Distributionally Robust Optimization via Iterative Algorithms in Continuous Probability Spaces

Linglingzhi Zhu
Georgia Institute of Technology
llzzhu@gatech.edu

Yunqin Zhu
Georgia Institute of Technology
yzhu812@gatech.edu

Yao Xie
Georgia Institute of Technology
yao.xie@isye.gatech.edu

Abstract

We consider a minimax problem motivated by distributionally robust optimization (DRO) when the worst-case distribution is continuous, leading to significant computational challenges due to the infinite-dimensional nature of the optimization problem. Leveraging Brenier's theorem, we represent the worst-case distribution as a transport map of a continuous reference measure and reformulate the regularized discrepancy-based DRO as a minimax problem in Wasserstein space. We further propose an algorithmic framework with global convergence guarantees and complexity bounds for obtaining approximate stationary points. Under this continuous formulation, the proposed algorithms overcome the scalability, generalization, and worst-case inference limitations of discrete DRO approaches. Numerical results with neural network-based transport maps demonstrate that the proposed method enables both stable training of robust classifiers and effective worst-case inference for classification tasks.

1 Introduction

Distributionally robust optimization (DRO) addresses decision-making problems under uncertainty of the data distribution for stochastic optimization models. Since parameterized uncertainty set construction, e.g., moments [2, 6] and deviations [5] ambiguity sets may encompass an overly broad range of distributions significantly different from the reference, recent research has increasingly focused on non-parametric discrepancy-based ambiguity sets [1, 12, 16, 11, 3, 7], which have a tunable radius that quantifies the uncertainty level. In this paper, we focus on this category of discrepancy-based DRO:

$$\inf_{f \in \mathcal{F}} \sup_{\mathbb{Q} \in \mathcal{B}_\delta(\mathbb{P})} \mathbb{E}_{\xi \sim \mathbb{Q}}[\ell(f, \xi)]. \tag{1}$$

Here $\mathbb{P}$ is d-dimensional reference distribution with a finite second moment and $\mathcal{B}_\delta(\mathbb{P}) := \{\mathbb{Q} \in \mathcal{P}_2 : \mathcal{D}(\mathbb{Q}, \mathbb{P}) \le \delta\}$ for some $\delta > 0$ is the uncertainty set defined by the discrepancy function $\mathcal{D}$. Meanwhile, $\mathcal{F}$ is a measurable decision function class and $\ell : \mathcal{F} \times \mathbb{R}^d \to \mathbb{R}$ is the loss function.

Although the theory of DRO is well established and efficient solvers have recently emerged [14, 13, 9, 10], comparatively little attention has been paid to explicitly generating the worst-case distribution (i.e., least favorable distribution (LFD)). Existing approaches often rely on dual reformulations under discrete reference distributions [12, 11, 3, 7] or perturbation methods based on stochastic gradients [15, 4]. However, the above mentioned approaches faces several challenges when directly applied

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ML × OR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

to generate the LFD. First, there is a significant computational challenge. The dual formulation of DRO is not scalable to large datasets; for example, constructing the LFD under the Wasserstein case will require solving a linear program with the number of decision variables to be $\mathcal{O}(N^2)$, where N is the total number of training data points and the complexity of solving a linear program is typically quadratic on the number of the decision variable. Such computational complexity for problems with thousands of training data points can be prohibitive, which leads to the current DRO formulation usually can only be used to find discrete LFDs for small sample settings. Second, the discrete LFD will limit the generalization capability of the resulting algorithm. Since the LFD is discrete and supported solely on the training dataset, the corresponding optimal detector is confined to training points. This also creates a challenge when generating samples from the LFD, as the existing point-wise perturbations method requires retraining for each unseen data point to obtain the corresponding LFD sample.

In this paper, we are motivated to explore the DRO problem (1) as a minimax optimization in the continuous probability space for the LFD generation. Assuming the solution of the inner maximization problem is attainable, we focus on the following regularized formulation of (1):

$$\min_{\phi\in\Phi}\max_{\mathbb{Q}\in\mathcal{P}_2} \mathbb{E}_{\xi\sim\mathbb{Q}}[\ell(f_\phi,\xi)] - \lambda\cdot\mathcal{D}(\mathbb{Q},\mathbb{P}). \tag{P}$$

Here, we incorporate a parameterized decision function f_ϕ with $\phi\in\Phi\subseteq\mathbb{R}^m$ and a predetermined Lagrangian multiplier $\lambda\geq 0$. Different from traditional discrete DRO approaches and sampling-based methods in Wasserstein space, we leverage the existence of an optimal transport (OT) map guaranteed by Brenier's theorem thanks to our continuous setting. This allows us to derive an equivalent optimization problem over the transport map. Once the minimax optimal transport map for problem (P) is learned, we can simultaneously derive a robust classifier and efficiently generate worst-case samples by passing inputs through the learned map. This reformulation enables a practical implementation via neural network parameterizations [17], which rely on the change-of-variables formula and explicitly model data distributions through transport maps.

Theoretically, we study iterative algorithms for solving the regularized DRO problem (P) in infinite-dimensional Wasserstein spaces. Our main framework (Algorithm 1) updates both the decision variable and the LFD iteratively, and we establish global convergence to stationary points under mild assumptions. To enhance training stability, we further propose a proximal-regularized alternating scheme (Algorithm 2), inspired by the Jordan–Kinderlehrer–Otto (JKO) framework. Our analysis provides its convergence guarantees in Wasserstein space under weaker geometric conditions and characterizes the total number of gradient and JKO iterations required to achieve approximate stationarity.

Numerical results on a 2D binary classification example show that our alternating scheme parameterized with neural transport maps leads to more stable convergence and better exploration of complex LFDs than its non-alternating counterpart and particle-based perturbations.

2 Minimax Algorithmic Framework for Distributionally Robust Optimization

We present the main algorithmic framework for solving the DRO problem (P) in the continuous probability space. The detailed iterative procedure is given in Algorithm 1.

Algorithm 1 Minimax Algorithmic Framework

```
Input: Initialization $\phi_0$, step size $\eta > 0$, regularization parameter $\lambda > 0$, data $\xi \sim \mathbb{P}$
for $k = 0$ to $K - 1$ do
    $\mathbb{Q}_k := \epsilon\text{-}\operatorname{argmax}_{\mathbb{Q}\in\mathcal{P}_2} \{\mathbb{E}_{\xi\sim\mathbb{Q}}[\ell(f_{\phi_k},\xi)] - \lambda\cdot\mathcal{D}(\mathbb{Q},\mathbb{P})\}$
    $\phi_{k+1} := \operatorname{proj}_\Phi(\phi_k - \eta_k\cdot\zeta(\phi_k,\mathbb{Q}_k))$ with $\zeta(\phi_k,\mathbb{Q}_k) \in \partial_\phi \mathbb{E}_{\xi\sim\mathbb{Q}_k}[\ell(f_{(\cdot)},\xi)](\phi_k)$
end for
```

From a computational perspective, the key challenge for practical implementation is solving the inexact maximization step in updating the probability measure $\mathbb{Q}$. In fact, with fixed decision function f_ϕ, we can consider the following equivalent transport map maximization problem:

$$\max_{T\in L^2(\mathbb{P})} \mathbb{E}_{\xi\sim\mathbb{P}}[\ell(f_\phi,T(\xi))] - \lambda\cdot\mathcal{D}(T_\#\mathbb{P},\mathbb{P}). \tag{2}$$

Hence, we can solve the inexact maximization step by parameterizing the transport map. To derive the convergence result, we introduce the notation:

- **Minimax loss function:** $\mathcal{H}(\phi, \mathbb{Q}) := \mathbb{E}_{\xi \sim \mathbb{Q}}[\ell(f_\phi, \xi)] - \lambda \cdot \mathcal{D}(\mathbb{Q}, \mathbb{P})$;
- **Worst-case value function:** $\mathcal{V}(\phi) := \max_{\mathbb{Q} \in \mathcal{P}_2} \mathcal{H}(\phi, \mathbb{Q})$.

The following mild assumption is required to establish the convergence guarantees of Algorithm 1.

Assumption 2.1. *$\ell(f_{(\cdot)}, \xi)$ is ρ-weakly convex and L-Lipschitz on the convex set Φ for any $\xi \in \mathbb{R}^d$.*

Theorem 2.2 (Convergence theorem)**.** *Suppose Assumption 2.1 holds. Then for the sequence generated by Algorithm 1 with $\eta = 1/\sqrt{K}$, there exists a $k \in \{0, 1, \ldots, K-1\}$ such that*

$$\|\nabla \mathcal{V}_{1/2\rho}(\phi_k)\| \leq \mathcal{O}(K^{-1/4}) + \mathcal{O}(\sqrt{\epsilon}),$$

where $\mathcal{V}_{1/2\rho}$ is the Moreau envelope of $\mathcal{V}$.

3 Alternating Scheme and Oracle Complexity

To improve stability and efficiency in solving (2), we introduce an alternating scheme inspired by the JKO framework [8], which yields a stable discrete-time approximation of Wasserstein gradient flows.

Algorithm 2 Alternating Algorithm

```
Input: Initialization $\phi_0$ and $\mathbb{Q}_0$, step size $\eta > 0$, regularization parameter $\lambda > 0$, data $\xi \sim \mathbb{P}$
for $k = 0, 1, \ldots, K-1$ do
    $\mathbb{Q}_{k+1} := \mathrm{argmax}_{\mathbb{Q} \in \mathcal{P}_2} \{\mathcal{H}(\phi_k, \mathbb{Q}) - \frac{1}{2\gamma} \cdot \mathcal{W}_2^2(\mathbb{Q}, \mathbb{Q}_k)\}$
    $\phi_{k+1} := \phi_k - \eta \cdot \nabla_\phi \mathcal{H}(\phi_k, \mathbb{Q}_k)$
end for
```

An approximate JKO solution can be obtained through a parameterized transport map maximization via Brenier's theorem, initialized with a map T_0 that pushes a continuous anchor distribution ν to $\mathbb{P}$:

$$T_{k+1} := \underset{T \in L^2(\nu)}{\mathrm{argmax}} \left\{ \mathbb{E}_{\xi \sim \nu} \left[\ell(f_{\phi_k}, T(\xi)) - \frac{1}{2\gamma} \|T(\xi) - T_k(\xi)\|^2 \right] - \lambda \cdot \mathcal{D}(T_\# \nu, \mathbb{P}) \right\}. \tag{3}$$

Proposition 3.1 (Equivalent JKO step via transport maps)**.** *Suppose that $T_k \in L^2(\nu)$ is invertible. If T_{k+1} is the solution of* (3) *and $\mathbb{Q}_k = (T_k)_\# \nu$, then the pushforward $(T_{k+1})_\# \nu$ satisfies*

$$(T_{k+1})_\# \nu = \underset{\mathbb{Q} \in \mathcal{P}_2}{\mathrm{argmax}} \left\{ \mathcal{H}(\phi, \mathbb{Q}) - \frac{1}{2\gamma} \cdot \mathcal{W}_2^2(\mathbb{Q}, \mathbb{Q}_k) \right\}.$$

To proceed with the complexity analysis, we impose the following assumptions.

Assumption 3.2. *The following assumptions on the problem* (P) *hold:*

(i) *$\ell(f_{(\cdot)}, \cdot)$ is continuously differentiable with L-Lipschitz gradient on the convex set $\mathbb{R}^m \times \mathbb{R}^d$;*

(ii) *$\mathcal{D}(\cdot, \mathbb{P})$ is 1-strongly convex along generalized geodesics centered at $\nu \in \mathcal{P}_2^r$ with $\lambda > \rho$, and the support of $\mathrm{argmax}_{\mathbb{Q} \in \mathcal{P}_2} \mathcal{H}(\phi, \mathbb{Q})$ belongs to a compact set $\mathcal{X}$ for any $\phi \in \mathbb{R}^m$.*

The strong convexity along fixed generalized geodesics is satisfied by common discrepancy measures such as the squared Wasserstein-2 distance and the KL divergence with a log-concave reference density induced by a strongly convex potential.

Assumption 3.3 (Approximate k-th JKO step solution)**.** *Let $\epsilon' \geq 0$. For k-th step, there exists*

$$\xi_{k+1} \in -\partial_{\mathcal{W}_2} \mathcal{H}(\phi, \mathbb{Q}_{k+1}) + (I - T_{k+1}^k)/\gamma$$

such that $\|\xi_{k+1}\|_{\mathbb{Q}_{k+1}} \leq \epsilon'$, where T_{k+1}^k is the OT map from $\mathbb{Q}_{k+1}$ to $\mathbb{Q}_k$.

Theorem 3.4 (Oracle complexity for alternating algorithm)**.** *Suppose that Assumptions 3.2 and 3.3 hold. Then for the sequence generated by Algorithm 2 with step size $\eta = \mathcal{O}(1/L)$ and $\gamma = \Omega(1/(\lambda - \rho))$, there exists a $k \in \{0, 1, \ldots, K-1\}$ such that*

$$\|\nabla \mathcal{V}(\phi_k)\| \leq \mathcal{O}(K^{-1/2}) + \mathcal{O}(\epsilon').$$

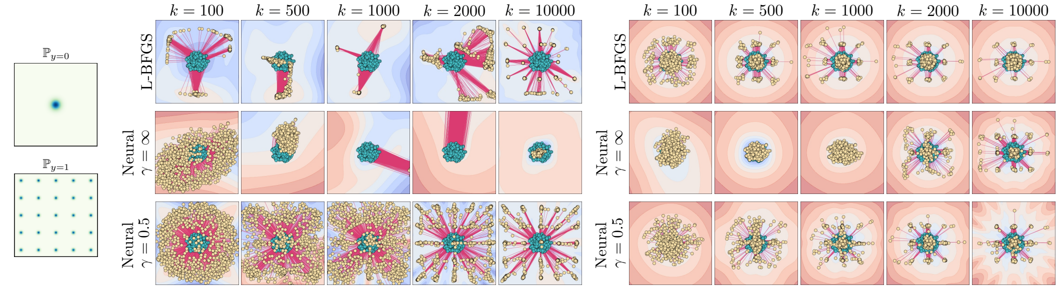
Moreover, Algorithm 2 will return a solution ϕ^* such that $\|\nabla \mathcal{V}(\phi^*)\| \leq \varepsilon$ within $\mathcal{O}(\varepsilon^{-2})$ gradient oracle calls and $\mathcal{O}(\varepsilon^{-2})$ inexact JKO steps in Assumption 3.3 with accuracy $\epsilon' = \mathcal{O}(\varepsilon)$.

4 Numerical Experiments

Problem Setup Our algorithms offer a high-level framework for DRO with convergence guarantees. A central difficulty, however, is that JKO steps over continuous distributions lack a general analytic solution. Fortunately, when $\mathcal{D} = \mathcal{W}_2^2$, the transport map problem (3) can be approximately solved by

$$\max_{T \in L^2(\nu)} \mathbb{E}_{\xi \sim \nu}\left[\ell(f_\phi, T(\xi)) - \lambda \|T(\xi) - T_0(\xi)\|^2 - \frac{1}{2\gamma}\|T(\xi) - T_k(\xi)\|^2\right]. \tag{4}$$

We illustrate this with a binary classification task, where f_ϕ is trained with the cross-entropy loss $\ell(f_\phi, \xi) = -\log \sigma(f_\phi(\xi)) - \mathbb{E}_{\xi' \sim \mathbb{P}'}\left[\log(1 - \sigma(f_\phi(\xi')))\right]$. We let $\nu = \mathbb{P}$ be standard Gaussian, and define the opposite class $\mathbb{P}'$ as a 25-component Gaussian mixture. The goal is to show convergence of both the transport map T and classifier f_ϕ under minimax optimization.



(a) Data Dist. (b) $\lambda = 0$ (c) $\lambda = 0.1$

Figure 1: **Minimax optimization on 2D synthetic data.** We compare $\lambda = 0$ vs. $\lambda = 0.1$ under three methods. (a) shows class distributions. In (b)–(c), each column is a training step k: original samples in blue, worst-case samples in yellow, transport displacements in red. Background contours show classifier logits f_{ϕ_k} (blue toward class 0, red toward class 1).

Compared Methods We evaluate two implementations of the worst-case transport map (4). (i) **L-BFGS** applied directly to particles $\xi \sim \mathbb{P}$, considering the non-alternating $\gamma = \infty$ case, which reduces to the method of [15]. (ii) A **neural network** parameterization of the form $T(\xi) = \xi + F_\theta(\xi)$, with either $\gamma = \infty$ (non-alternating [17]) or $\gamma = 0.5$ (alternating). Here, F_θ is a ResNet with 3 hidden blocks of width 768. The parameters θ_k are updated to θ_{k+1} using 5 mini-batches per iteration. We adopt a relatively large network to ensure sufficient expressivity and fast convergence. The classifier f_ϕ is an MLP with 3 hidden layers of width 768. Both networks use smooth SiLU activation and are trained using Adam with learning rate 10^{-4} and batch size $m = 1000$.

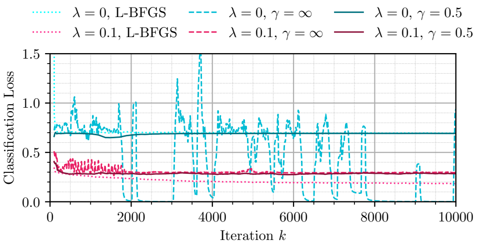


Figure 2: Classification loss on 2D synthetic data (moving average over 100 mini-batches).

Results Figures 1 and 2 show that the alternating scheme with $\gamma = 0.5$ substantially stabilizes the convergence of both the neural transport map and the classifier, compared to using an infinite step size. The parameterized LFD also explores the multi-modal distribution $\mathbb{P}'$ more quickly and effectively than the discrete particle-based solver under $\lambda = 0$, where the latter tends to become trapped around a few modes. These advantages arise from two main factors: (i) the alternating scheme constrains the update step size of the continuous LFD, reducing noise in the early stages of minimax optimization and encouraging stable, progressive exploration of the input space; and (ii) neural transport maps can be warm-started across iterations, so that knowledge from the previously learned $\mathbb{Q}_k$ naturally transfers forward, lowering the cost of repeated inner maximization while implicitly regularizing the learning process. Overall, these results highlight the strength of our continuous distribution formulation, which motivates the development of iterative alternating scheme and naturally leads to expressive neural network parameterizations.

References

[1] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[2] Dimitris Bertsimas and Jay Sethuraman. Moment problems and semidefinite optimization. In *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, pages 469–509. Springer, 2000.

[3] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[4] Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research*, 47(2):1500–1529, 2022.

[5] Xin Chen, Melvyn Sim, and Peng Sun. A robust optimization perspective on stochastic programming. *Operations Research*, 55(6):1058–1071, 2007.

[6] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[7] Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

[8] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[9] Daniel Kuhn, Soroosh Shafieezadeh-Abadeh, and Wolfram Wiesemann. Distributionally robust optimization. *arXiv preprint arXiv:2411.02549*, 2024.

[10] Jiashuo Liu, Tianyu Wang, Henry Lam, Hongseok Namkoong, and Jose Blanchet. Dro: A python library for distributionally robust optimization in machine learning. *arXiv preprint arXiv:2505.23565*, 2025.

[11] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

[12] Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in Neural Information Processing Systems*, 29, 2016.

[13] Hamed Rahimian and Sanjay Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, 2022.

[14] Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

[15] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[16] Zizhuo Wang, Peter W Glynn, and Yinyu Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13:241–261, 2016.

[17] Chen Xu, Jonghyeok Lee, Xiuyuan Cheng, and Yao Xie. Flow-based distributionally robust optimization. *IEEE Journal on Selected Areas in Information Theory*, 2024.