OPT2024: 16th Annual Workshop on Optimization for Machine Learning

A Stochastic Algorithm for Sinkhorn Distance-Regularized Distributionally Robust Optimization

Yufeng Yang YUFENG.YANG@TAMU.EDU
Yi Zhou YIZHOU_TAMU@EDU
Texas A& M University, College Station

Zhaosong Lu ZHAOSONG@UMN.EDU
University of Minnesota, Twin Cities

Abstract

Distributionally Robust Optimization (DRO) is a powerful modeling technique to tackle the challenge caused by data distribution shifts. This paper focuses on Sinkhorn distance regularized DRO. We generalize Sinkhorn distance allowing broader function choices to model ambiguity set and derive the lagrangian dual taking the form of nested stochastic programming. We also design the algorithm based on stochastic gradient descent with easy-to-implement constant learning rate. Unlike previous work doing algorithm analysis for convex and bounded loss function, our algorithm provides convergence guarantee for non-convex and possible unbounded loss function under proper choice of sampling batch-size. The resultant sample complexity for finding ϵ-stationary point reveals independent relationship with data size and parameter dimension, and thus our modeling and algorithms are suitable for large-scale applications.

Keywords: stochastic algorithm, distributionally robust optimization, Sinkhorn distance

1. Introduction

In classic machine learning, the primary goal is to achieve good predictive performance on the test set after training the model on a designated training set. The training problem is typically formulated by the expected risk minimization problem. In particular, stochastic gradient descent (SGD) [17] and its variants [13, 21, 22] have been introduced to solve this type of problem. However, expected risk minimization assumes that the training set and the test set follow the same underlying distribution, which is often unrealistic and may result in bad test performance when data distribution shift exists.

The shifts on data distribution is prevalent in real-world scenarios. It can be caused by many factors such as sampling bias, presence of anomalies, data merging and change of measurements, etc. To tackle this challenge, Distributionally Robust Optimization (DRO) [18] was proposed, which formulates the objective function as a min-max problem. DRO aims to learn a robust model by minimizing the expected risk over the worst-case data distribution within a predefined ambiguity set. This formulation offers a principled framework to learn the optimal resilient solution in the face of distribution uncertainty.

One key factor in DRO problems is the selection of an appropriate divergence measure for modeling the ambiguity set. Specifically, the divergence measure should not only be computationally tractable but also yield a solution that avoids excessive conservatism. In the existing literature, various divergence-based ambiguity sets have been studied. In [14, 15, 25, 33], the authors focus on

reformulating the expressions of loss functions under the worst-case distributions into a tractable form and exploring possible algorithms to tackle DRO problems under Wasserstein metric based ambiguity sets. For more information, we refer the readers to [23] for a comprehensive survey on Wasserstein DRO. In [12, 19, 24], the authors analyze alternative expressions of loss functions under the worst-case distribution and develop algorithms to solve DRO problems under Kullback–Leibler (KL) divergence-based and f-divergence-based ambiguity sets. However, the aforementioned divergence measures have certain limitations. For example, it is known that DRO with Wasserstein distance requires high computational complexity [5, 28]. Both KL and f-divergence are not symmetric when assessing distributions. Furthermore, these two divergence measures require that the distributions share the same probability support, a strong condition that may fail to capture extreme distributions at certain points.

The Sinkhorn distance, first introduced in [10], was designed to address the aforementioned limitations. Sinkhorn distance is symmetric and allows distributions from the same sample space to have different probability support. Furthermore, Sinkhorn distance is a convex function with respect to distributions, ensuring computation tractability and efficiency for large-scale problems. In [37], Sinkhorn DRO was initially investigated. Specifically, they derived the dual formulation of a constrained Sinkhorn DRO problem, which can be effectively solved by the mirror descent algorithm. However, the convergence analysis conducted in their work assumed that the loss function is convex and bounded, which may not hold in practical modern machine learning applications.

Motivated by these limitations, in this study, we consider the regularized Sinkhorn DRO problem (see (1)) with nonconvex and possibly unbounded loss. Our contributions are summarized as follows.

- We introduce a generalized Sinkhorn distance based on the class of f-divergence measures. This generalized notion not only retains the advantages of the original Sinkhorn distance but also allows to use a broader range of divergences to model the ambiguity set.
- We derive an equivalent dual formulation of the regularized Sinkhorn DRO problem with strong duality guarantee. The dual problem takes the form of nested stochastic programming.
- To solve the nested stochastic problem with nonconvex and unbounded loss, we design a Nested-SGD algorithm with guaranteed convergence. Our Nested-SGD is specifically tailored for solving large-scale regularized Sinkhorn DRO problems.

2. Related Work

DRO. The DRO framework shares strong connections with contrastive learning [38], multiple instance learning [32], and anomaly detection [6]. The key challenge during modelings is the choice of ambiguity set. The first stream focuses on using information divergence to construct ambiguity set. Commonly employed divergence measures include the Wasserstein metric [14, 15, 25, 33]; KL divergence [19, 35]; f-divergence [12, 20, 24, 26] and Sinkhorn distance [37]. Another stream for constructing ambiguity set is using special statistics, such as geometry shape constraints [7] and statistical moments [9, 11, 18] etc.

Sinkhorn Distance. Sinkhorn distance has successful applications in areas like generative models [16, 27], matrix factorization [30], image segmentation[31] etc. In [10], Sinkhorn matrix scaling algorithms was proposed to compute optimal transport map under Sinkhorn distance objective.

Later, in [1, 4], greedy and stochastic variants of Sinkhorn scaling algorithms were proposed to clarify relationship between algorithm convergence and input dimensions. Some works also study Sinkhorn distance computation over data samples with special structures. In [2, 36], their algorithms specially applies to data samples over compact Riemannian manifolds and Euclidean balls respectively.

Algorithms for Solving DRO. For Wasserstein metric ambiguity set, several works reformulate primal problem into tractable forms such as convex programming[33], semi-definite programming [25] and mixed integer programming [32]. Subsequent works [32, 33] directly use software toolbox to solve their problems. In [25], they use cutting-surface method to solve semi-definite programming for general nonlinear loss and branch-and-bound algorithms for bilinear loss. Another common technique to transform DRO is using lagrangian duality [15, 24]. Through this way, computation of shifted distribution can be avoided. In [29], projected SGD and acceleration is used to solve the dual form of KL divergence constrained DRO. In [20], normalized SGD with momentum is used to solve the dual of f-divergence regularized DRO. In [40], stochastic Frank-Wolfe is used to solve approximation for the dual of general Cressie-Read family divergence constrained DRO. In [37], stochastic mirror descent is used to solve the dual form of Sinkhorn distance constrained DRO.

3. DRO with Generalized Sinkhorn Distance

In DRO, the goal is to learn a model that achieves good and robust performance under uncertainty of the underlying data distribution. Specifically, consider a machine learning problem with the loss function denoted by $\ell(x;\xi)$, where $x \in \mathbf{R}^d$ denotes the collection of model parameters and ξ corresponds to a data sample that follows an underlying nominal distribution $\mathbb{Q}$. Then, with a regularization parameter $\lambda > 0$, we study the following regularized DRO problem

$$\min_{x\in\mathbf{R}^d} \sup_{\mathbb{Q}} \Big\{ \mathbb{E}_{\xi\sim\mathbb{Q}}\big[\ell(x;\xi)\big] - \lambda W_\epsilon(\mathbb{P},\mathbb{Q}) \Big\}, \tag{1}$$

where $W_\epsilon(\mathbb{P},\mathbb{Q})$ denotes a certain function (with parameter $\epsilon > 0$) that measures the distance between the distributions $\mathbb{P}$ and $\mathbb{Q}$. In particular, the operation $\min_x \sup_{\mathbb{Q}}$ aims to optimize the model under the worst-case data distribution $\mathbb{Q}$ to enhance model robustness against distribution shift.

In this work, we consider the following generalized Sinkhorn distance to quantify the distribution shift. Throughout the paper, we consider a sample space Ω and σ-algebra $\mathcal{F}$. We assume that the distribution $\mathbb{Q}$ over a measurable subset of $\mathcal{F}$ is absolutely continuous with regard to a reference measure ν, i.e., $\mathbb{Q} \ll \nu$.

Definition 1 (Generalized Sinkhorn Distance) *Denote $\Gamma(\mathbb{P},\mathbb{Q})$ as the set of joint distributions that have marginal distributions $\mathbb{P},\mathbb{Q}$. For a fixed regularization parameter $\epsilon > 0$ and a cost metric $c : \Omega \times \Omega \to \mathbf{R}$, the generalized Sinkhorn distance is defined as*

$$W_\epsilon(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \Big\{ \mathbb{E}_{(\zeta,\xi)\sim\gamma}\big[c(\zeta,\xi)\big] + \epsilon D_f(\gamma \mid \mathbb{P}\otimes\nu) \Big\},$$

where D_f corresponds to the f-divergence[1]*, that is,*

$$D_f(\gamma \mid \mathbb{P} \otimes \nu) = \int f\big(\frac{\mathrm{d}\gamma(\zeta,\xi)}{\mathrm{d}\mathbb{P}(\zeta)\mathrm{d}\nu(\xi)}\big)\mathrm{d}\nu(\xi)\mathrm{d}\mathbb{P}(\zeta).$$

And $\frac{\mathrm{d}\gamma(\zeta,\xi)}{\mathrm{d}\mathbb{P}(\zeta)\mathrm{d}\nu(\xi)}$ represents density ratio of γ with respect to $\mathbb{P}\otimes\nu$ evaluated at (ζ,ξ).

Remark 2 *Typical choices of the reference measure ν include the Lebesgue measure or the Gaussian measure. We note that the divergence term $D_f(\gamma \mid \mathbb{P}\otimes\nu)$ is equivalent to $D_f(\gamma \mid \mathbb{P}\otimes\mathbb{Q})$ up to a constant, and we consider the former term for simplicity.*

The proposed generalized Sinkhorn distance allows the data distributions $\mathbb{P}$ and $\mathbb{Q}$ to have different probability support. This provides more flexibility to model distribution uncertainty compared to other divergence-based measures [20, 24]. Moreover, our generalized Sinkhorn distance is based on the f-divergence, which generalizes the KL-divergence adopted in the definition of the standard Sinkhorn distance [37].

The primal regularized DRO problem in (1) is hard to solve, since it is challenging to obtain an analytical form of the worst-case distribution $\mathbb{Q}$. In the next section, we study its dual formulation and develop efficient estimators of its stochastic gradients.

4. Dual Formulation and Gradient Estimation

The generalized Sinkhorn distance involves special structures that can transform the primal regularized DRO problem in (1) into a simpler dual form. To elaborate, we first decompose the joint distribution as $\gamma(\zeta,\xi) = \gamma_\zeta(\xi)\mathbb{P}(\zeta)$, where γ_ζ corresponds to the conditional distribution over ξ. Moreover, by the interchangeability [34] between $\mathbb{E}_{\zeta\sim\mathbb{P}}$ and $\sup_{\gamma_\zeta}$, the primal problem in (1) can be rewritten as

$$\min_{x\in\mathbf{R}^d} \mathbb{E}_{\zeta\sim\mathbb{P}}\Big[\sup_{\gamma_\zeta}\Big(\mathbb{E}_{\xi\sim\gamma_\zeta}\big[\ell(x;\xi) - \lambda c(\zeta,\xi)\big] - \lambda\epsilon D_f(\gamma_\zeta \mid \nu)\Big)\Big]. \tag{2}$$

Following inverse c.d.f sampling argument in [24], the inner supremum term $\sup_{\gamma_\zeta}(\cdot)$ has the following equivalent dual formulation

$$\min_{\eta\in\mathbf{R}}\Big\{L_\zeta(x,\eta) := \lambda\epsilon\mathbb{E}_{\xi\sim\nu}\Big[f^*\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta}{\lambda\epsilon}\Big)\Big] + \eta\Big\},$$

where η is the dual variable and f^* denotes the conjugate function of f [2]. For simplicity of presentation, we denote $\eta_x^*(\zeta) \in \arg\min_\eta L_\zeta(x,\eta)$ and define

$$\Psi_\zeta(x) := L_\zeta(x,\eta_x^*(\zeta)), \quad L_{\zeta,\xi}(x,\eta) := \lambda\epsilon f^*\big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta}{\lambda\epsilon}\big)+\eta.$$

Then, the problem (1) can be written as the following nested stochastic problem

$$\min_{x\in\mathbf{R}^d} \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big], \text{ where } \Psi_\zeta(x) = \min_\eta L_\zeta(x,\eta). \tag{3}$$

1. For f-divergence, the function $f:[0,+\infty)\to[-\infty,+\infty]$ is convex and satisfies $f(1)=0$ and $f(0)=\lim_{t\to 0+} f(t)$.
2. The conjugate function is defined as $f^*(v) = \sup_{t\in\mathrm{dom}(f)}\{vt - f(t)\}$

The hardness of the above problem is due to the nested structure, where both the inner and outer problems are stochastic optimization problems. Moreover, it is different from a standard bi-level optimization problem, since the inner optimizer $\eta_x^*(\zeta)$ varies with regard to the data ζ.

To make sure the problem is well-defined, we adopt the following assumptions throughout the paper.

Assumption 4.1 *The functions in problem (3) satisfy:*

- *For every ξ, $\ell(\cdot;\xi)$ is G-Lipschitz continuous, and $\ell(\cdot;\xi)$ is differentiable and L-smooth.*
- *Function $f^*(\cdot)$ is differentiable and M-smooth.*
- *The function $\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big]$ is bounded below.*

Note that the loss function $\ell(x;\xi)$ is not necessary to be convex.

In addition, we adopt the following assumptions that both the loss function and the cost metric have bounded variance.

Assumption 4.2 *There exists $\sigma, \delta > 0$ such that:*

- *For every x, the variance of $\ell(x;\cdot)$ is bounded by σ^2.*
- *For every ζ, the variance of $c(\zeta,\cdot)$ is bounded by δ^2. And for every ξ, the variance of $c(\cdot,\xi)$ is bounded by δ^2.*

To solve the nested stochastic optimization problem (3), intuitively, one can apply the standard SGD algorithm to solve the outer-level problem $\min_{x\in\mathbf{R}^d}\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big]$. The following lemma, proved in [20], provides an analytical formula for computing the exact stochastic gradient.

Lemma 3 *Let Assumption 4.1 hold and consider any fixed x and ζ. Then, the function $\Psi_\zeta(x)$ is differentiable and satisfies $\nabla\Psi_\zeta(x) = \nabla_1 L_\zeta(x,\eta_x^*(\zeta))$, where $\eta_x^*(\zeta) \in \arg\min_\eta L_\zeta(x,\eta)$.*

However, as can be seen from the above lemma, calculating the stochastic gradient requires access to the exact minimizer $\eta_x^*(\zeta)$ of the inner stochastic problem, which is often hard to obtain in practice. To address this issue, in the following theorem, we develop an approximation of the stochastic gradient based on an inexact solution of the inner stochastic problem.

Theorem 4 *Suppose we obtain x and $\eta_x(\zeta)$ such that the gradient taken over second argument satisfy*

$$\big|\nabla_2 L_\zeta(x,\eta_x(\zeta))\big| \le \epsilon_1. \tag{4}$$

Then, for any ζ, the gradient taken over first argument satisfy

$$\|\nabla\Psi_\zeta(x) - \nabla_1 L_\zeta(x,\eta_x(\zeta))\| \le G\epsilon_1. \tag{5}$$

The above theorem indicates that, an accurate estimate of the stochastic gradient can be constructed based on an inexact solution $\eta_x(\zeta)$ of the inner problem. Note that the stationary condition in (4) is an optimality condition due to convexity of the inner problem.

5. Nested-SGD and Convergence Analysis

Theorem 4 naturally inspires us to design a SGD-like algorithm with nested loops, i.e., the inner SGD loop solves the inner stochastic problem for ever sampled ζ to obtain the inexact solution $\eta_x(\zeta)$ (see Algorithm 2), and the outer SGD loop solves the outer stochastic problem based on the stochastic gradient estimator proposed in the above theorem (see Algorithm 1).

Specifically, in the Nested-SGD Algorithm 1, we design a mini-batch SGD algorithm to update x. At each iteration t, the algorithm samples one ζ and a batch of $\{\xi\}_{B_1}$ with batch size B_1 to construct a stochastic gradient estimator taking the form

$$\hat{g}_t^B = \frac{1}{B_1}\sum_{i=1}^{B_1} f^{*'}\Big(\frac{\ell(x_t;\xi_i) - c(\zeta,\xi_i) - \eta_{x_t}(\zeta)}{\lambda\epsilon}\Big)\nabla\ell(x_t;\xi_i). \tag{6}$$

In Algorithm 2, we use mini-batch SGD to find an approximation of $\eta_x^*(\zeta)$. At each iteration d, the algorithm samples a batch of $\{\xi\}_{B_2}$ with batch size B_2 to construct a stochastic gradient estimator taking the form

$$v_d^B = 1 - \frac{1}{B_2}\sum_{i=1}^{B_2} f^{*'}\Big(\frac{\ell(x;\xi_i) - c(\zeta,\xi_i) - \eta_x^d(\zeta)}{\lambda\epsilon}\Big). \tag{7}$$

Since $L_\zeta(x,\eta)$ is a convex function with regard to η, when output $\eta_x^{\bar{d}}(\zeta)$ with the minimal gradient norm, it can be guaranteed that the obtained $\eta_x(\zeta)$ is close to $\eta_x^*(\zeta)$.

Algorithm 1: Nested-SGD for solving $\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big]$

Data: $T \in \mathbb{N}$, initialization x_0, η_0, learning rate γ_t
for $t = 0....T-1$ **do**
  Sample $\{\zeta\}$and$\{\xi\}_{B_1}$ with batch size B_1
  Construct estimator $\eta_{x_t}(\zeta)$ via Algorithm 2
  Compute gradient estimator $\hat{g}_t^B$ via equation (6)
  Update $x_{t+1} = x_t - \gamma_t\hat{g}_t^B$
end
Result: Output $x_{\bar{t}}$, where $\bar{t}$ is sampled from $\{0....T-1\}$ uniformly at random

Algorithm 2: Construct Estimator $\eta_x(\zeta)$

Data: $D \in \mathbb{N}$, learning rate α_d
for $d = 0....D-1$ **do**
  Utilize the ζ sampled in Algorithm 1
  Sample $\{\xi\}_{B_2}$ with batch size B_2
  Compute gradient estimator v_d^B via equation (7)
  Update $\eta_{x_t}^{d+1}(\zeta) = \eta_{x_t}^d(\zeta) - \alpha_d v_d^B$
end
Result: Output $\eta_{x_t}^{\bar{d}}(\zeta)$, where $\bar{d} \in \{0,...,D-1\}$ corresponds to the index with minimal gradient norm

To analyze the convergence of Nested-SGD, we first prove some smoothness conditions for the functions $\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big]$ and $L_\zeta(x,\eta)$. Note that when x and η are arbitrary chosen, the function $\mathbb{E}_{\zeta\sim\mathbb{P}}\big[L_\zeta(x,\eta)\big]$ satisfies generalized smoothness condition proposed in [8, 39]. However, if η is carefully chosen such that $\eta_x^*(\zeta) \in \arg\min_\eta L_\zeta(x,\eta)$, the generalized smoothness condition reduces to the smoothness condition as shown in the following Lemma 5.

Lemma 5 *The following smoothness conditions hold.*

- *For any x, x', it holds that*
$$\mathbb{E}_{\zeta\sim\mathbb{P}}\big\|\nabla\Psi_\zeta(x) - \nabla_1 L_\zeta(x', \eta_x^*(\zeta))\big\|^2 \le K^2\big\|x - x'\big\|^2, \tag{8}$$
where $K = G^2(\lambda\epsilon)^{-1}M + L$.
- *For any x and any η, η', it holds that*
$$\mathbb{E}_{\xi\sim\nu}\big\|\nabla_2 L_{\zeta,\xi}(x,\eta) - \nabla_2 L_{\zeta,\xi}(x,\eta')\big\|^2 \le K'^2\big\|\eta - \eta'\big\|^2, \tag{9}$$
where $K' = M(\lambda\epsilon)^{-1}$.

We also develop bounds for the second moment of the proposed gradient estimators.

Lemma 6 *For mini-batch gradient estimator $\hat{g}_t^B$ used in Algorithm 1, it satisfies*
$$\mathbb{E}_{\zeta\sim\mathbb{P},\xi_B\sim\nu}\big\|\hat{g}_t^B\big\|^2 \le R_{B_1} + \frac{8G^2\epsilon_1^2}{B_1} + \big\|\nabla_1\mathbb{E}_{\zeta\sim\mathbb{P}}\big[L_\zeta(x_t,\eta_{x_t}(\zeta))\big]\big\|^2, \tag{10}$$
where $R_{B_1} = \mathcal{O}(\frac{G^2+G^2M^2(\lambda\epsilon)^{-2}\sigma^2}{B_1} + G^2M^2\epsilon^{-2}\delta^{-2})$.
For mini-batch gradient estimator v_d^B used in Algorithm 2, it satisfies
$$\mathbb{E}_{\xi_B\sim\nu}\big\|v_d^B\big\|^2 \le \frac{R_2}{B_2} + \big\|\nabla_2 L_\zeta(x_t,\eta_{x_t}^d(\zeta))\big\|^2, \tag{11}$$
where $R_2 = 2M^2(\lambda\epsilon)^{-2}(\sigma^2 + \lambda^2\delta^2)$.

Based on the above lemma, we obtain the following convergence result of Nested-SGD for minimizing $\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x)]$.

Theorem 7 *Let Assumptions 4.1 and 4.2 hold. Denote $\Delta = \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_0)\big] - \inf_x \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big]$. Run Nested-SGD for T iterations with learning rate $\gamma_t = \min\big\{\frac{1}{3K}, \sqrt{\frac{2\Delta}{KR_{B_1}T}}\big\}$ and error threshold $\epsilon_1(t) = \Theta(G^{-1}T^{-\frac{1}{2}})$ for all t. Then, the convergence result is*
$$\mathbb{E}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2 \le \mathcal{O}\Big(\sqrt{\frac{\Delta KR_{B_1}}{T}}\Big) + \mathcal{O}\Big(\frac{\Delta K}{T}\Big) + \mathcal{O}\Big(\frac{B_1^{-1}\sqrt{\Delta K/R_{B_1}}}{T^{3/2}}\Big). \tag{12}$$
Moreover, to achieve $\mathbb{E}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\| \le \delta_1$, choose $B_1 = \Theta(1)$, then the sample complexity of Algorithm 1 is $\Omega(\Delta KR_{B_1}\delta_1^{-4})$.

For Algorithm 2, the convergence analysis and learning rate choice follow the standard SGD [3, 17]. The difference is we use mini-batch version to ensure convergence.

Theorem 8 *Let Assumptions 4.1 and 4.2 hold. Denote $\widehat{\Delta} = L_\zeta(x_t,\eta_{x_t}^0(\zeta)) - \Psi_\zeta(x_t)$. Run Algorithm 2 for D iterations with learning rate $\alpha_d = \min\Big\{\frac{1}{K'}, \sqrt{\frac{2\widehat{\Delta}}{K'(R_2/B_2)D}}\Big\}$ for all d. Then, the convergence result is*
$$\mathbb{E}\big|\nabla_2 L_\zeta(x_t,\eta_{x_t}^d(\zeta))\big|^2 \le \mathcal{O}\Big(\sqrt{\frac{\widehat{\Delta}K'R_2}{DB_2}}\Big) + \mathcal{O}\Big(\frac{\hat{\Delta}K'}{D}\Big). \tag{13}$$
In particular, choose $B_2 = \Theta(\epsilon_1^{-2})$, then Algorithm 2 outputs $\eta_{x_t}^{\bar{d}}(\zeta)$ satisfy $\big|\nabla_2 L_\zeta(x,\eta_x^{\bar{d}}(\zeta))\big| \le \epsilon_1$ after $\Omega(\widehat{\Delta}K'R_2\epsilon_1^{-2})$ iterations. The overall sample complexity of Alogrithm 2 is $\Omega(\widehat{\Delta}K'R_2\epsilon_1^{-4})$.

However, we noticed both algorithms are sub-optimal in terms of convergence rate. In the future, we intend to enhance current algorithms with improved analysis and conduct experiments to evaluate the performance and theoretical guarantees of proposed algorithms.

Acknowledgment

Y. Yang and Y. Zhou's work are supported by the National Science Foundation under grants CCF-2106216, DMS-2134223, ECCS-2237830 (CAREER).

References

[1] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, 2018.

[2] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Niles-Weed. Massively scalable sinkhorn distances via the nyström method. In *Advances in Neural Information Processing Systems*, 2019.

[3] Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214, 2022.

[4] Genevay Aude, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Adavances in Neural Information Processing Systems*, 2016.

[5] Khanh Do Ba, Huy L Nguyen, Huy N Nguyen, and Ronitt Rubinfeld. Sublinear time algorithms for earth mover's distance. *Theory of Computing Systems*, 48:428–442, 2010.

[6] Ruidi Chen and Ioannis Ch. Paschalidis. A distributionally robust optimization approach for outlier detection. In *IEEE Conference on Decision and Control (CDC)*, pages 352–357, 2018.

[7] Xi Chen, Simai He, Bo Jiang, Christopher Thomas Ryan, and Teng Zhang. The discrete moment problem with nonconvex shape constraints. *Operations Research*, 69(1):279–296, 2021.

[8] Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, volume 202, 2023.

[9] Meysam Cheramin, Jianqiang Cheng, Ruiwei Jiang, and Kai Pan. Computationally efficient approximations for distributionally robust optimization under moment and wasserstein ambiguity. *INFORMS Journal on Computing*, 34(3):1768–1794, 2022.

[10] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

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

[12] John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *the Annals of Statistics*, 49:1378–1406, 2020.

[13] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

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

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

[16] Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1608–1617, 09–11 Apr 2018.

[17] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23:2341–2368, 2013.

[18] Scarf Herbert. A min-max solution of an inventory problem. in: Studies in the mathematical theory of inventory and production, 1957.

[19] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Optimization Online*, 1(2):9, 2013.

[20] Jikai Jin, Bohang Zhang, Haiyang Wang, and Liwei Wang. Non-convex distributionally robust optimization: Non-asymptotic analysis. In *Adavances in Neural Information Processing Systems*, 2021.

[21] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26, 2013.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[23] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. *Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning*, chapter 6, pages 130–166. Informs, 2019.

[24] Daniel Levy, Yair Carmon, John C. Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems*, 2020.

[25] Fengqiao Luo and Sanjay Mehrotra. Decomposition algorithm for distributionally robust optimization using wasserstein metric with an application to a class of regression models. *European Journal of Operational Research*, 278(1):20–35, 2019.

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

[27] Giorgio Patrini, Rianne Van den Berg, Patrick Forre, Marcello Carioni, Samarth Bhargav, Max Welling, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. In *Uncertainty in Artificial Intelligence*, pages 733–743, 2020.

[28] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *IEEE International Conference on Computer Vision*, pages 460–467, 2009.

[29] Qi Qi, Jiameng Lyu, Er Wei Bai, Tianbao Yang, et al. Stochastic constrained dro with a complexity independent of sample size. *Transactions on Machine Learning Research*, 2023.

[30] Wei Qian, Bin Hong, Deng Cai, Xiaofei He, Xuelong Li, et al. Non-negative matrix factorization with sinkhorn distance. In *International Joint Conference on Artificial Intelligence*, pages 1960–1966, 2016.

[31] Julien Rabin and Nicolas Papadakis. Convex color image segmentation with optimal transport distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 256–269, 2015.

[32] Hitesh Sapkota, Yiming Ying, Feng Chen, and Qi Yu. Distributionally robust optimization for deep kernel multiple instance learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2188–2196. PMLR, 2021.

[33] Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[34] Alexander Shapiro. *Chapter 9: Background Material*, chapter 9, pages 403–495. Society for Industrial and Applied Mathematics, 2009.

[35] Alexander Shapiro, Enlu Zhou, and Yifan Lin. Bayesian distributionally robust optimization. *SIAM Journal on Optimization*, 33(2):1279–1304, 2023.

[36] Evgeny Tenetov, Gershon Wolansky, and Ron Kimmel. Fast entropic regularized optimal transport using semidiscrete cost approximation. *SIAM Journal on Scientific Computing*, 40 (5):A3400–A3422, 2018.

[37] Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *arXiv: 2109.11926*, 2023.

[38] Junkang Wu, Jiawei Chen, Jiancan Wu, Wentao Shi, Xiang Wang, and Xiangnan He. Understanding contrastive learning via distributionally robust optimization. In *Adavances in Neural Information Processing Systems*, 2023.

[39] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

[40] Qi Zhang, Yi Zhou, Ashley Prater-Bennette ASHLEY, Lixin Shen, and Shaofeng Zou. Large-scale non-convex stochastic constrained distributionally robust optimization. In *NeurIPS Workshop on Optimization for Machine Learning*, 2023.