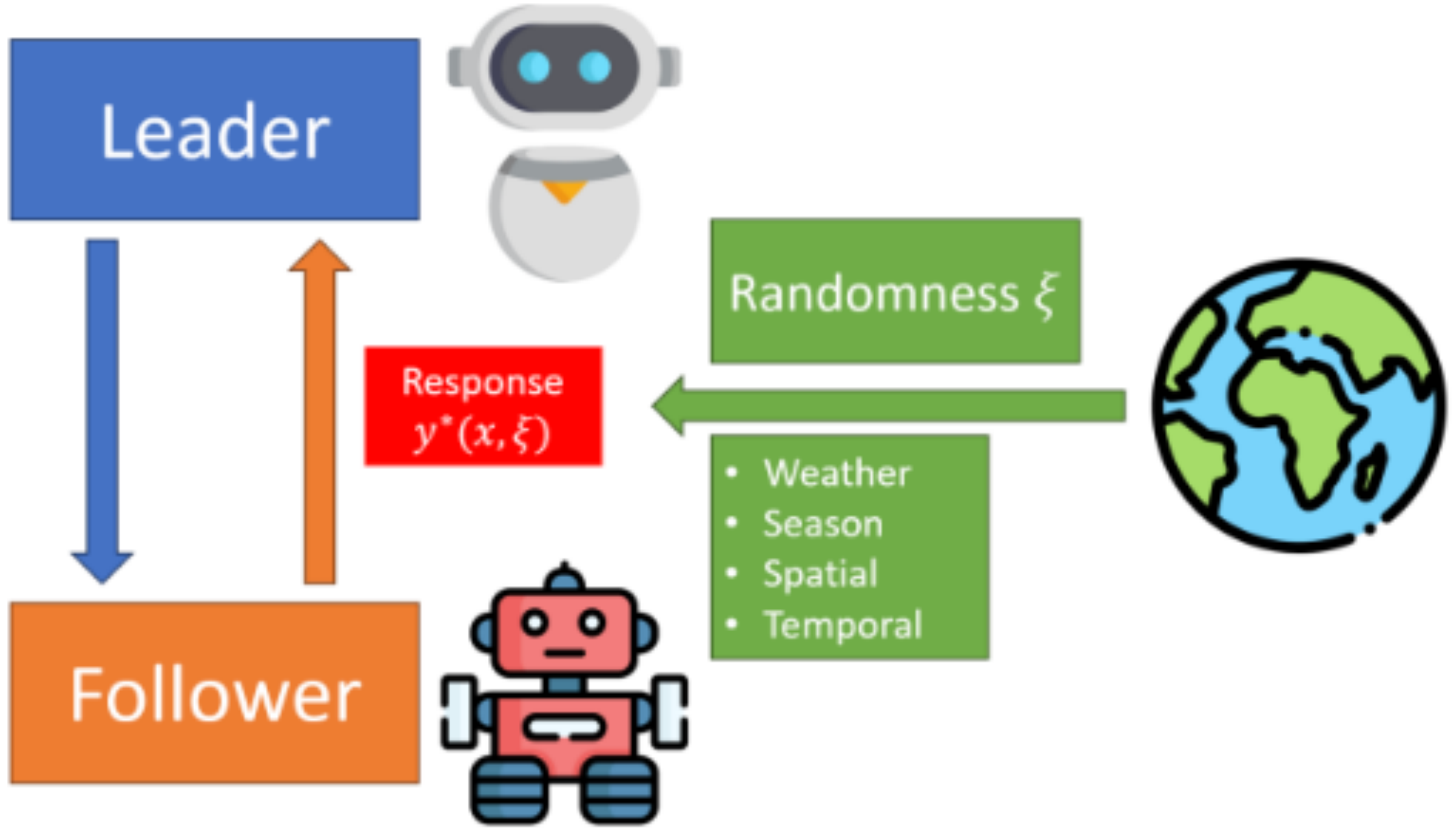
General Optimization Results

- **Goal**: $\min_{\theta} F(\theta)$, whereas **unbiased** gradient of $F(\theta)$ is **not available**!
- **Assumption**:
  - Gradient of approximation objective F^l is easy to obtain
  - $|F^l(\theta) - F(\theta)| = O(2^{-l})$ or $\|\nabla F^l(\theta) - \nabla F(\theta)\|^2 = O(2^{-l})$
- **Examples**:

Contextual Bilevel Optimization

$$\text{minimize} \quad F(\theta) \triangleq \mathbb{E}_{\xi}\left[f(\theta, y^*(\theta;\xi))\right]$$

$$\text{where} \quad y^*(\theta;\xi) \triangleq \operatorname{argmin}_y \mathbb{E}_{\eta\sim\mathbb{P}_{\eta|\xi}}\left[g(x,y;\eta,\xi)\right], \quad \forall \xi$$



Hu Y, **Wang J**, Xie Y, Krause A, Kuhn D (2023) Contextual stochastic bilevel optimization. *NIPS'23*

Hu Y, **Wang J**, Chen X, He N (2024) Multi-level Monte-Carlo Gradient Methods for Stochastic Optimization with Biased Oracles. *arXiv preprint*

General Optimization Results

- **Goal**: min $F(\theta)$, whe... **able**!

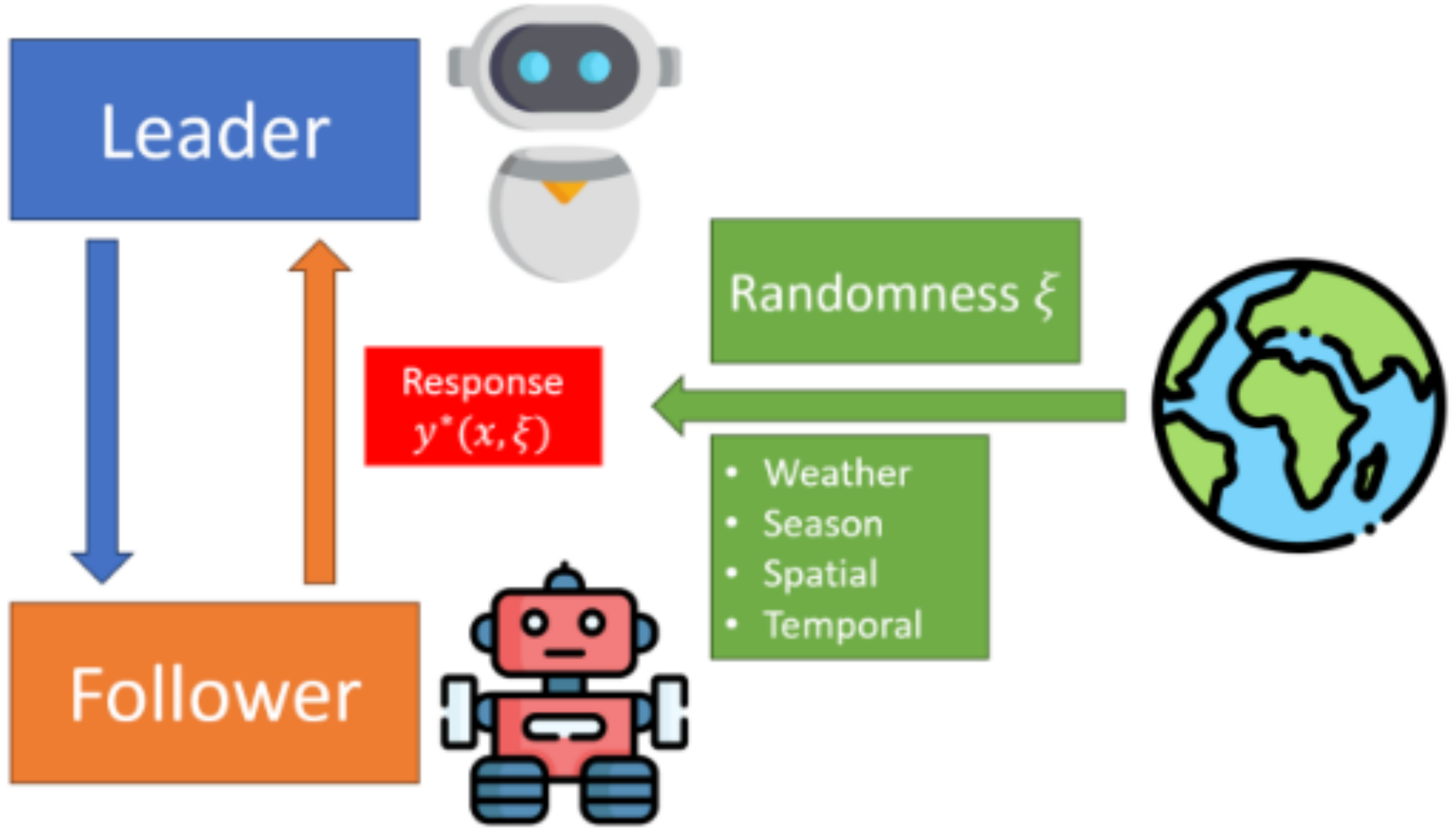
Random Sampling Gradient Estimator Achieves Optimal Complexity on those examples!

- **As**...
  - ...
  - ... $\|^2 = O(2^{-l})$
- **Examples**:

Contextual Bilevel Optimization

$$\text{minimize} \quad F(\theta) \triangleq \mathbb{E}_\xi\left[f(\theta, y^*(\theta;\xi))\right]$$

$$\text{where} \quad y^*(\theta;\xi) \triangleq \operatorname{argmin}_y \mathbb{E}_{\eta\sim\mathbb{P}_{\eta|\xi}}\left[g(x,y;\eta,\xi)\right], \quad \forall \xi$$



Hu Y, **Wang J**, Xie Y, Krause A, Kuhn D (2023) Contextual stochastic bilevel optimization. *NIPS'23*

Hu Y, **Wang J**, Chen X, He N (2024) Multi-level Monte-Carlo Gradient Methods for Stochastic Optimization with Biased Oracles. *arXiv preprint*

4. Numerical Study and Discussion

Extension

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

p-Wasserstein DRO Approximation

[Sinha, Namkoong, Volpi, Duchi, 2020]

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

$$\mathcal{W}_p(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})}\left\{\left(\mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|^p]\right)^{1/p}\right\}$$

Extension

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

p-Wasserstein DRO Approximation

[Sinha, Namkoong, Volpi, Duchi, 2020]

$$\min_{\theta\in\Theta,\lambda\ge 0} \left\{ \lambda\rho^p + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\sup_z \left\{\ell(z;\theta) - \lambda\|z-x\|^p\right\}\right] \right\}$$

- Easy to optimize for large choice of λ

Extension

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

p-**Wasserstein DRO**

$$\min_{\theta\in\Theta,\lambda\ge 0}\left\{\lambda\rho^p+\mathbb{E}_{x\sim\mathbb{P}_n}\left[\sup_z\left\{\ell(z;\theta)-\lambda\|z-x\|^p\right\}\right]\right\}$$

Entropic Regularized *p*-**Wasserstein DRO Approximation**

[Wang, Gao, Xie, 2021]

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

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

Extension

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

p-**Wasserstein DRO**

$$\min_{\theta\in\Theta,\lambda\ge 0} \left\{ \lambda\rho^p + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\sup_z \left\{ \ell(z;\theta) - \lambda\|z-x\|^p \right\} \right] \right\}$$

Entropic Regularized *p*-Wasserstein DRO Approximation

[Wang, Gao, Xie, 2021]

$$\min_{\theta\in\Theta,\lambda\ge 0} \left\{ \lambda\bar{\rho} + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right] \right] \right\}$$

$$\frac{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)}{\mathrm{d}z} \propto e^{-\|z-x\|^p/\epsilon}$$

- **Wang J**, Gao R, Xie Y (2024) Regularization for Adversarial Robust Learning. *arXiv preprint arXiv:2109.11926*

Extension

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

How about adding regularization directly?

***p*-Wasserstein DRO**

$$\min_{\theta\in\Theta,\lambda\ge 0}\left\{\lambda\rho^p+\mathbb{E}_{x\sim\mathbb{P}_n}\left[\sup_z\left\{\ell(z;\theta)-\lambda\|z-x\|^p\right\}\right]\right\}$$

Entropic Regularized *p*-Wasserstein DRO Approximation

[Wang, Gao, Xie, 2021]

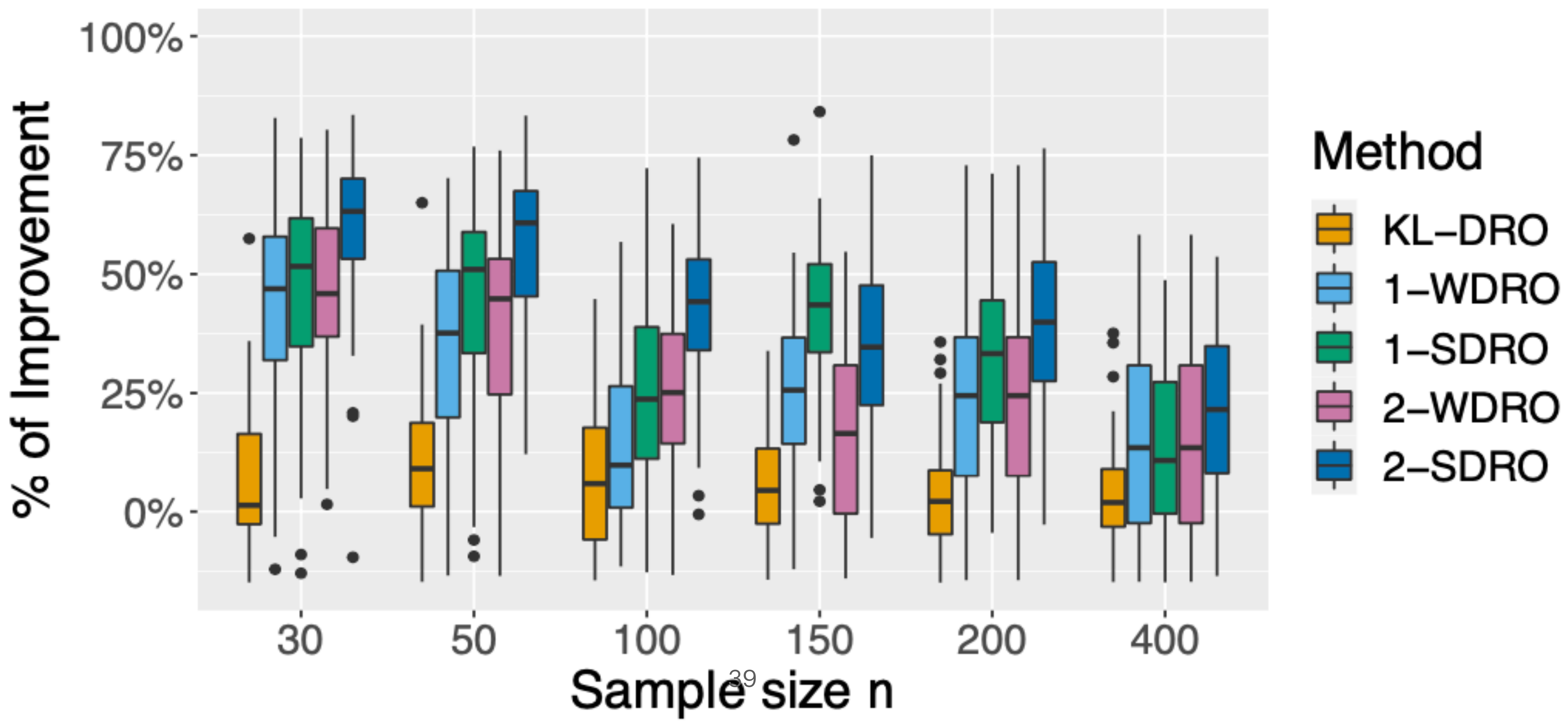
$$\min_{\theta\in\Theta,\lambda\ge 0}\left\{\lambda\bar{\rho}+\mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\right]\right\}$$

$$\frac{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)}{\mathrm{d}z}\propto e^{-\|z-x\|^p/\epsilon}$$

- **Wang J**, Gao R, Xie Y (2024) Regularization for Adversarial Robust Learning. *arXiv preprint arXiv:2109.11926*

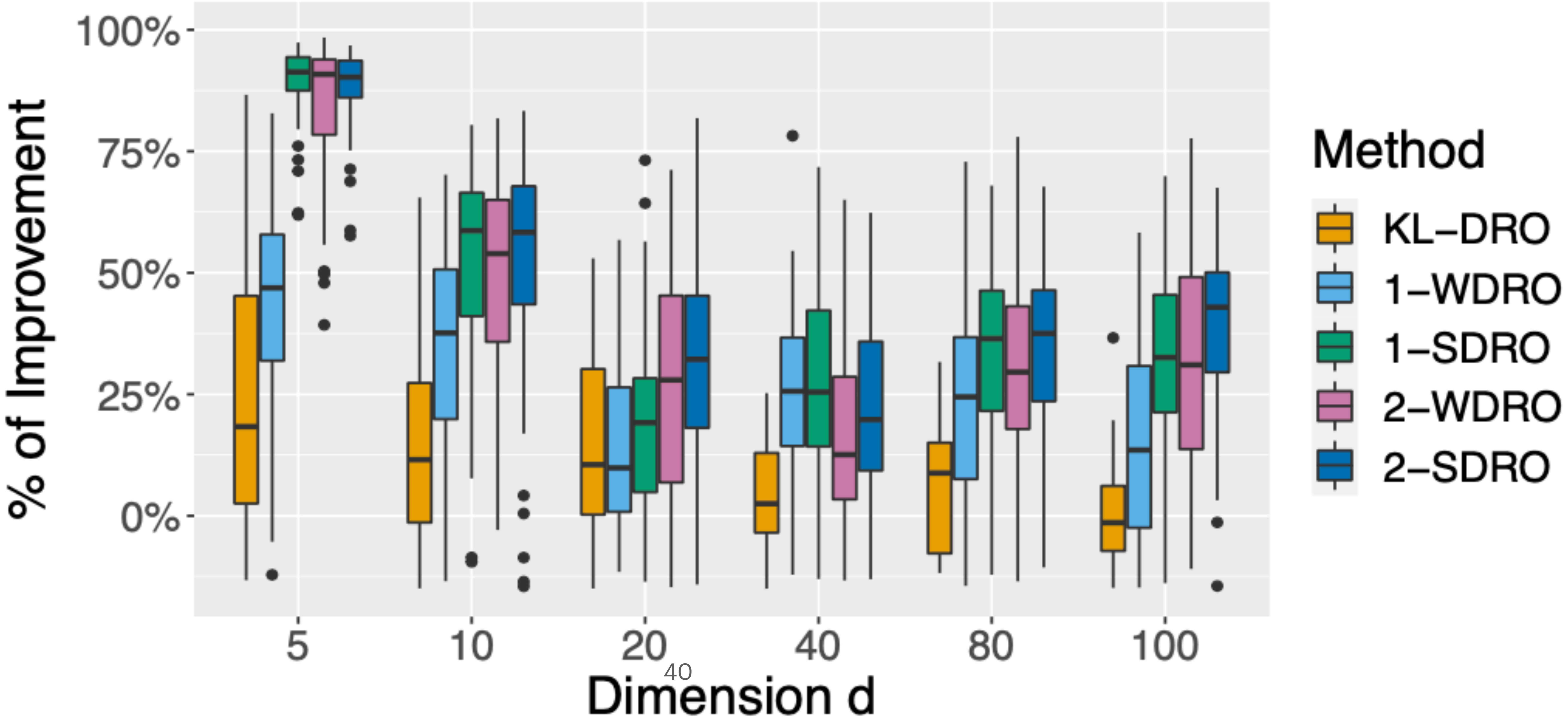
Numerical Results

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



Numerical Results

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



Numerical Results

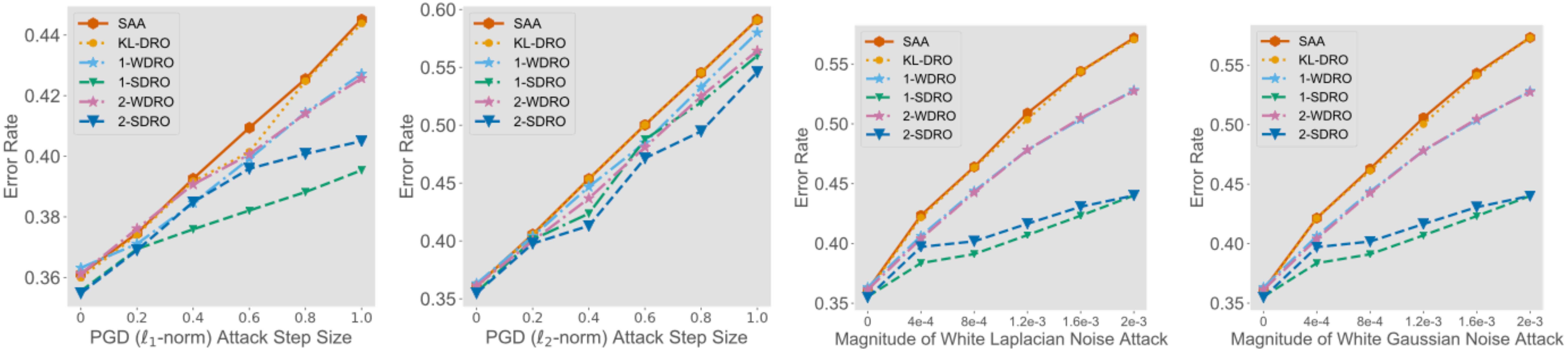
$$\min_{B \in \mathbb{R}^{d \times C}} \quad \mathbb{E}_{(x,\mathbf{y}) \sim \mathbb{P}_{\text{true}}} \left[h_B(x, \mathbf{y}) \right], \quad h_B(x, \mathbf{y}) = -\mathbf{y}^\top B^\top x + \log\left(1^\top e^{B^\top x}\right).$$

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

Numerical Results

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

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



- Computational 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

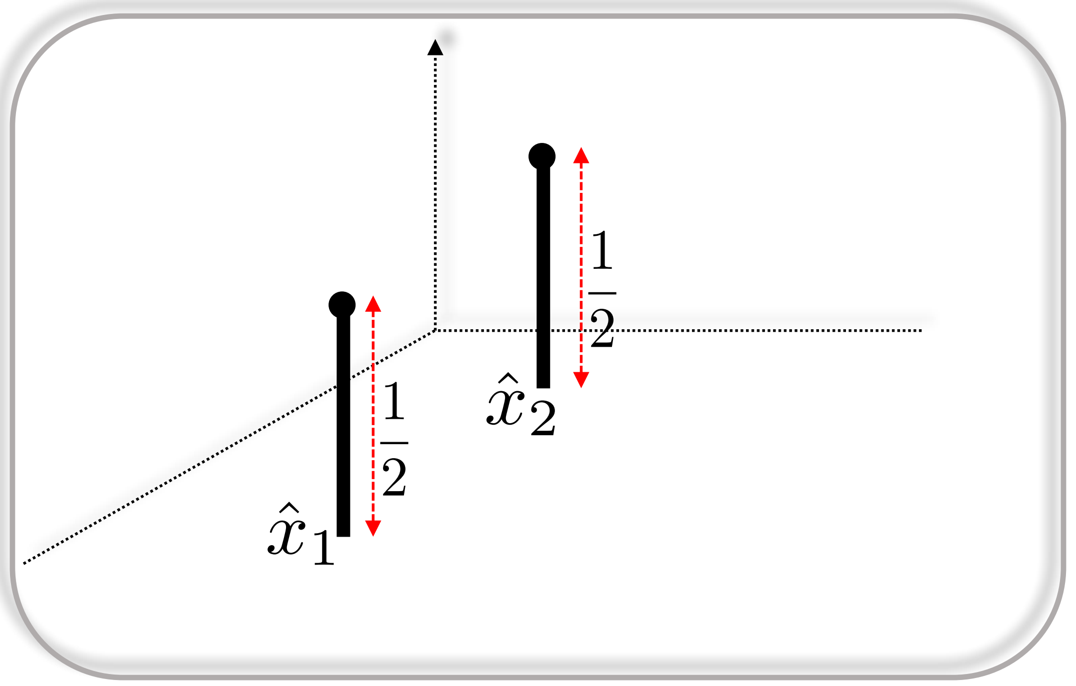
Conclusion

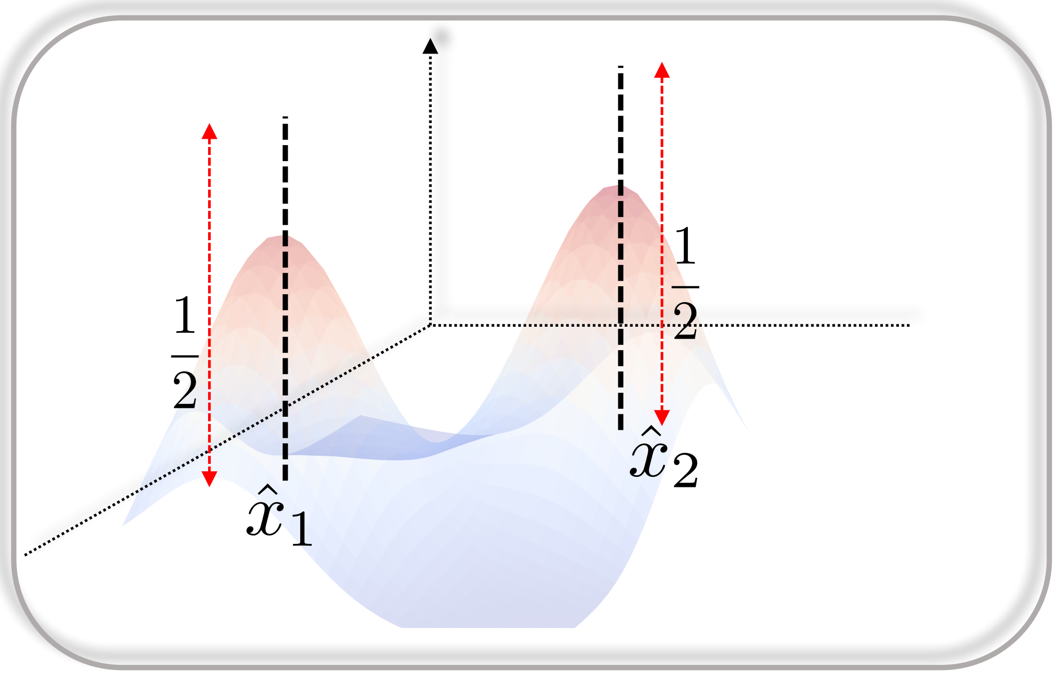
- Sinkhorn DRO is a great notion of DRO models:

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

2. Scalable computation by first-order method;

3. **Connections with regularized machine learning**





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

Related References

- **Wang J**, Gao R, Xie Y (2023) Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926 [Major Revision at Operations Research]*
- **Wang J**, Gao R, Xie Y (2024) Regularization for Adversarial Robust Learning. *arXiv preprint arXiv:2109.11926*
- **Wang J**, Gao R, Zha H. Reliable off-policy evaluation for reinforcement learning. Operations Research, 2024, 72(2): 699-716
- Hu Y, **Wang J**, Chen X, He N (2024) Multi-level Monte-Carlo Gradient Methods for Stochastic Optimization with Biased Oracles. *arXiv preprint arXiv:2408.11084*
- **Wang J** (2023) Reliable Offline Pricing in eCommerce Decision-Making: A Distributionally Robust Viewpoint, Finalist of INFORMS Pricing Competition
- Hu Y, **Wang J**, Xie Y, Krause A, Kuhn D (2023) Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems* 36