Iterative Sampling Methods for Solving Sinkhorn Distributionally Robust Optimization

Jie Wang

The Chinese University of Hong Kong, Shenzhen

School of Artificial Intelligence, School of Data Science

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

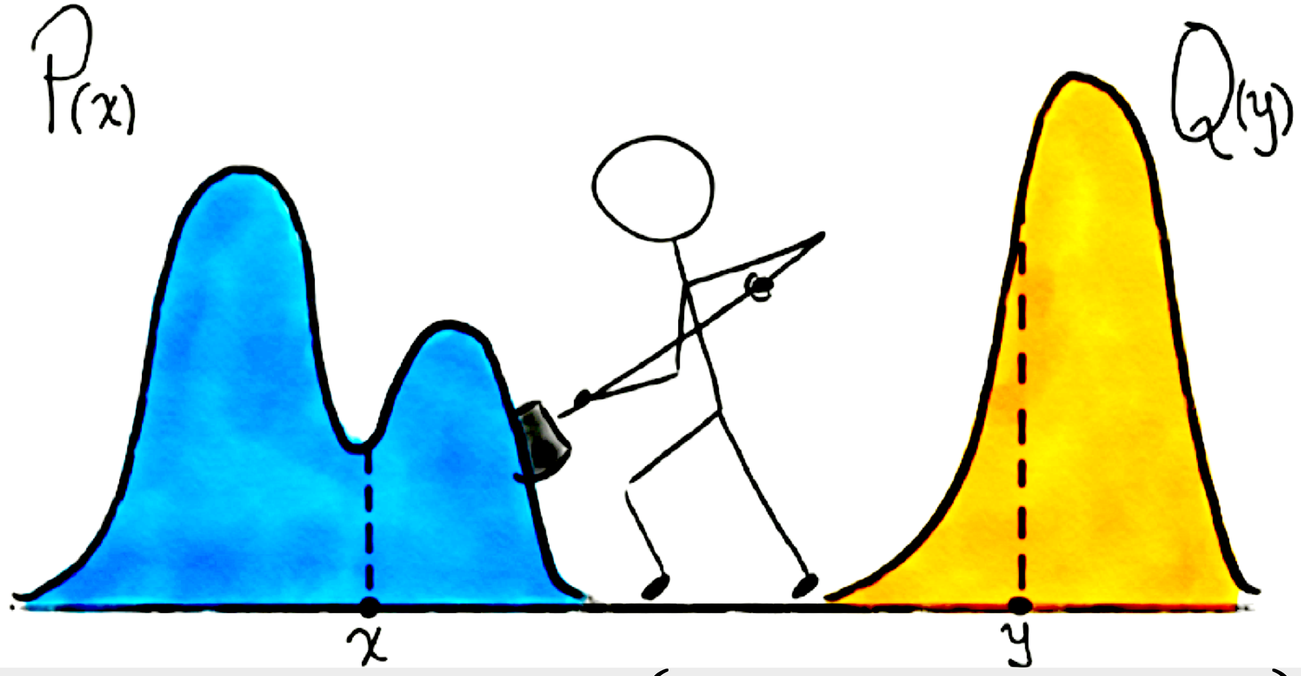
SCHOOL OF DATA SCIENCE
數據科學學院

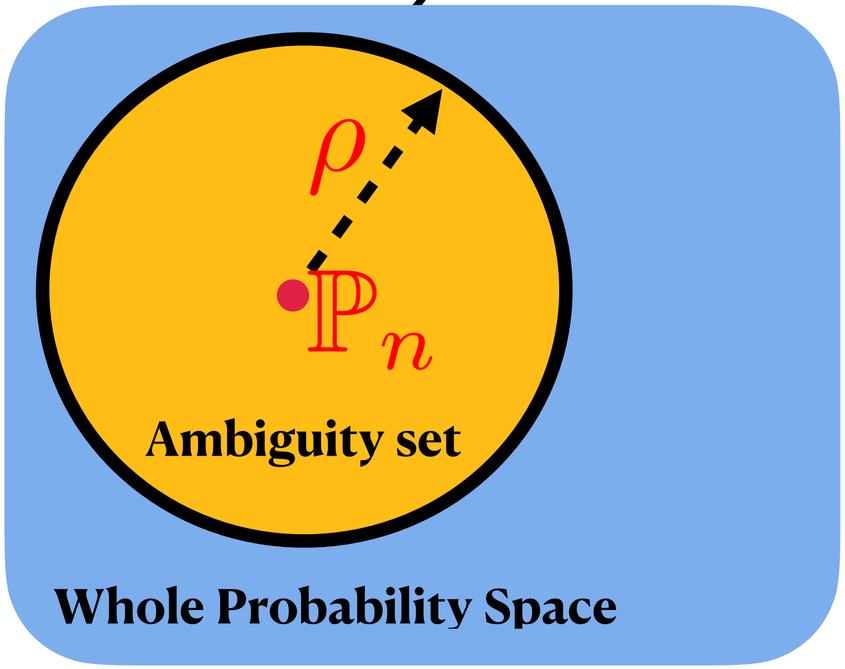
人工智能學院
School of Artificial Intelligence

APORS
Association of Asia-pacific Operational Research Societies
A Regional Grouping within IFORS

Wasserstein Distributionally Robust Optimization

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





$$\mathcal{W}(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{ \mathbb{E}_{(x,y)\sim\gamma}\left[\|x-y\|^2\right] \right\}_2$$

- 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

Literature Review on Infinite-Dimensional Optimization

$$\min_{\rho \in \mathscr{P}(\mathbb{R}^d)} \quad \mathscr{F}(\rho)$$

1. Grid and optimize: replace whole space by N points $\{x_1, \ldots, x_N\}$

$$\rho \approx \sum_{i=1}^{N} p_i \delta_{x_i}$$

Cons: Curse of dimensionality

Literature Review on Infinite-Dimensional Optimization

$$\min_{\rho \in \mathscr{P}(\mathbb{R}^d)} \quad \mathscr{F}(\rho)$$

1. Grid and optimize
2. Duality [Shapiro A, etc...]

Cons: Specialize to problem structures

Literature Review on Infinite-Dimensional Optimization

$$\min_{\rho \in \mathscr{P}(\mathbb{R}^d)} \quad \mathscr{F}(\rho)$$

1. Grid and optimize
2. Duality
3. Frank-wolfe algorithm **[Jaggi M, 2013]**:
   Iteratively,

Cons: Subproblems are generally difficult to solve

$$\nu = \operatorname*{argmin}_{\mu \in \mathscr{P}} \mathscr{F}'_{\mu}(\mu_k), \mu_{k+1} = (1-\beta)\mu_k + \beta\nu$$

Literature Review on Infinite-Dimensional Optimization

$$\min_{\rho \in \mathscr{P}(\mathbb{R}^d)} \quad \mathscr{F}(\rho)$$

1. Grid and optimize
2. Duality
3. Frank-wolfe algorithm
4. Particle Gradient Descent **[Chizat L, Bach F, 2018]**

Cons: Sharp convergence generally applicable to entropic regularized objective

Initialize $\mu = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$, update particles $\{x_i\}_i$ using gradient descent

Sinkhorn Discrepancy

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

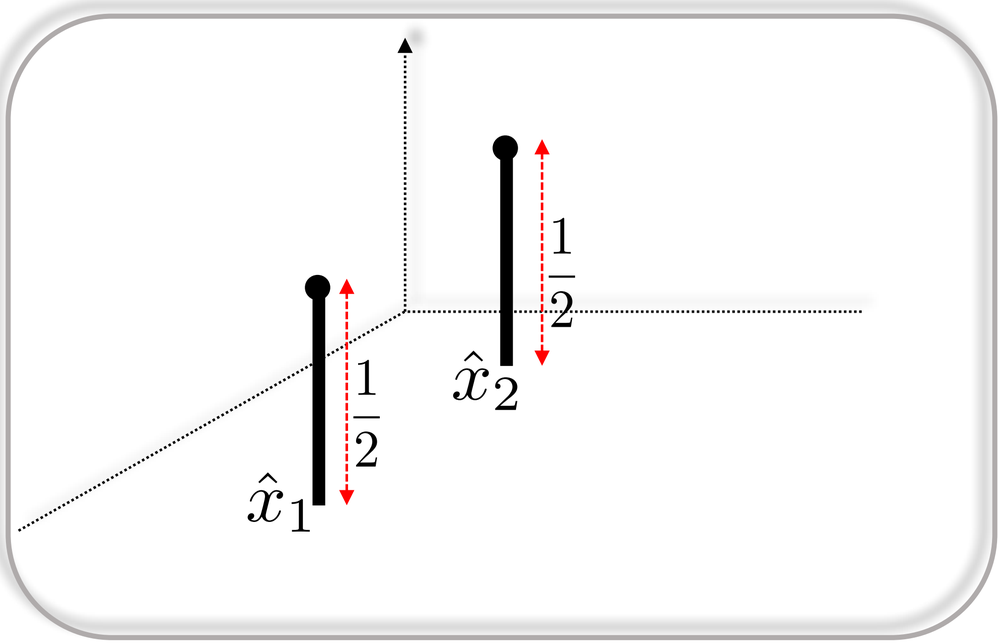
Historical Review:

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

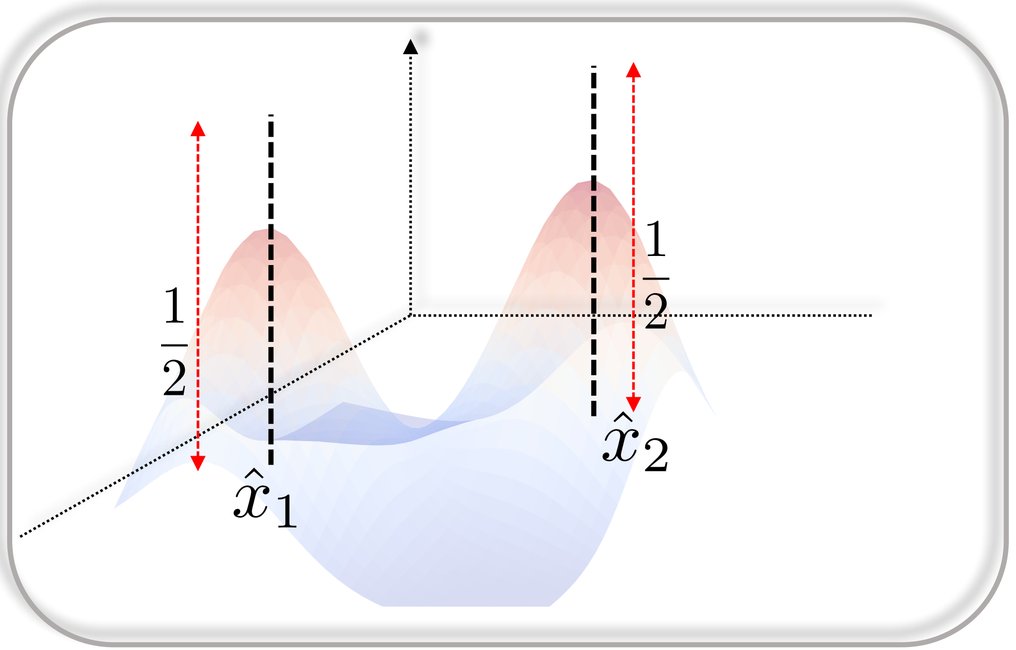
Main Framework

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

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**

Monte Carlo Sampling

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

- Jie Wang, Rui Gao, Yao Xie (2025) Sinkhorn Distributionally Robust Optimization. *Operations Research*

Monte Carlo Sampling

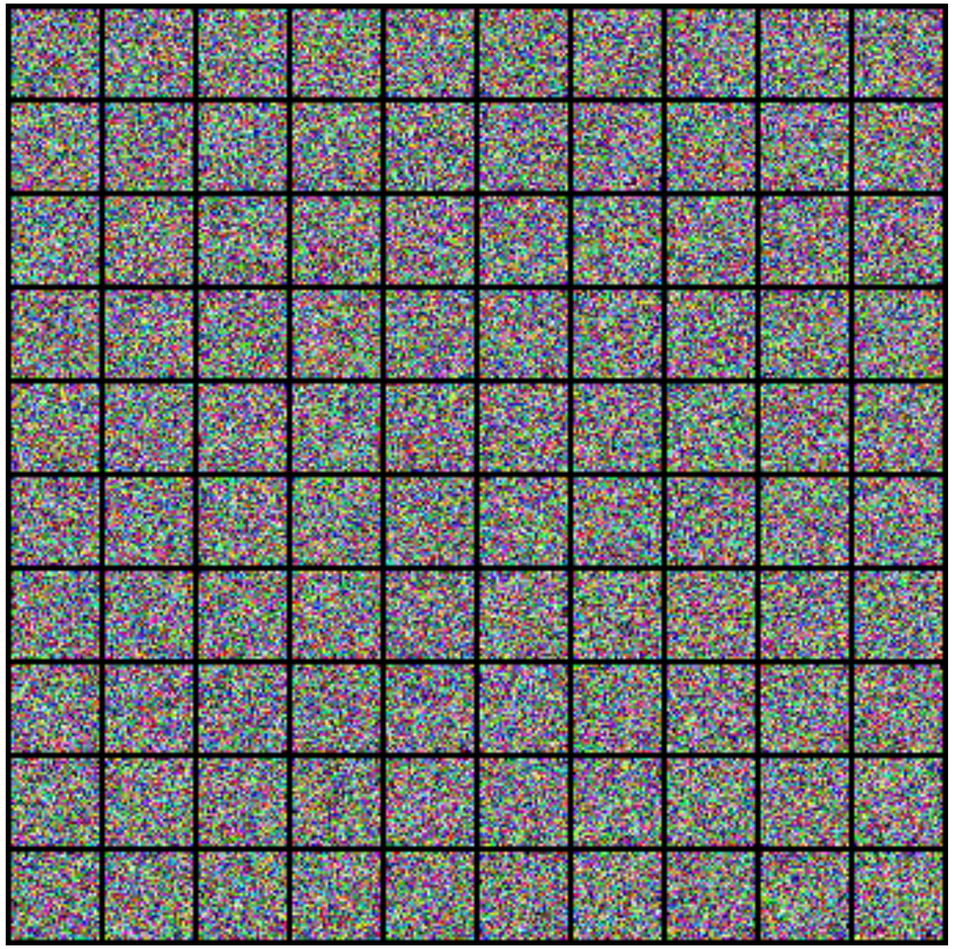
$$\min_{\theta,\lambda\geq 0} \left\{ \lambda\bar{\rho} + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbf{N}(x,\epsilon\mathbf{I})}\left[e^{\ell(z;\theta)/(\lambda\epsilon)}\right]\right] \right\}$$

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

- A. Shapiro

- Jie Wang, Rui Gao, Yao Xie (2025) Sinkhorn Distributionally Robust Optimization. *Operations Research*

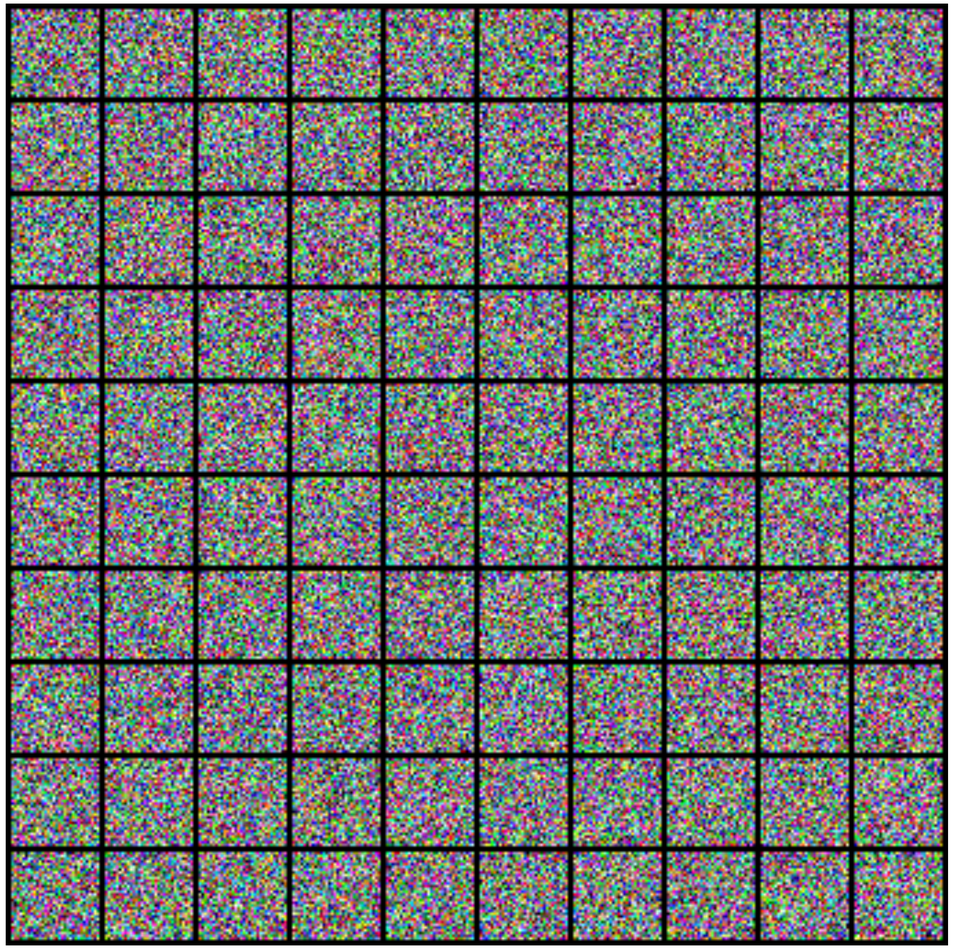
Diffusion Model



- Given training samples $\{x^{(1)}, \ldots, x^{(n)}\}$ i.i.d. from $\mathbb{P}_{\text{true}}$
- **Goal of Generative AI**: Generate a sample $\hat{x}$ with distribution close to $\mathbb{P}_{\text{true}}$
  - Step 1: Learn score function $s(x) \approx \log \mathbb{P}_{\text{true}}(x)$;
  - Step 2: Sampling from (unnormalized density) $\exp(s(x))$ using Langevin dynamics.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

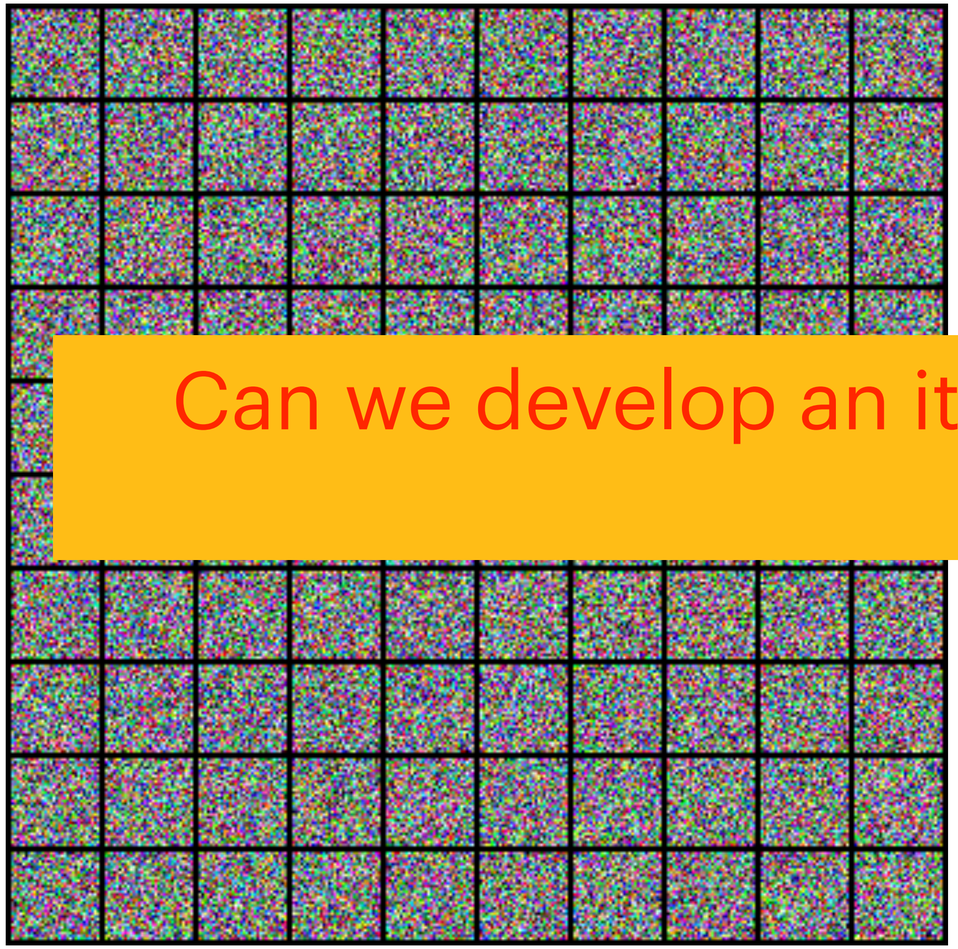
Diffusion Model



- Given training samples $\{x^{(1)}, \ldots, x^{(n)}\}$ i.i.d. from $\mathbb{P}_{\text{true}}$
- **Goal of Generative AI**: Generate a sample $\hat{x}$ with distribution close to $\mathbb{P}_{\text{true}}$
  - Step 1: Learn score function $s(x) \approx \log \mathbb{P}_{\text{true}}(x)$;
  - Step 2: Sampling from (unnormalized density) $\exp(s(x))$ using Langevin dynamics.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Diffusion Model



- Given training samples $\{x^{(1)}, \ldots, x^{(n)}\}$ i.i.d. from $\mathbb{P}_{\text{true}}$
- **Goal of Generative AI**: Generate a sample

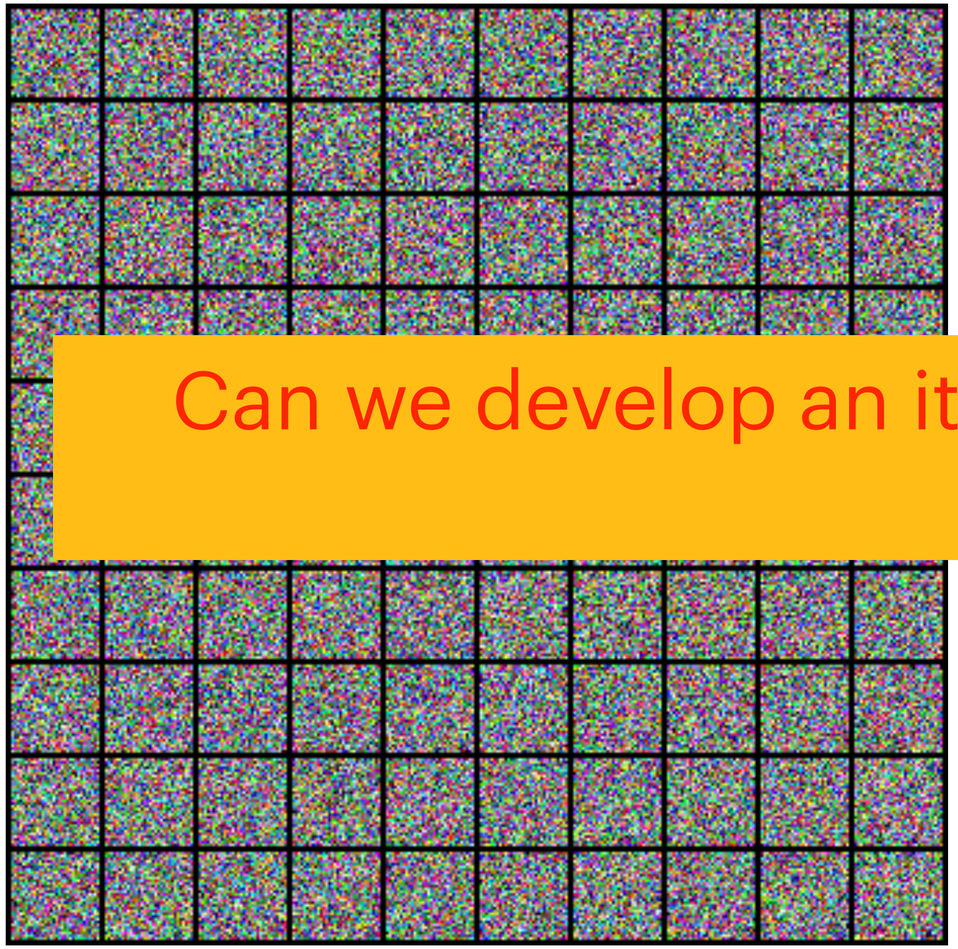
Can we develop an iterative sampling approach to solving Sinkhorn DRO?

$s(x) \approx \log \mathbb{P}_{\text{true}}(x)$;

- Step 2: Sampling from (unnormalized density) $\exp(s(x))$ using Langevin dynamics.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Diffusion Model



- Given training samples $\{x^{(1)}, \ldots, x^{(n)}\}$ i.i.d. from $\mathbb{P}_{\text{true}}$
- **Goal of Generative AI**: Generate a sample

Can we develop an iterative sampling approach to solving Sinkhorn DRO?

$s(x) \approx \log \mathbb{P}_{\text{true}}(x)$;

- Step 2: Sampling from (unnormalized density) $\exp(s(x))$ using Langevin dynamics.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Diffusion for Sinkhorn DRO

$$\min_{\theta} \ \mathbb{E}_{z\sim\mathbb{P}^*_\theta}[\ell(\theta; z)]$$

where $\mathbb{P}^*_\theta = \frac{1}{n}\sum_{i\in[n]} \mathbb{P}^*_{i,\theta}$

and $$\boxed{\frac{\mathrm{d}\mathbb{P}^*_{i,\theta}}{\mathrm{d}z}(z) \propto \exp\left(\frac{\ell(z;\theta)}{\lambda\epsilon} - \frac{\|x^{(i)} - z\|_2^2}{2\epsilon}\right)}$$

Efficient sampling from $\mathbb{P}^*_{i,\theta}$ requires **log-Sobolev inequality (LSI)**

If converged?

False

Sample $i \in [n]$ and next sample $z \sim \mathbb{P}^*_{i,\theta}$

Generate gradient estimator $\nabla\ell(\theta, z)$ and update θ

Sampling via Langevin Dynamics

$$\frac{\mathrm{d}\mathbb{P}^*_{i,\theta}}{\mathrm{d}z}(z) \propto \exp\left(\frac{\ell(z;\theta)}{\lambda\epsilon} - \frac{\|x^{(i)} - z\|_2^2}{2\epsilon}\right)$$

$$\mathbb{P}^*_{i,\theta} = \mathrm{argmax}_{\mathbb{P}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \frac{\lambda}{2}\mathbb{E}_{z\sim\mathbb{P}}[\|z - x^{(i)}\|_2^2] - \lambda\epsilon\mathscr{H}(\mathbb{P}) \right\}$$

$$z_{k+1} = z_k + \eta \cdot \left(\nabla_z \ell(z;\theta) - \lambda(z - x^{(i)})\right) + \sqrt{2\eta\lambda\epsilon} \cdot \mathbf{N}(0, I_d)$$

Demo

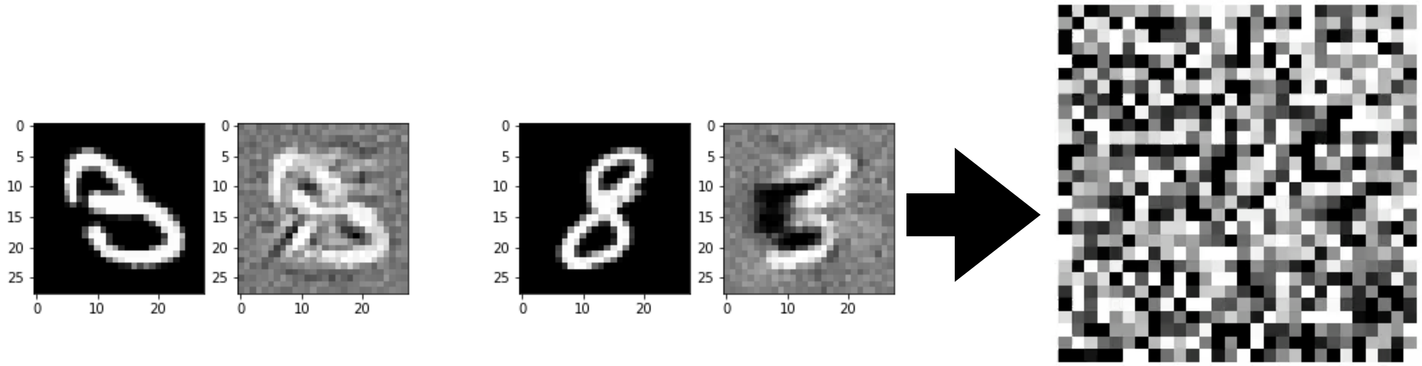
$$\min_{\theta} \left\{ \sup_{\mathbb{P}:\ \mathcal{W}_\epsilon(\mathbb{P}_n, \mathbb{P}) \leq \rho} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] \right\}$$

- $\ell(z;\theta)$ is a convolutional neural net with logistic loss
- Classification task with 5 samples from MNIST digit 3 and 5 samples from digit 8

Demo

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

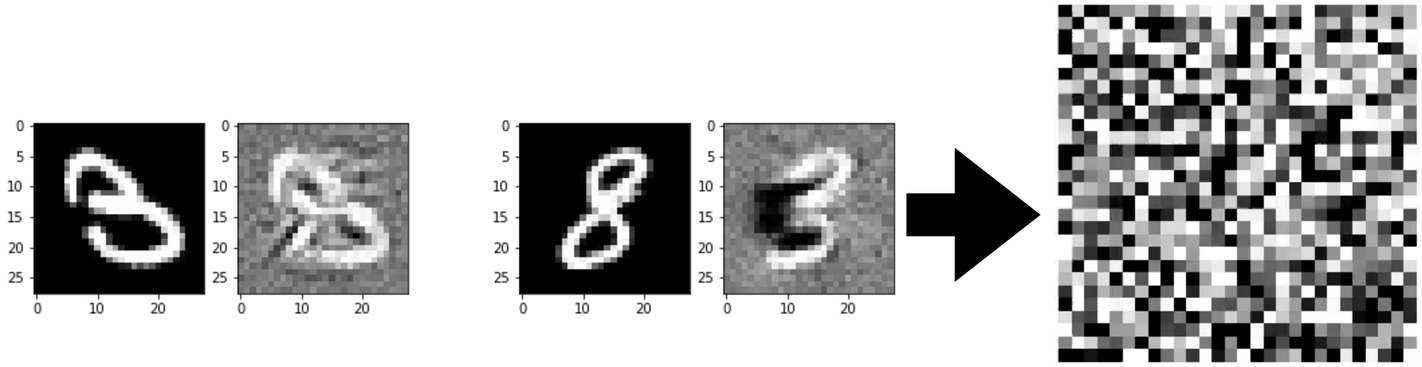
- $\ell(z;\theta)$ is a convolutional neural net with logistic loss
- Classification task with 5 samples from MNIST digit 3 and 5 samples from digit 8



Demo

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

- $\ell(z;\theta)$ is a convolutional neural net with logistic loss
- Classification task with 5 samples from MNIST digit 3 and 5 samples from digit 8



1. Sinkhorn DRO can be viewed as a "task-aware" generative AI (GenAI)
2. Conventional GenAI: target distribution is given as a *priori*

Bilevel Reformulation for Sinkhorn DRO

$$\min_{\theta} \; F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mu_{i,\theta}^*}[\ell(z;\theta)]$$

$$\mu_{i,\theta}^* = \operatorname*{argmin}_{\mu \in \mathscr{P}} \mathscr{D}_{\mathrm{KL}}\big(\mu \| u_{i,\theta}(\,\cdot\,)\big), \quad i \in [n]$$

$$u_{i,\theta}(z) \propto \exp\left(\frac{\ell(z;\theta)}{\lambda\epsilon} - \frac{1}{2\epsilon}\|z - x^{(i)}\|_2^2\right)$$

Double Loop Algorithm:
Iteratively,

1. Fix upper-level decision
2. Run inner-loop to find near-optimal lower-level solution
3. Update upper-level solution

For large n, randomly select a batch of lower-level problems

Bilevel Reformulation for Sinkhorn DRO

$$\min_{\theta} \; F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mu^*_{i,\theta}}[\ell(z;\theta)]$$

$$\mu^*_{i,\theta} = \operatorname*{argmin}_{\mu \in \mathscr{P}} \mathscr{D}_{\mathrm{KL}}\big(\mu \| u_{i,\theta}(\,\cdot\,)\big), \quad i \in [n]$$

$$u_{i,\theta}(z) \propto \exp\left(\frac{\ell(z;\theta)}{\lambda\epsilon} - \frac{1}{2\epsilon}\|z - x^{(i)}\|_2^2\right)$$

Single Loop Algorithm:
Iteratively,

1. Fix upper-level decision
2. One-step update of lower-level solution
3. One-step update of upper-level solution

Randomly update a batch of lower-level problem solutions, leave others unchanged.

Bilevel Reformulation for Sinkhorn DRO

$$\min_{\theta} \; F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mu^*_{i,\theta}}[\ell(z;\theta)]$$

$$\mu^*_{i,\theta} = \operatorname*{argmin}_{\mu \in \mathscr{P}} \mathscr{D}_{\mathrm{KL}}\big(\mu \| u_{i,\theta}(\,\cdot\,)\big), \quad i \in [n]$$

$$u_{i,\theta}(z) \propto \exp\left(\frac{\ell(z;\theta)}{\lambda\epsilon} - \frac{1}{2\epsilon}\|z - x^{(i)}\|_2^2\right)$$

Single Loop Algorithm:
Iteratively,

1. Fix upper-level decision
2. One-step update of lower-level solution
3. One-step update of upper-level solution

- Generate gradient estimator for upper-level solution,
- Construct a moving average estimator,
- Do the update.

Single-Loop Algorithm for Sinkhorn DRO

Algorithm 3 Mean-Field Single-Loop Iterative Sampling Algorithm

Require: Stepsize parameters η, τ, initial guess $\mu_0^{(i)}, i \in [n]$, moving average estimator r_0, moving average parameter β_0, number of iterations T.

for $k = 0, 1, 2, \ldots, T-1$ **do**
  Randomly sample indices $I_k \subseteq [n]$
  for $i \in [n]$ **do**
    Update $\mu_{k+1}^{(i)}$ according to (10)
  end for
  Update gradient estimator v_{k+1} according to (11)
  Update $r_{k+1} = (1-\beta_0) r_k + \beta_0 v_{k+1}$
  Update $\theta_{k+1} = \theta_k - \tau \eta r_{k+1}$.
end for
Output $\widehat{\theta}$ uniformly selected from $\{\theta_1, \ldots, \theta_T\}$.

Convergence Analysis

Under LSI assumption on worst-case distribution, it holds that:

Double-loop algorithm finds ϱ-stationary solution with complexity $\mathcal{O}(\varrho^{-6})$.

- Solving Sinkhorn DRO from the dual requires boundness of loss function $\ell(z;\theta)$
- Existing algorithm has complexity $\mathcal{O}(\varrho^{-4})$ **[Hu et. al 2023]** by Multi-level Monte Carlo Technique.

Convergence Analysis

Under LSI assumption on worst-case distribution, it holds that:

Single-loop algorithm finds ϱ-stationary solution with complexity $\mathcal{O}(\varrho^{-6})$.

Proof Outline

Lemma 4.2 (Descent Lemma). *Assume Assumption 3.6(II) holds and the stepsize parameters satisfy* $\eta\tau \leq \frac{1}{2L_{f,2}}$. *Consider the update in Step 8 of Algorithm 3, it holds that*

$$F(\theta_{k+1}) \leq F(\theta_k) + \frac{\tau\eta}{2}\|\nabla F(\theta_k) - r_{k+1}\|_2^2 - \frac{\tau\eta}{2}\|\nabla F(\theta_k)\|_2^2 - \frac{\tau\eta}{4}\|r_{k+1}\|_2^2.$$

Critical Step: Bound the difference between **true hypergradient** $\nabla F(\theta_k)$ and our **gradient estimator** r_{k+1} used in k-th iteration

Proof Outline

Lemma 4.3 (Gradient Difference Lemma). *Assume Assumption 3.6(II) holds, and the parameter* $\beta_0 \in (0, 1]$, *then it holds that*

$$\begin{aligned}\mathbb{E}\left[\|\nabla F(\theta_k) - r_{k+1}\|_2^2 \middle| \mathcal{G}_{k-1}\right] \leq (1-\beta_0)\|\nabla F(\theta_{k-1}) - r_k\|_2^2 + \frac{4L_{f,2}^2\tau^2\eta^2}{\beta_0}\|r_k\|_2^2 \\ + 4\beta_0\left\|\nabla F(\theta_k) - \nabla F(\theta_k; \mu_k^{(1:n)})\right\|_2^2 + \beta_0^2\mathbb{E}\left[\left\|\nabla F(\theta_k; \mu_k^{(1:n)}) - v_{k+1}\right\|_2^2 \middle| \mathcal{G}_{k-1}\right].\end{aligned} \tag{13}$$

On the right-hand-side of (13), it can be shown that the last component can be bounded in terms of the variance of v_{k+1}, and the third component can be bounded as

$$\left\|\nabla F(\theta_k) - \nabla F(\theta_k; \mu_k^{(1:n)})\right\|_2^2 \leq \frac{L_{f,1}^2}{n}\sum_{i\in[n]}\mathcal{D}_{\mathrm{KL}}(\mu_k^{(i)}, \mu_*^{i,\theta_k}).$$

Major Difficulty: Bound difference between true lower-level solution μ_*^{i,θ_k} and estimated lower-level solution $\mu_k^{(i)}$

Proof Outline

Lemma 4.4 (KL-Divergence Bound). *Assume Assumptions 3.2 and 3.6 hold, then it holds that*

$$\sum_{k=0}^{T} \mathbb{E}\left[\mathcal{D}_{\mathrm{KL}}(\mu_k^{(i)} \| \mu_*^{i,\theta_k})\right]$$

$$\leq \frac{4n}{\alpha\tau|I_k|}\mathcal{D}_{\mathrm{KL}}(\mu_0^{(i)} \| \mu_*^{i,\theta_0}) + \frac{12\eta L_{f,1}^2 Tn}{\lambda\epsilon\alpha|I_k|} + \frac{32\tau\epsilon d L_{G,2}^2 T}{\alpha} + \frac{20\eta n}{\lambda\epsilon\alpha|I_k|}\sum_{k=0}^{T-1}\mathbb{E}\left[\|r_{k+1}\|_2^2\right].$$

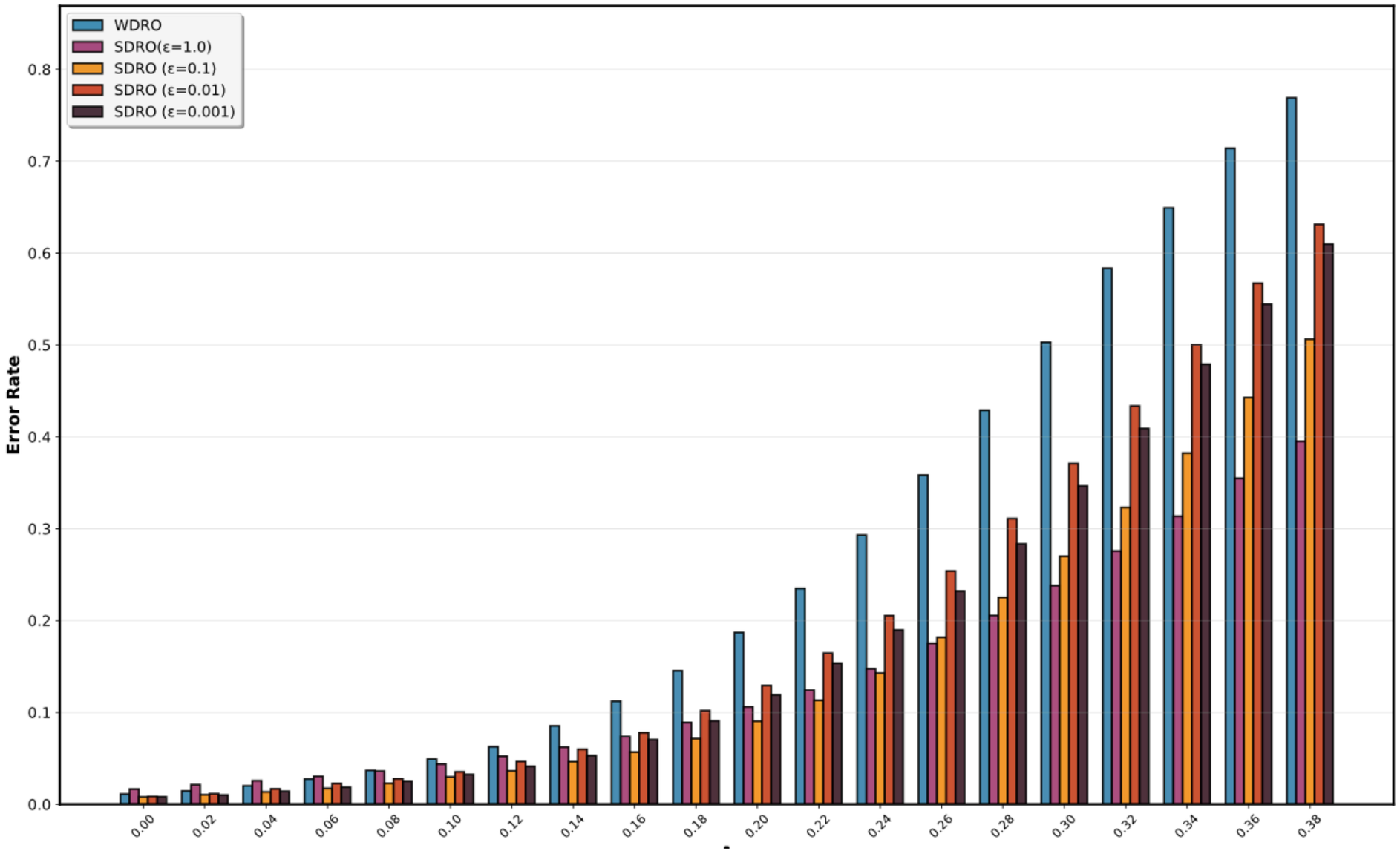
- Many tricks in finite-dimensional bilevel program are **not applicable**, e.g., triangular inequality
- Use KL-divergence as a metric. Analyze the trajectory

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_{\mathrm{KL}}(\rho_t \| \mu_*^{i,\vartheta_t})$$

$$\vartheta_t = \theta_k - t\eta r_{k+1}$$

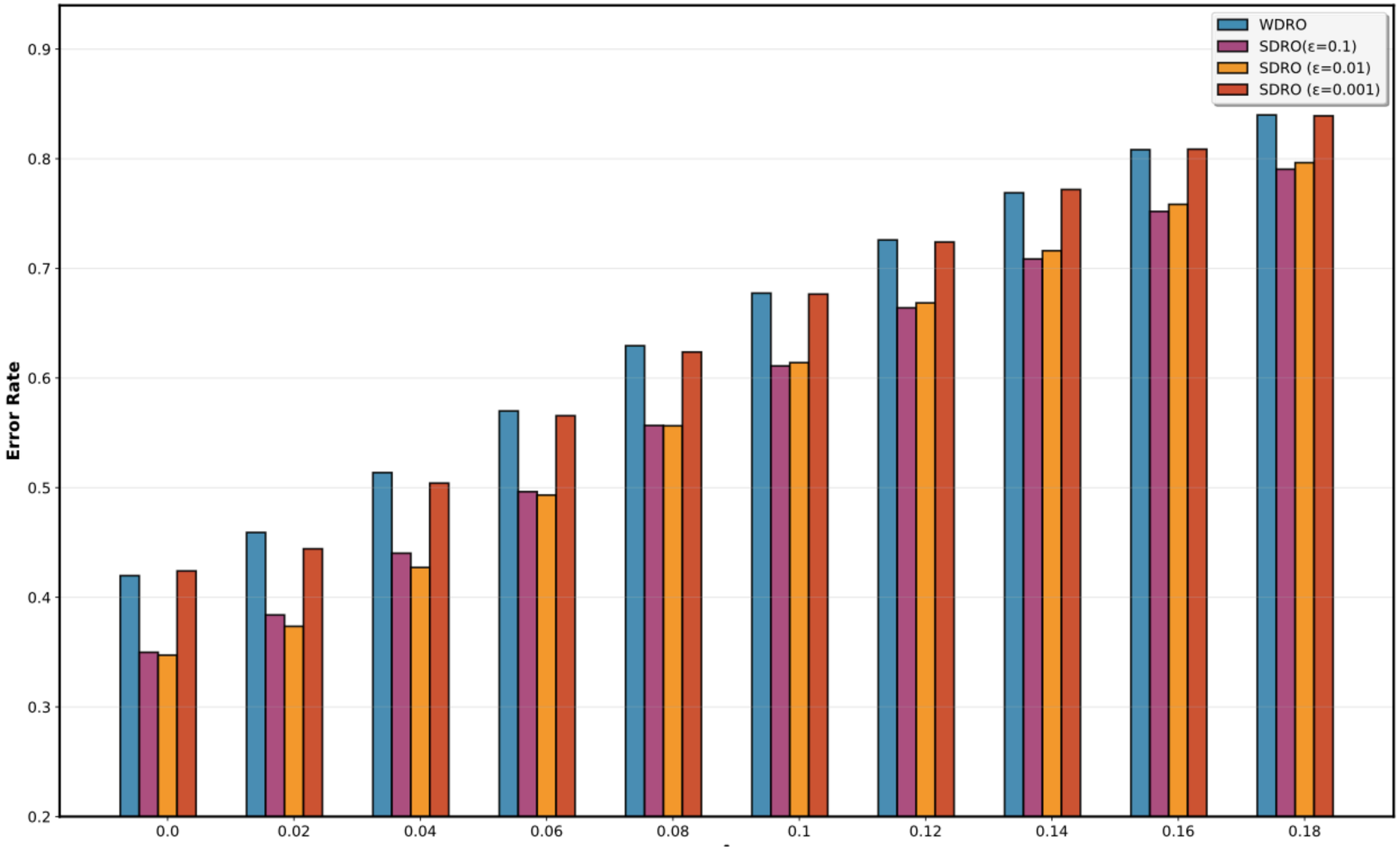
$$\mathrm{d}z_t = -\left(-\frac{\nabla f_{\theta_k}(z_t)}{\lambda} + (z_t - x^{(i)})\right)\mathrm{d}t + \sqrt{2\epsilon}\mathrm{d}\mathbf{W}_t, \quad \mathrm{Law}(z_t) = \rho_t, \rho_0 = \mu_k^{(i)}, \rho_\tau = \mu_{k+1}^{(i)}$$

Numerical Study



- MNIST Dataset
- *x*-axis: level of perturbation
- *y*-axis: error rate on testing dataset

Numerical Study



- CIFAR-10 Dataset
- *x*-axis: level of perturbation
- *y*-axis: error rate on testing dataset

Summary

- Solving Sinkhorn DRO from the **primal perspective**
- Reformulate Sinkhorn DRO as **infinite-dimensional bilevel program** with **multiple lower-level problems**.
- Double-loop algorithm is a "**noisy**" counterpart of the Wasserstein DRO algorithm by **[Sinha et. al 2018]**
- Single-loop algorithm requires **ODE analysis** in addition to finite-dimensional optimization tricks.