Sinkhorn Distributionally Robust Optimization

Jie Wang
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, jwang3163@gatech.edu

Rui Gao
Department of Information, Risk, and Operations Management, University of Texas at Austin, Austin, TX 78712, rui.gao@mccombs.utexas.edu

Yao Xie
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, yao.xie@isye.gatech.edu

We study distributionally robust optimization (DRO) with Sinkhorn distance—a variant of Wasserstein distance based on entropic regularization. We derive convex programming dual reformulation for general nominal distributions, transport costs, and loss functions. Compared with Wasserstein DRO, our proposed approach offers enhanced computational tractability for a broader class of loss functions, and the worst-case distribution exhibits greater plausibility in practical scenarios. To solve the dual reformulation, we develop a stochastic mirror descent algorithm with biased gradient oracles. Remarkably, this algorithm achieves near-optimal sample complexity for both smooth and nonsmooth loss functions, nearly matching the sample complexity of the Empirical Risk Minimization counterpart. Finally, we provide numerical examples using synthetic and real data to demonstrate its superior performance.

Key words: Wasserstein distributionally robust optimization, Sinkhorn distance, Duality theory, Stochastic mirror descent

1. Introduction

Decision-making problems under uncertainty arise in various fields such as operations research, machine learning, engineering, and economics. In these scenarios, uncertainty in the data arises from factors like measurement error, limited sample size, contamination, anomalies, or model misspecification. Addressing this uncertainty is crucial to obtain reliable and robust solutions. In recent years, Distributionally Robust Optimization (DRO) has emerged as a promising data-driven approach to tackle these challenges. DRO aims to find a minimax robust optimal decision that minimizes the expected loss under the most adverse distribution within a predefined set of relevant distributions, known as an ambiguity set. This approach provides a principled framework to handle uncertainty and obtain solutions that are resilient to distributional variations. It goes beyond the traditional sample average approximation (SAA) method used in stochastic programming and offers improved out-of-sample performance. For a comprehensive overview of DRO, we refer interested readers to the recent survey by [96].

At the core of distributionally robust optimization lies the crucial task of selecting an appropriate ambiguity set. An ambiguity set should strike a balance between computational tractability and practical interpretability while being rich enough to encompass relevant distributions and avoiding unnecessary ones that may lead to overly conservative decisions. In the literature, various formulations of DRO have been proposed, among which the ambiguity set based on Wasserstein distance has gained significant attention in recent years [126, 78, 18, 48]. The Wasserstein distance incorporates the geometry of the sample space, making it suitable for comparing distributions with non-overlapping supports and hedging against data perturbations [48]. The Wasserstein ambiguity set has received substantial theoretical attention, with provable performance guarantees [103, 17, 20, 19, 46]. Empirical success has also been demonstrated across a wide range of applications, including operations research [14, 29, 108, 85, 107, 123], machine learning [104, 24, 75, 15, 84, 114], stochastic control [128, 1, 111, 39, 129, 122], and more. For a comprehensive discussion and further references, we refer interested readers to [67].

However, the current Wasserstein DRO framework has its limitations. First, from a *computational efficiency* perspective, the tractability of Wasserstein DRO is typically only achieved under somewhat stringent conditions on the loss function, as its dual formulation involves a subproblem that requires the global supremum of some regularized loss function over the sample space. Table 1 summarizes the known tractable cases for solving Wasserstein DRO, $\min_{\theta\in\Theta}\max_{\mathbb{P}\in\mathfrak{M}}\ \mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)]$, where the loss function $f_\theta(z)$ is convex in θ belonging to a closed and convex feasible region Θ, and the ambiguity set $\mathfrak{M}$ is centered around a nominal distribution $\widehat{\mathbb{P}}$ and contains distributions supported on a space $\mathcal{Z}$. One general approach to solving Wasserstein DRO is to use a finite and discrete grid of scenarios to approximate the entire sample space. This involves solving the formulation restricted to the approximated sample space [92, 26, 73], but suffers from the curse of dimensionality. Simplified convex reformulations are known when the loss function can be expressed as a pointwise maximum of finitely many concave functions [43, 48, 102], or when the loss is the generalized linear model [104, 130, 101, 102]. In addition, efficient first-order algorithms have been developed for Wasserstein DRO with strongly convex transport cost, smooth loss functions, and sufficiently small radius (or equivalently, sufficiently large Lagrangian multiplier) so that the involved subproblem becomes strongly convex [109, 21]. However, beyond these conditions on the loss function and the transport cost, solving Wasserstein DRO becomes a computationally challenging task. Second, from a *modeling* perspective, in data-driven Wasserstein DRO, where the nominal distribution is finitely supported (usually the empirical distribution), the worst-case distribution is shown to be a discrete distribution [48] (which is unique when the regularized loss function has a unique maximizer). This is the case even though the underlying true distribution in many practical applications may be continuous. Consequently, concerns arise regarding whether Wasserstein DRO hedges the right family of distributions and whether it induces overly conservative solutions.

Table 1 Existing tractability result of Wasserstein DRO

References	Loss function $f_\theta(z)$	Transport cost	Nominal distribution $\widehat{\mathbb{P}}$	Support $\mathcal{Z}$
[92, 26, 73]	General	General	General	Discrete and finite set
[43, 48, 102]	Piecewise concave in z	General	Empirical distribution	General
[104, 130, 101, 102]	Generalized linear model in (z,θ)	General	General	Whole Euclidean space[1]
[109, 21]	$z\mapsto f_\theta(z)-\lambda^* c(x,z)$ is strongly concave[2]	Strongly convex function[3]	General	General

To address the aforementioned concerns while retaining the advantages of Wasserstein DRO, we propose a novel approach called Sinkhorn DRO. Sinkhorn DRO leverages the Sinkhorn distance [36], which hedges against distributions that are close to a given nominal distribution in Sinkhorn distance. The Sinkhorn distance can be viewed as a smoothed version of the Wasserstein distance and is defined as the minimum transport cost between two distributions associated with an optimal transport problem with entropic regularization (see Definition 1 in Section 2). To the best of our knowledge, this paper is the first to explore the DRO formulation using the Sinkhorn distance. Our work makes several contributions, which are summarized below:

(I) We derive a strong duality reformulation for Sinkhorn DRO (Theorem 1) in a highly general setting, where the loss function, transport cost, nominal distribution, and probability support are allowed to be arbitrary. The dual objective of Sinkhorn DRO smooths the dual objective of Wasserstein DRO, where the level of smoothness is controlled by the entropic regularization parameter (Remark 3). Additionally, the dual objective of Sinkhorn DRO is upper bounded by that of the KL-divergence DRO, with the nominal distribution being a kernel density estimator (Remark 5).

(II) Our duality proof yields an insightful characterization of the worst-case distribution in Sinkhorn DRO (Remark 4). Unlike Wasserstein DRO, where the worst-case distribution is typically discrete and finitely supported, the worst-case distribution in Sinkhorn DRO is absolutely continuous with respect to a pre-specified reference measure, such as the Lebesgue or counting measure. This characteristic of Sinkhorn DRO highlights its flexibility as a modeling choice and provides a more realistic representation of uncertainty that better aligns with the underlying true distribution in practical scenarios.

(III) The dual reformulation of Sinkhorn DRO can be viewed as a conditional stochastic optimization problem, which has been the subject of recent research [59, 61, 60]. In our work, we introduce and analyze an efficient stochastic mirror descent algorithm with biased gradient oracles to solve this problem efficiently (Section 4). The proposed algorithm leverages the trade-off between bias and second-order moment of stochastic gradient estimators. By carefully balancing these factors, our algorithm achieves a near-optimal sample complexity of $\widetilde{\mathcal{O}}(\delta^{-2})$ and storage cost of $\widetilde{\mathcal{O}}(1)$ for finding a δ-optimal solution[4]. In contrast to Wasserstein DRO, the dual problem of Sinkhorn DRO offers computational tractability for a significantly broader range of loss functions, transport costs, nominal distributions, and probability support.

(IV) To validate the effectiveness and efficiency of the proposed Sinkhorn DRO model, we conduct a series of experiments in Section 5, including the newsvendor problem, mean-risk portfolio optimization, and multi-class adversarial classification. Using synthetic and real datasets, we compare the Sinkhorn DRO model against benchmarks such as SAA, Wasserstein DRO, and KL-divergence DRO. The results demonstrate that the Sinkhorn DRO model consistently outperforms the benchmarks in terms of out-of-sample performance and computational speed.

Related Literature

In the following, we first compare our work with the four most closely related papers that appear recently.

Feng and Schlögl [44] studied the Wasserstein DRO formulation with an additional differential entropy constraint on the optimal transport mapping, which can be viewed as a variation of Sinkhorn distance. They derived a weak dual formulation and characterized the worst-case distribution under the assumption that strong duality holds. It is important to note that such an assumption cannot be taken for granted for the considered infinite-dimensional problem. Moreover, their results heavily depend on the assumption that the nominal distribution $\widehat{\mathbb{P}}$ is absolutely continuous with respect to the Lebesgue measure. This limits the applicability of their formulation in data-driven settings where $\widehat{\mathbb{P}}$ is discrete. Since the initial submission of our work, Azizian et al. [6] have presented a duality result similar to ours, but with different assumptions. Their results apply to more general regularization beyond entropic regularization, but they assume a continuous loss function and a compact probability space under the Slater condition. Song et al. [112] have recently explored the application of Sinkhorn DRO in reinforcement learning. Their duality proof rely on the boundedness of the loss function and the discreteness of the probability support. These three papers do not present numerical algorithms to solve the dual formulation. Blanchet and Kang [16, Section 3.2] solved a log-sum-exp approximation of the Wasserstein DRO dual formulation. This smooth approximation can be viewed as a special case of the dual reformulation of our Sinkhorn DRO model. However, their study did not specifically explore the primal form of Sinkhorn DRO. Their algorithm employed unbiased gradient estimators, even though the second-order moment could be unbounded. The paper did not provide explicit theoretical convergence guarantees for their algorithm. Additionally, numerical comparisons detailed in Appendix EC.2.1 suggest that our proposed algorithm outperforms theirs in terms of empirical performance.

Next, we review papers on several related topics.

On DRO models. In the literature on DRO, there are two main approaches to constructing ambiguity sets. The first approach involves defining ambiguity sets based on descriptive statistics, such as support information [12], moment conditions [100, 37, 54, 134, 125, 28, 13], shape constraints [94, 117], marginal distributions [45, 80, 2, 40], etc. The second approach, which has gained popularity in recent years, involves considering distributions within a pre-specified statistical distance from a nominal distribution. Commonly used statistical distances in the literature include ϕ-divergence [62, 10, 124, 9, 41], Wasserstein distance [92, 126, 78, 132, 18, 48, 27, 127], and maximum mean discrepancy [113, 133]. Our proposed Sinkhorn DRO can be seen as a variant of the Wasserstein DRO. In the literature on Wasserstein DRO, researchers have also explored the regularization effects and statistical inference of the approach. In particular, it has been shown that Wasserstein DRO is asymptotically equivalent to a statistical learning problem with variation regularization [47, 17, 103]. When the radius is chosen properly, the worst-case loss of Wasserstein DRO serves as an upper confidence bound on the true loss [17, 20, 46, 19]. Variants of Wasserstein DRO have been proposed by combining it with other information, such as moment information [50, 119] or marginal distributions [49, 42] to enhance its modeling capabilities.

On Sinkhorn distance. Sinkhorn distance [36] was proposed to improve the computational complexity of Wasserstein distance, by regularizing the original mass transportation problem with relative entropy penalty on the transport plan. It has been demonstrated to be beneficial because of lower computational cost in various applications, including domain adaptations [33, 34, 32], generative modeling [52, 90, 74, 89], dimension reduction [71, 120, 121, 63], etc. In particular, this distance can be computed from its dual form by optimizing two blocks of decision variables alternatively, which only requires simple matrix-vector products and therefore significantly improves the computation speed [91, 77, 72, 3]. Such an approach first arises in economics and survey statistics [66, 131, 38, 7], and its convergence analysis is attributed to the mathematician Sinkhorn [110], which gives the name of Sinkhorn distance. It is important to note that the computation of Sinkhorn distance and solving Sinkhorn DRO formulation are two different computational tasks. Indeed, the two problems have intrinsically different structures from the dual formulation point of view: the former is a standard stochastic optimization problem where an unbiased gradient estimate can be easily obtained, whereas the latter can be viewed as a conditional stochastic optimization (CSO) [59] involving an expectation of nonlinear transformation of a conditional expectation, where the unbiased gradient is challenging to compute. Therefore, existing approaches for computing Sinkhorn distance cannot be directly applied to solve the Sinkhorn DRO. To our knowledge, the study of Sinkhorn distance is new in the DRO literature.

On algorithms for solving DRO models. In the introduction, we have elaborated on the literature that proposes efficient optimization algorithms for solving the Wasserstein DRO dual formulation [132, 26, 73, 109, 43, 48, 104, 130, 101, 21, 102], in which the tractability is limited to a certain class of loss functions, transport costs, and nominal distributions. To solve the ϕ-divergence DRO, one common approach is to employ sample average approximation (SAA) to approximate the dual formulation. However, SAA requires storing the entire set of samples, making it inefficient in terms of storage usage. An alternative approach is to use first-order stochastic gradient algorithms, which are more storage-efficient. These algorithms have the advantage of complexity that can be independent of the sample size of the nominal distribution [68, 79, 95]. Our derived dual reformulation of Sinkhorn DRO can be seen as an instance of the CSO problem [59, 61, 60]. In this context, we have developed stochastic mirror descent algorithms with biased gradient oracles. Notably, our proof can be adjusted to show that the proposed algorithm achieves near-optimal complexity for general CSO problems with both smooth and nonsmooth loss functions, marking an improvement over the state-of-the-art [61, Theorem 3.2] that is sub-optimal for nonsmooth loss functions.

The rest of the paper is organized as follows. In Section 2, we describe the main formulation for the Sinkhorn DRO model. In Section 3, we develop its strong dual reformulation. In Section 4, we

propose a first-order optimization algorithm that solves the reformulation efficiently. We report several numerical results in Section 5, and conclude the paper in Section 6. All omitted proofs can be found in Appendices.

2. Model Setup

Notation. Assume that the logarithm function $\log$ is taken with base e. For a positive integer N, we write $[N]$ for $\{1, 2, \ldots, N\}$. For a measurable set $\mathcal{Z}$, denote by $\mathcal{M}(\mathcal{Z})$ the set of measures (not necessarily probability measures) on $\mathcal{Z}$, and $\mathcal{P}(\mathcal{Z})$ the set of probability measures on $\mathcal{Z}$. Given a probability distribution $\mathbb{P}$ and a measure μ, we denote $\operatorname{supp}\mathbb{P}$ the support of $\mathbb{P}$, and write $\mathbb{P} \ll \mu$ if $\mathbb{P}$ is absolutely continuous with respect to μ. Given a measure $\mu \in \mathcal{M}(\mathcal{Z})$ and a measurable variable $f\colon\ \mathcal{Z} \to \mathbb{R}$, we write $\mathbb{E}_{z\sim\mu}[f(z)]$ for $\int f(z)\,\mathrm{d}\mu(z)$. For a given element x, denote by δ_x the one-point probability distribution supported on $\{x\}$. Denote $\mathbb{P}\otimes\mathbb{Q}$ as the product measure of two probability distributions $\mathbb{P}$ and $\mathbb{Q}$. Denote by $\mathrm{Proj}_{1\#}\gamma$ and $\mathrm{Proj}_{2\#}\gamma$ the first and the second marginal distributions of γ, respectively. For a given function $\omega\colon\ \Theta \to \mathbb{R}$, we say it is κ-strongly convex with respect to norm $\|\cdot\|$ if $\langle \theta' - \theta, \nabla\omega(\theta') - \nabla\omega(\theta)\rangle \geq \kappa\|\theta' - \theta\|^2, \forall \theta, \theta' \in \Theta$.

We first review the definition of Sinkhorn distance.

Definition 1 (Sinkhorn Distance). Let $\mathcal{Z}$ be a measurable set. Consider distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{Z})$, and let $\mu, \nu \in \mathcal{M}(\mathcal{Z})$ be two reference measures such that $\mathbb{P} \ll \mu$, $\mathbb{Q} \ll \nu$. For regularization parameter $\epsilon \geq 0$, the *Sinkhorn distance* between two distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\mathcal{W}_\epsilon(\mathbb{P}, \mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{\mathbb{E}_{(x,y)\sim\gamma}[c(x,y)] + \epsilon H(\gamma \,|\, \mu\otimes\nu)\right\},$$

where $\Gamma(\mathbb{P}, \mathbb{Q})$ denotes the set of joint distributions whose first and second marginal distributions are $\mathbb{P}$ and $\mathbb{Q}$ respectively, $c(x, y)$ denotes the transport cost, and $H(\gamma \,|\, \mu\otimes\nu)$ denotes the relative entropy of γ with respect to the product measure $\mu\otimes\nu$:

$$H(\gamma \,|\, \mu\otimes\nu) = \mathbb{E}_{(x,y)\sim\gamma}\left[\log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)}\right)\right],$$

where $\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)}$ stands for the density ratio of γ with respect to $\mu\otimes\nu$ evaluated at (x, y). ♢

Remark 1 (Variants of Sinkhorn Distance). Sinkhorn distance in Definition 1 is based on general reference measures μ and ν. Special forms of distance have been investigated in the literature. For instance, the entropic regularized optimal transport distance $\mathcal{W}_\epsilon^{\mathtt{Ent}}(\mathbb{P}, \mathbb{Q})$ [36, Equation (2)] chooses μ and ν as the Lebesgue measure when the corresponding $\mathbb{P}$ and $\mathbb{Q}$ are continuous, or counting measures if $\mathbb{P}$ and $\mathbb{Q}$ are discrete. For given $\mathbb{P}$ and $\mathbb{Q}$, one can check the two distances above are equivalent up to a constant:

$$\begin{aligned}\mathcal{W}_\epsilon^{\mathtt{Ent}}(\mathbb{P}, \mathbb{Q}) &= \mathcal{W}_\epsilon(\mathbb{P}, \mathbb{Q}) + \mathbb{E}_{(x,y)\sim\gamma}\left[\log\left(\frac{\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)}{\mathrm{d}x\,\mathrm{d}y}\right)\right] \\ &= \mathcal{W}_\epsilon(\mathbb{P}, \mathbb{Q}) + \mathbb{E}_{x\sim\mathbb{P}}\left[\log\left(\frac{\mathrm{d}\mu(x)}{\mathrm{d}x}\right)\right] + \mathbb{E}_{y\sim\mathbb{Q}}\left[\log\left(\frac{\mathrm{d}\nu(y)}{\mathrm{d}y}\right)\right].\end{aligned}$$

Another variant is to chose μ and ν to be $\mathbb{P}, \mathbb{Q}$, respectively [51, Section 2]. A hard-constrained variant of the relative entropy regularization has been discussed in [36, Definition 1] and [8]:

$$\mathcal{W}_R^{\mathtt{Info}}(\mathbb{P}, \mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{\mathbb{E}_{(X,Y)\sim\gamma}[c(X,Y)]\colon\ H(\gamma \,|\, \mathbb{P}\otimes\mathbb{Q}) \leq R\right\},$$

where $R \geq 0$ quantifies the upper bound for the relative entropy between distributions γ and $\mathbb{P}\otimes\mathbb{Q}$. ♣

REMARK 2 (CHOICE OF REFERENCE MEASURES). We discuss below our choice of the two reference measures μ and ν in Definition 1. For the reference measure μ, observe from the definition of relative entropy and the law of probability, we can see that the regularization term in $\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})$ can be written as

$$\begin{aligned} H(\gamma \,|\, \mu \otimes \nu) &= \mathbb{E}_{(x,y)\sim\gamma}\left[\log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}\nu(y)}\right) + \log\left(\frac{\widehat{\mathbb{P}}(x)}{\mathrm{d}\mu(x)}\right)\right] \\ &= \mathbb{E}_{(x,y)\sim\gamma}\left[\log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}\nu(y)}\right)\right] + \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\left(\frac{\widehat{\mathbb{P}}(x)}{\mathrm{d}\mu(x)}\right)\right]. \end{aligned}$$

Therefore, any choice of the reference measure μ satisfying $\widehat{\mathbb{P}} \ll \mu$ is equivalent up to a constant. For simplicity, in the sequel we will take $\mu = \widehat{\mathbb{P}}$. For the reference measure ν, observe that the worst-case solution $\mathbb{P}$ in (Primal) should satisfy that $\mathbb{P} \ll \nu$ since otherwise the entropic regularization in Definition 1 is undefined. As a consequence, we can choose ν which the underlying true distribution is absolutely continuous with respect to and is easy to sample from. For example, if we believe the underlying distribution is continuous, then we can choose ν to be the Lebesgue measure or Gaussian measure, or if we believe the underlying distribution is discrete, we can choose ν to be a counting measure. We refer to [93, Section 3.6] for the construction of a general reference measure. ♣

In this paper, we study the Sinkhorn DRO model. Given a loss function f, a nominal distribution $\widehat{\mathbb{P}}$ and the Sinkhorn radius ρ, the primal form of the worst-case expectation problem of Sinkhorn DRO is given by

$$V := \sup_{\mathbb{P}\in\mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})} \mathbb{E}_{z\sim\mathbb{P}}[f(z)], \tag{Primal}$$

where $\mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}}) := \left\{\mathbb{P} : \mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P}) \le \rho\right\}$ is the Sinkhorn ball of the radius ρ centered at the nominal distribution $\widehat{\mathbb{P}}$. Due to the convex entropic regularize $\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})$ with respect to $\mathbb{P}$ [35], the Sinkhorn distance $\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})$ is convex in $\mathbb{P}$, i.e., it holds that $\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \lambda\mathbb{P}_1 + (1-\lambda)\mathbb{P}_2) \le \lambda\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P}_1) + (1-\lambda)\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P}_2)$ for all probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ and all $0 \le \lambda \le 1$. Therefore, the Sinkhorn ball is a convex set, and the problem (Primal) is an (infinite-dimensional) convex program.

Our goal for the rest of the paper is to derive tractable reformulations and efficient algorithms to solve the Sinkhorn DRO model.

3. Strong Duality Reformulation

Problem (Primal) is an infinite-dimensional optimization problem over probability distributions. To obtain a more tractable form, in this section, we derive a strong duality result for (Primal). Our main goal is to derive the strong dual program

$$V_{\mathrm{D}} := \inf_{\lambda\ge 0}\left\{\lambda\rho + \lambda\epsilon\,\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda c(x,z))/(\lambda\epsilon)}\right]\right]\right\}, \tag{1}$$

where the dual variable λ corresponds to the Sinkhorn ball constraint in (Primal), and by convention, we define the dual objective evaluated at $\lambda = 0$ as the limit of the objective values with $\lambda \downarrow 0$, which equals the essential supremum of the objective function with respect to the measure ν. Or equivalently, by defining the constant

$$\overline{\rho} := \rho + \epsilon\,\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\mathbb{E}_{z\sim\nu}\left[e^{-c(x,z)/\epsilon}\right]\right], \tag{2}$$

and the kernel probability distribution

$$\mathrm{d}\mathbb{Q}_{x,\epsilon}(z) := \frac{e^{-c(x,z)/\epsilon}}{\mathbb{E}_{u\sim\nu}\left[e^{-c(x,u)/\epsilon}\right]}\,\mathrm{d}\nu(z), \tag{3}$$

we have

$$V_{\mathrm{D}} = \inf_{\lambda \geq 0} \left\{ \lambda \overline{\rho} + \lambda \epsilon \, \mathbb{E}_{x \sim \widehat{\mathbb{P}}} \Big[\log \mathbb{E}_{z \sim \mathbb{Q}_{x,\epsilon}} \big[e^{f(z)/(\lambda\epsilon)} \big] \Big] \right\}. \tag{Dual}$$

The rest of this section is organized as follows. In Section 3.1, we summarize our main results on the strong duality reformulation of Sinkhorn DRO. Next, we provide detailed discussions in Section 3.2. In Section 3.4, we provide a proof sketch of our main results.

3.1. Main Theorem

To make the above primal (Primal) and dual (Dual) problems well-defined, we introduce the following assumptions on the transport cost c, the reference measure ν, and the loss function f.

ASSUMPTION 1. (I) $\nu\{z:\ 0 \leq c(x,z) < \infty\} = 1$ *for* $\widehat{\mathbb{P}}$*-almost every* x*;*
(II) $\mathbb{E}_{z\sim\nu}\left[e^{-c(x,z)/\epsilon}\right] < \infty$ *for* $\widehat{\mathbb{P}}$*-almost every* x*;*
(III) $\mathcal{Z}$ *is a measurable space, and the function* $f:\ \mathcal{Z} \to \mathbb{R} \cup \{\infty\}$ *is measurable.*
(IV) *For every joint distribution* γ *on* $\mathcal{Z} \times \mathcal{Z}$ *with first marginal distribution* $\widehat{\mathbb{P}}$*, it has a regular conditional distribution* γ_x *given the value of the first marginal equals* x*.*

Assumption 1(I) implies that $0 \leq c(x,y) < \infty$ for $\widehat{\mathbb{P}} \otimes \nu$-almost every (x,y). By [91, Proposition 4.1], the Sinkhorn distance has an equivalent formulation

$$\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P}) = \min_{\gamma \in \Gamma(\widehat{\mathbb{P}}, \mathbb{P})} \int \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\mathcal{K}}(x,y)\right) \mathrm{d}\gamma(x,y), \quad \text{where } \mathrm{d}\mathcal{K}(x,y) = e^{-c(x,y)/\epsilon}\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y).$$

Therefore Assumption 1(I) ensures that the reference measure $\mathcal{K}$ is well-defined. Assumption 1(II) ensures the optimal transport mapping γ_* for Sinkhorn distance $\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})$ exists with density value $\frac{\mathrm{d}\gamma_*(x,y)}{\mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}\nu(y)} \propto e^{-c(x,y)/\epsilon}$. Hence, Assumption 1(I) and 1(II) together ensure the Sinkhorn distance is well-defined. Assumption 1(III) ensures the expected loss $\mathbb{E}_{z\sim\mathbb{P}}[f(z)]$ to be well-defined for any distribution $\mathbb{P}$. Assumption 1(IV) ensures the joint distribution γ can be written as $\mathrm{d}\gamma(x,z) = \mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}\gamma_x(z)$ and the law of total expectation holds; we refer to [64, Chapter 5] for the concept of the regular conditional distribution.

To distinguish the cases $V_D < \infty$ and $V_D = \infty$, we introduce the light-tail condition on f in Condition 1. In Appendix EC.3, we present sufficient conditions for Condition 1 that are easy to verify.

CONDITION 1. *There exists* $\lambda > 0$ *such that* $\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[e^{f(z)/(\lambda\epsilon)}] < \infty$ *for* $\widehat{\mathbb{P}}$*-almost every* x*.*

In the following, we provide the main results of the strong duality reformulation.

THEOREM 1 **(Strong Duality)**. *Let* $\widehat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$*, and assume Assumption 1 holds. Then the following holds:*
(I) *The primal problem* (Primal) *is feasible if and only if* $\overline{\rho} \geq 0$*;*
(II) *Whenever* $\overline{\rho} \geq 0$*, it holds that* $V = V_D$*.*
(III) *If, in addition, Condition 1 holds and* $\overline{\rho} > 0$*, then* $V = V_{\mathrm{D}} < \infty$*; otherwise* $V = V_{\mathrm{D}} = \infty$*.*

We remark that if $\overline{\rho} < 0$, by convention, $V = -\infty$ and $V_D = -\infty$ as well by Lemma EC.3 in Appendix EC.5. Therefore, we have $V = V_D$ as long as Assumption 1 holds.

3.2. Discussions

In the following, we make several remarks regarding the strong duality result.

REMARK 3 (COMPARISON WITH WASSERSTEIN DRO). As the regularization parameter $\epsilon \to 0$, the dual objective of the Sinkhorn DRO (Dual) converges to (see Appendix EC.4 for details)

$$\lambda\rho + \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\sup_{z\in\operatorname{supp}\nu}\big\{f(z)-\lambda c(x,z)\big\}\Big],$$

which essentially follows from the fact that the log-sum-exp function is a smooth approximation of the supremum. Particularly, when $\operatorname{supp}\nu = \mathcal{Z}$, the dual objective of the Sinkhorn DRO converges to the dual objective of the Wasserstein DRO [48, Theorem 1]. The main computational difficulty in Wasserstein DRO is solving the maximization problem inside the expectation above. All results in Table 1 ensure the tractability of this inner maximization. In contrast, as Sinkhorn DRO does not need to solve this maximization, it does not need stringent assumptions on $f(\cdot)$ and thus is tractable for a larger class of loss functions, as we will elaborate on in Section 4.

We also remark that Sinkorn DRO and Wasserstein DRO result in different conditions for finite worst-case values. From Condition 1 we see that the Sinkhorn DRO is finite if and only if under a light-tail condition on f, while the Wasserstein DRO is finite if and only if the loss function satisfies a growth condition [48, Theorem 1 and Proposition 2]: $f(z) \le L_f c(z, z_0) + M, \forall z \in \mathcal{Z}$ for some constants $L_f, M > 0$ and some $z_0 \in \mathcal{Z}$. ♣

REMARK 4 (WORST-CASE DISTRIBUTION). Assume the optimal Lagrangian multiplier in (Dual) $\lambda^* > 0$. As we will demonstrate in the proof of Theorem 1, the worst-case distribution for (Primal) maps every $x \in \operatorname{supp}\widehat{\mathbb{P}}$ to a (conditional) distribution whose density function (with respect to ν) at z is

$$\alpha_x \cdot \exp\Big(\big(f(z)-\lambda^* c(x,z)\big)/(\lambda^*\epsilon)\Big),$$

where $\alpha_x := \big(\mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)}\big]\big)^{-1}$ is a normalizing constant to ensure the conditional distribution well-defined. As such, the density of worst-case distribution becomes

$$\frac{\mathrm{d}\mathbb{P}_*(z)}{\mathrm{d}\nu(z)} = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\alpha_x \cdot \exp\Big(\big(f(z)-\lambda^* c(x,z)\big)/(\lambda^*\epsilon)\Big)\right],$$

from which we can see that the worst-case distribution shares the same support as the measure ν. Particularly, when $\widehat{\mathbb{P}}$ is the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{\hat{x}_i}$ and ν is any continuous distribution on $\mathbb{R}^d$, the worst-case distribution $\mathbb{P}_*$ is supported on the entire $\mathbb{R}^d$. In contrast, the worst-case distribution for Wasserstein DRO is supported on at most $n+1$ points [48].

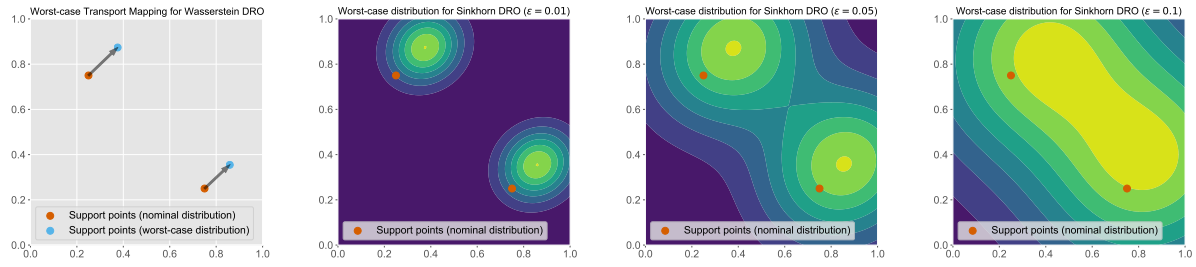

Figure 1 Visualization of worst-case distributions from Wasserstein DRO (left plot) and Sinkhorn DRO models (right three plots) with varying choices of ϵ.

In Figure 1 we visualize the worst-case distributions from Wasserstein/Sinkhorn DRO models. The loss function and transport cost used in this plot follow the setup described in Example 2. The Wasserstein ball radius, Sinkhorn ball radius, and entropic regularization value are fine-tuned to

ensure that the optimal dual multipliers for all instances equal 5. Notably, the support points of the worst-case distributions from the Wasserstein DRO model correspond to the modes of the continuous worst-case distributions from the Sinkhorn DRO model.

The above demonstrates another difference, or advantage possibly, of Sinkhorn DRO compared with Wasserstein DRO. Indeed, for many practical problems, the underlying distribution is modeled as a continuous distribution. The worst-case distribution for Wasserstein DRO is often finitely supported, raising the concern of whether it hedges against the wrong family of distributions and thus results in suboptimal solutions. The numerical results in Section 5 demonstrate some empirical advantages of Sinkhorn DRO. ♣

REMARK 5 (CONNECTION WITH KL-DRO). Using Jensen's inequality, we can see that the dual objective function of the Sinkhorn DRO model can be upper-bounded as

$$\lambda\overline{\rho} + \lambda\epsilon \log\left(\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\right),$$

which corresponds to the dual objective for the following KL-divergence DRO [10]

$$\sup_{\mathbb{P}} \left\{\mathbb{E}_{z\sim\mathbb{P}}[f(z)] :\ D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{P}^0) \le \overline{\rho}/\epsilon\right\},$$

where $\mathbb{P}^0$ satisfies $\mathrm{d}\mathbb{P}^0(z) = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}[\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)]$, which can be viewed as a non-parametric kernel density estimation constructed from $\widehat{\mathbb{P}}$. Particularly, when $\widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^n \delta_{\hat{x}_i}$, $\mathcal{Z} = \mathbb{R}^d$ and $c(x,y) = \|x-y\|_2^2$, $\mathbb{P}^0$ is kernel density estimator with Gaussian kernel and bandwidth ϵ:

$$\frac{\mathrm{d}\mathbb{P}^0(z)}{\mathrm{d}z} = \frac{1}{n}\sum_{i=1}^n K_\epsilon(z - x_i), \quad z\in\mathbb{R}^d,$$

where $K_\epsilon(x) \propto \exp(-\|x\|_2^2/\epsilon)$ represents the Gaussian kernel. By Lemma 2 and divergence inequality [35, Theorem 2.6.3], we can see the Sinkhorn DRO with $\overline{\rho} = 0$ is reduced to the following SAA model based on the distribution $\mathbb{P}^0$:

$$V = \mathbb{E}_{z\sim\mathbb{P}^0}[f(z)] = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[f(z)]. \tag{4}$$

In non-parameteric statistics, the optimal bandwidth to minimize the mean-squared-error between the estimated distribution $\mathbb{P}_0$ and the underlying true one is at rate $\epsilon = O(n^{-1/(d+4)})$ [56, Theorem 4.2.1]. However, such an optimal choice for the kernel density estimator may not be the optimal choice for optimizing the out-of-sample performance of the Sinkhorn DRO. In our numerical experiments in Section 5, we select ϵ based on cross-validation. ♣

REMARK 6 (CONNECTION WITH BAYESIAN DRO). The recent Bayesian DRO framework [106] proposed to solve

$$\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\sup_{\mathbb{P}}\left\{\mathbb{E}_{z\sim\mathbb{P}}[f(z)] :\ \mathbb{P}\in\mathcal{P}_x\right\}\right],$$

where $\widehat{\mathbb{P}}$ is a special posterior distribution constructed from collected observations, and the ambiguity set $\mathcal{P}_x$ is typically constructed as a KL-divergence ball, i.e., $\mathcal{P}_x := \{\mathbb{P} :\ D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}_x) \le \eta\}$, with $\mathbb{Q}_x$ being the parametric distribution conditioned on x. According to [106, Section 2.1.3], a relaxation of the Bayesian DRO dual formulation is given by

$$\inf_{\lambda\ge 0}\left\{\lambda\eta + \lambda\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\mathbb{E}_{z\sim\mathbb{Q}_x}\left[e^{f(z)/\lambda}\right]\right]\right\}.$$

When specifying the parametric distribution $\mathbb{Q}_x$ as the kernel probability distribution in (3) and applying the change-of-variable technique such that λ is replaced with $\lambda\epsilon$, this relaxed formulation becomes

$$\inf_{\lambda\ge 0}\left\{\lambda(\eta\epsilon) + \lambda\epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\mathbb{E}_{z\sim\mathbb{Q}_x}\left[e^{f(z)/(\lambda\epsilon)}\right]\right]\right\}.$$

In comparison with (Dual), we find the Sinkhorn DRO model can be viewed as a special relaxation formulation of the Bayesian DRO model. ♣

3.3. Examples

In the following, we provide several cases in which our strong dual reformulation (Dual) can be simplified into more tractable formulations.

EXAMPLE 1 (LINEAR LOSS). Suppose that the loss function $f(z) = a^{\mathrm{T}} z$, support $\mathcal{Z} = \mathbb{R}^d$, ν is the corresponding Lebesgue measure, and the transport cost is the Mahalanobis distance, i.e., $c(x,y) = \frac{1}{2}(x-y)^{\mathrm{T}}\Omega(x-y)$, where Ω is a positive definite matrix. In this case, we have the reference measure $\mathbb{Q}_{x,\epsilon} \sim \mathcal{N}(x, \epsilon\Omega^{-1})$. As a consequence, the dual problem can be written as

$$V_{\mathrm{D}} = \inf_{\lambda>0} \left\{ \lambda\overline{\rho} + \lambda\epsilon \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\Lambda_x(\lambda)\Big] \right\},$$

where

$$\Lambda_x(\lambda) = \log \mathbb{E}_{z\sim\mathcal{N}(x,\epsilon\Omega^{-1})}\left[e^{a^\top z/(\lambda\epsilon)}\right] = \frac{a^{\mathrm{T}}x}{\lambda\epsilon} + \frac{a^{\mathrm{T}}\Omega^{-1}a}{2\lambda^2\epsilon}.$$

Therefore

$$V_{\mathrm{D}} = a^{\mathrm{T}}\mathbb{E}_{x\sim\widehat{\mathbb{P}}}[x] + \sqrt{2\overline{\rho}}\sqrt{a^{\mathrm{T}}\Omega^{-1}a} := \mathbb{E}_{x\sim\widehat{\mathbb{P}}}[a^{\mathrm{T}}x] + \sqrt{2\overline{\rho}}\cdot\|a\|_{\Omega^{-1}}.$$

This indicates that the Sinkhorn DRO is equivalent to an empirical risk minimization with norm regularization, and can be solved efficiently using algorithms for the second-order cone program. ♣

EXAMPLE 2 (QUADRATIC LOSS). Consider the example of linear regression with quadratic loss $f_\theta(z) = (a^{\mathrm{T}}\theta - b)^2$, where $z := (a,b)$ denotes the predictor-response pair, $\theta \in \mathbb{R}^d$ denotes the fixed parameter choice, and $\mathcal{Z} = \mathbb{R}^{d+1}$. Taking ν as the Lebesgue measure and the transport cost as $c((a,b),(a',b')) = \frac{1}{2}\|a-a'\|_2^2 + \infty|b-b'|$. In this case, the dual problem becomes

$$V_{\mathrm{D}} = \mathbb{E}_{z\sim\widehat{\mathbb{P}}}[(a^{\mathrm{T}}\theta - b)^2] + \inf_{\lambda>2\|\theta\|_2^2} \left\{ \lambda\overline{\rho} + \frac{\mathbb{E}_{z\sim\widehat{\mathbb{P}}}[(a^{\mathrm{T}}\theta-b)^2]}{\frac{1}{2}\lambda\|\theta\|_2^{-2} - 1} - \frac{\lambda\epsilon}{2}\log\det\left(I - \frac{\theta\theta^{\mathrm{T}}}{\frac{1}{2}\lambda}\right) \right\}.$$

In comparison with the corresponding Wasserstein DRO formulation with radius ρ (see, e.g., [19, Example 4])

$$V_{\mathrm{D}}^{\text{Wasserstein}} = \mathbb{E}_{z\sim\widehat{\mathbb{P}}}[(a^{\mathrm{T}}\theta - b)^2] + \inf_{\lambda>2\|\theta\|_2^2} \left\{ \lambda\rho + \frac{\mathbb{E}_{z\sim\widehat{\mathbb{P}}}[(a^{\mathrm{T}}\theta-b)^2]}{\frac{1}{2}\lambda\|\theta\|_2^{-2} - 1} \right\},$$

one can check in this case the Sinkhorn DRO formulation is equivalent to the Wasserstein DRO with log-determinant regularization. ♣

When the support $\mathcal{Z}$ is finite, the following result presents a conic programming reformulation.

COROLLARY 1 **(Conic Reformulation for Finite Support)**. *Suppose that the support contains $L_{\max}$ elements, i.e., $\mathcal{Z} = \{z_\ell\}_{\ell=1}^{L_{\max}}$, and the nominal distribution $\widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^n \delta_{\hat{x}_i}$. If Condition 1 holds and $\overline{\rho} \geq 0$, the dual problem (Dual) can be formulated as the following conic optimization:*

$$\begin{aligned} V_{\mathrm{D}} = \min_{\substack{\lambda\geq 0, s\in\mathbb{R}^n,\\ a\in\mathbb{R}^{n\times L}}} \quad & \lambda\overline{\rho} + \frac{1}{n}\sum_{i=1}^n s_i \\ \text{s.t.} \quad & \lambda\epsilon \geq \sum_{\ell=1}^{L_{\max}} q_{i,\ell} a_{i,\ell}, i\in[n], \\ & (\lambda\epsilon, a_{i,\ell}, f(z_\ell) - s_i) \in \mathcal{K}_{\exp}, i\in[n], \ell\in[L]. \end{aligned} \tag{5}$$

where $q_{i,\ell} := \Pr_{z\sim\mathbb{Q}_{\hat{x}_i,\epsilon}}\{z = z_\ell\}$, with the distribution $\mathbb{Q}_{\hat{x}_i,\epsilon}$ defined in (3), and $\mathcal{K}_{\exp}$ denotes the exponential cone $\mathcal{K}_{\exp} = \{(\nu,\lambda,\delta)\in\mathbb{R}_+\times\mathbb{R}_+\times\mathbb{R}:\ \exp(\delta/\nu)\leq\lambda/\nu\}$.

Problem (5) is a convex program that minimizes a linear function with respect to linear and conic constraints, which can be solved using interior point algorithms [83, 118]. We will develop an efficient first-order optimization algorithm in Section 4 that is able to solve a more general problem (without a finite support).

3.4. Proof of Theorem 1

In this subsection, we present a sketch of the proof for Theorem 1. We begin with the weak duality result in Lemma 1, which can be shown by applying the Lagrangian weak duality.

LEMMA 1 **(Weak Duality)**. *Under Assumption 1, it holds that*

$$V \leq \inf_{\lambda \geq 0} \left\{ \lambda\rho + \lambda\epsilon\, \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\log \mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))/(\lambda\epsilon)}\big]\Big] \right\} = V_{\mathrm{D}}.$$

Proof of Lemma 1. Based on Definition 1 of Sinkhorn distance, we reformulate V as

$$V = \sup_{\gamma\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z}):\mathrm{Proj}_{1\#}\gamma=\widehat{\mathbb{P}}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[f(z)] : \mathbb{E}_{(x,z)\sim\gamma}\left[c(x,z) + \epsilon \log\left(\frac{\mathrm{d}\gamma(x,z)}{\mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}\nu(z)} \right)\right] \leq \rho \right\}.$$

By the law of total expectation, the constraint is equivalent to

$$\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}\left[c(x,z) + \epsilon \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)} \right)\right] \leq \rho,$$

and the primal problem is equivalent to

$$V = \sup_{\{\gamma_x\}_{x\in\mathrm{supp}\,\widehat{\mathbb{P}}}\subset\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}[f(z)] :\ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}\left[c(x,z) + \epsilon \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)} \right)\right] \leq \rho \right\}. \tag{6}$$

Introducing the Lagrange multiplier λ associated to the constraint, we reformulate V as

$$V = \sup_{\{\gamma_x\}_{x\in\mathrm{supp}\,\widehat{\mathbb{P}}}\subset\mathcal{P}(\mathcal{Z})} \left\{ \inf_{\lambda\geq 0} \left\{ \lambda\rho + \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}\left[f(z) - \lambda c(x,z) - \lambda\epsilon \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)} \right)\right]\right\}\right\}.$$

Interchanging the order of the supremum and infimum operators, we have that

$$V \leq \inf_{\lambda\geq 0} \left\{ \lambda\rho + \sup_{\{\gamma_x\}_{x\in\mathrm{supp}\,\widehat{\mathbb{P}}}\subset\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}\left[f(z) - \lambda c(x,z) - \lambda\epsilon \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)} \right)\right]\right\}\right\}.$$

Since the optimization over $\{\gamma_x\}_x$ is separable for each x, by defining

$$v_x(\lambda) := \sup_{\gamma_x\in\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{z\sim\gamma_x}\left[f(z) - \lambda c(x,z) - \lambda\epsilon \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)} \right)\right]\right\},$$

and swap the supremum and the integration, we obtain

$$V \leq \inf_{\lambda\geq 0} \left\{ \lambda\rho + \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[v_x(\lambda)\Big]\right\}. \tag{7}$$

Since $v_x(\lambda)$ is an entropic regularized linear optimization, by Lemma EC.2, when there exists $\lambda > 0$ such that Condition 1 is satisfied, it holds that $\mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))/(\lambda\epsilon)}\big] < \infty$, $v_x(\lambda)$ is measurable with respect to $x\sim\widehat{\mathbb{P}}$ and

$$v_x(\lambda) = \lambda\epsilon \log \mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))/(\lambda\epsilon)}\big] < \infty.$$

In this case, the desired result holds. If for any $\lambda > 0$,

$$\widehat{\mathbb{P}}\left\{x :\ \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right] = \infty\right\} = \widehat{\mathbb{P}}\left\{x :\ \mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))/(\lambda\epsilon)}\big] = \infty\right\} > 0,$$

then intermediately we obtain

$$V \leq \inf_{\lambda\geq 0} \left\{ \lambda\rho + \lambda\epsilon\, \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\log \mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))/(\lambda\epsilon)}\big]\Big] \right\} = \infty,$$

and the weak duality still holds. □

Next, we show the feasibility result in Theorem 1(I). The key observation is that the primal problem (Primal) can be reformulated as a generalized KL-divergence DRO problem.

LEMMA 2 **(Reformulation of (Primal)).** *Under Assumption 1, it holds that*

$$V = \sup_{\{\gamma_x\}_{x\in\operatorname{supp}\widehat{\mathbb{P}}}\subset\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}[f(z)] :\ \epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}\left[\log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)}\right)\right] \le \overline{\rho} \right\}.$$

Due to Lemma 2, Theorem 1(I) holds because of the non-negativity of KL-divergence.

Finally, we develop the strong duality. In the following, we provide the proof of the first part of Theorem 1(III) for the most representative case where $\overline{\rho} > 0$, the dual minimizer λ^* exists with $\lambda^* > 0$, and Condition 1 holds. Proofs of other cases are moved in Appendix EC.5. Below we develop the optimality condition when the dual minimizer $\lambda^* > 0$ in Lemma 3, by simply setting the derivative of the dual objective function to be zero.

LEMMA 3 **(First-order Optimality Condition when $\lambda^* > 0$).** *Suppose $\overline{\rho} > 0$ and Condition 1 is satisfied, and assume further that the dual minimizer $\lambda^* > 0$, then λ^* satisfies*

$$\overline{\rho} + \epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda^*\epsilon)}\right]\right] = \frac{1}{\lambda^*}\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\frac{\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)} f(z)\right]}{\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)}\right]}\right]. \tag{8}$$

We prove the strong duality by constructing the worst-case distribution.

Proof of Theorem 1(III) for the case where Condition 1 holds and $\overline{\rho} > 0, \lambda^* > 0$. We take the transport mapping γ_* such that

$$\frac{\mathrm{d}\gamma_*(x,z)}{\mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}\nu(z)} = \alpha_x \cdot \exp\Big(\big(f(z) - \lambda^* c(x,z)\big)/(\lambda^*\epsilon)\Big),$$

and $\alpha_x := \left(\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)}\right]\right)^{-1}$ is a normalizing constant such that $\mathrm{Proj}_{1\#}\gamma_* = \widehat{\mathbb{P}}$. Also define the primal (approximate) optimal distribution $\mathbb{P}_* := \mathrm{Proj}_{2\#}\gamma_*$. Recall the expression of the Sinkhorn distance in Definition 1, one can verify that

$$\begin{aligned}
&\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P}_*) = \inf_{\gamma\in\Gamma(\widehat{\mathbb{P}},\mathbb{P}_*)} \left\{ \mathbb{E}_{(x,z)\sim\gamma}\left[c(x,z) + \epsilon\log\left(\frac{\mathrm{d}\gamma(x,z)}{\mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}\nu(z)}\right)\right]\right\} \\
&= \inf_{\gamma\in\Gamma(\widehat{\mathbb{P}},\mathbb{P}_*)} \left\{ \mathbb{E}_{(x,z)\sim\gamma}\left[\epsilon\log\left(\frac{e^{c(x,z)/\epsilon}\,\mathrm{d}\gamma(x,z)}{\mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}\nu(z)}\right)\right]\right\} \\
&\le \mathbb{E}_{(x,z)\sim\gamma_*}\left[\epsilon\log\left(\frac{e^{c(x,z)/\epsilon}\,\mathrm{d}\gamma_*(x,z)}{\mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}\nu(z)}\right)\right] = \mathbb{E}_{(x,z)\sim\gamma_*}\left\{\frac{1}{\lambda^*}f(z) + \epsilon\log(\alpha_x)\right\} \\
&= \frac{1}{\lambda^*}\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\frac{\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)} f(z)\right]}{\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)}\right]}\right] - \epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)}\right]\right]
\end{aligned}$$

where the inequality relation is because γ_* is a feasible solution in $\Gamma(\widehat{\mathbb{P}}, \mathbb{P}_*)$, and the last two relations are by substituting the expression of γ_*. Since $\overline{\rho} > 0$ and the dual minimizer $\lambda^* > 0$, the optimality condition in (8) holds, which implies that $\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P}_*) \le \rho$, i.e., the distribution $\mathbb{P}_*$ is primal feasible for the problem (Primal). Moreover, we can see that the primal optimal value is lower bounded by the dual optimal value:

$$\begin{aligned}
V &\ge \mathbb{E}_{\mathbb{P}_*}[f(z)] = \mathbb{E}_{(x,z)\sim\gamma_*}[f(z)] \\
&= \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\nu}\left[f(z)\left(\frac{\mathrm{d}\gamma_*(x,z)}{\mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}\nu(z)}\right)\right] = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\frac{\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)} f(z)\right]}{\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)}\right]}\right] \\
&= \lambda^*\left(\overline{\rho} + \epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda^*\epsilon)}\right]\right]\right) = V_{\mathrm{D}},
\end{aligned}$$

where the third equality is by substituting the expression of γ_*, and the last equality is based on the optimality condition in (8). This, together with the weak duality, completes the proof. □

4. Efficient First-order Algorithm for Sinkhorn Robust Optimization

Consider the Sinkhorn robust optimization problem

$$\inf_{\theta\in\Theta} \sup_{\mathbb{P}\in\mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})} \mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)]. \tag{9}$$

Here the feasible set $\Theta \subseteq \mathbb{R}^{d_\theta}$ is *closed and convex* containing all possible candidates of decision vector θ, and the Sinkhorn uncertainty set is centered around a given nominal distribution $\widehat{\mathbb{P}}$. Based on our strong dual expression (Dual), we reformulate (9) as

$$\inf_{\lambda\geq 0} \left\{ \lambda\overline{\rho} + \inf_{\theta\in\Theta} \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_\theta(z)/(\lambda\epsilon)} \right] \right] \right\}, \tag{D}$$

where the constant $\overline{\rho}$ and the distribution $\mathbb{Q}_{x,\epsilon}$ are defined in (2) and (3), respectively. In Example 1 and 2, we have seen special instances of (D) where we can get closed-form expressions for the above integration. For general loss functions when a closed-form expression is not available, we resort to an algorithmic approach, which is the main development in this section.

A typical and efficient approach for solving a stochastic optimization problem is the stochastic gradient descent method such as stochastic mirror descent [82]. However, unlike many other stochastic optimization problems, one salient feature of (D) is that the objective function of (10) involves a nonlinear transformation of the expectation. Consequently, based on a batch of simulated samples from $\mathbb{Q}_{x,\epsilon}$, an unbiased gradient estimate could be challenging to obtain. In Section 4.1, we will present a biased stochastic mirror descent algorithm to solve this problem. By properly balancing their bias and second-order moment trade-off in the biased gradient estimator, our algorithm can efficiently find a near-optimal decision of (10), as we will analyze in Section 4.2.

4.1. Main Algorithm

Define the objective value of the outer minimization in (D) as

$$F^*(\lambda) := \lambda\overline{\rho} + \inf_{\theta\in\Theta} \left\{ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_\theta(z)/(\lambda\epsilon)} \right] \right] \right\}. \tag{10}$$

We first present a Biased Stochastic Mirror Descent (BSMD) algorithm that solves the minimization over the decisions θ for a given Lagrangian multiplier λ in Section 4.1.1. Next, in Section 4.1.2 we present a bisection search algorithm for finding the optimal Lagrangian multiplier.

4.1.1. Biased Stochastic Mirror Descent Since the multiplier λ is fixed in (10), we omit the dependence of λ when defining objective or gradient terms in this subsection and set

$$F(\theta) := \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_\theta(z)/(\lambda\epsilon)} \right] \right]. \tag{11}$$

We first introduce several notations that are standard in the mirror descent algorithm. Let $\omega:\ \Theta\to\mathbb{R}$ be a distance generating function that is continuously differentiable and κ-strongly convex on Θ with respect to norm $\|\cdot\|$. This induces the Bregman divergence $D_\omega(\theta,\theta'):\ \Theta\times\Theta\to\mathbb{R}_+$:

$$D_\omega(\theta,\theta') = \omega(\theta') - \omega(\theta) - \langle\nabla\omega(\theta), \theta'-\theta\rangle.$$

We define the *prox-mapping* $\mathrm{Prox}:\ \mathbb{R}^{d_\theta}\to\Theta$ as

$$\mathrm{Prox}_\theta(y) = \arg\min_{\theta'\in\Theta} \left\{ \langle y, \theta'-\theta\rangle + D_\omega(\theta,\theta') \right\}.$$

Algorithm 1 BSMD for finding the optimal solution of (10) while fixing λ

Require: Maximal iteration T_{in}, constant step size γ, initial guess θ_1, fixed multiplier λ.

1: **for** $t = 0, 1, \ldots, T_{\text{in}} - 1$ **do**

2: Formulate a (biased) gradient estimate $v(\theta_t)$ of $F(\theta_t)$ using (13).

3: Update $\theta_{t+1} = \text{Prox}_{\theta_t}\big(\gamma v(\theta_t)\big)$.

4: **end for**

Output the estimate of optimal solution $\widehat{\theta}_i = \frac{1}{T_{\text{in}}} \sum_{t=1}^{T_{\text{in}}} \theta_t$.

With these notations in hand, we present our algorithm in Algorithm 1.

Algorithm 1 iteratively obtains a biased stochastic gradient estimate and performs a proximal gradient update. In Step 2 of Algorithm 1, we adopt the idea of multi-level Monte-Carlo (MLMC) simulation [60] to generate biased gradient estimates with controlled bias. To this end, we first construct a function $F^\ell(\theta)$ that approximate the original objective function $F(\theta)$ with $\mathcal{O}(2^{-\ell})$-gap:

$$F^\ell(\theta) = \mathbb{E}_{x^\ell \sim \widehat{\mathbb{P}}} \, \mathbb{E}_{\{z_j^\ell\}_{j\in[2^\ell]} \sim \mathbb{Q}_{x^\ell,\epsilon}} \left[\lambda\epsilon \log\Big(\frac{1}{2^\ell} \sum_{j\in[2^\ell]} e^{f_\theta(z_j^\ell)/(\lambda\epsilon)} \Big) \right], \tag{12}$$

where the random variable x^ℓ follows distribution $\widehat{\mathbb{P}}$, and given a realization of x^ℓ, $\{z_j^\ell\}_{j\in[2^\ell]}$ are independent and identically distributed samples from $\mathbb{Q}_{x^\ell,\epsilon}$. Unlike the original objective $F(\theta)$, unbiased gradient estimators of its approximation $F^\ell(\theta)$ can be easily obtained. Denote by $\zeta^\ell = (x^\ell, \{z_j^\ell\}_{j\in[2^\ell]})$ the collection of random sampling parameters, and

$$U_{n_1:n_2}(\theta, \zeta^\ell) := \lambda\epsilon \log\Big(\frac{1}{n_2 - n_1 + 1} \sum_{j\in[n_1:n_2]} e^{f_\theta(z_j^\ell)/(\lambda\epsilon)} \Big).$$

For a fixed parameter θ, we define

$$\begin{aligned} g^\ell(\theta, \zeta^\ell) &:= \nabla_\theta U_{1:2^\ell}(\theta, \zeta^\ell), \\ G^\ell(\theta, \zeta^\ell) &:= \nabla_\theta \left[U_{1:2^\ell}(\theta, \zeta^\ell) - \frac{1}{2} U_{1:2^{\ell-1}}(\theta, \zeta^\ell) - \frac{1}{2} U_{2^{\ell-1}+1:2^\ell}(\theta, \zeta^\ell) \right]. \end{aligned}$$

The random vector $g^\ell(\theta, \zeta^\ell)$ is an unbiased estimator of $\nabla F^\ell(\theta)$, while the vector $G^\ell(\theta, \zeta^\ell)$ is an unbiased estimator of $\nabla F^\ell(\theta) - \nabla F^{\ell-1}(\theta)$. The latter facilitates the reduction of second-order moment of the gradient estimator thanks to common random numbers. More specifically, we list the following choices of gradient estimators at a point θ:

– *Stochastic Gradient (SG) Estimator:* Get n_L° random vectors $\{g^L(\theta, \zeta_i^L)\}_{i=1}^{n_L^\circ}$, where $\{\zeta_i^L\}_i$ are i.i.d. copies of ζ^ℓ. Construct

$$v^{\text{SG}}(\theta) = \frac{1}{n_L^\circ} \sum_{i=1}^{n_L^\circ} g^L(\theta, \zeta_i^L). \tag{13a}$$

– *Randomized Truncation MLMC (RT-MLMC) Estimator [22]:* Sample n_L° random levels $\iota_1, \ldots, \iota_{n_L^\circ}$ i.i.d. from a distribution $\mathbb{P}(\iota = \ell) = \frac{2^{-\ell}}{2 - 2^{-L}}, \ell = 0, 1, \ldots, L$. Construct

$$v^{\text{RT-MLMC}}(\theta) = \frac{1}{n_L^\circ} \sum_{i=1}^{n_L^\circ} \frac{1}{\mathbb{P}(\iota = \iota_i)} G^{\iota_i}(\theta, \zeta^{\iota_i}). \tag{13b}$$

Remark 7 (Sampling from $\mathbb{Q}_{x,\epsilon}$). In many cases, generating samples from $\mathbb{Q}_{x,\epsilon}$ is easy. For example, when the transport cost $c(\cdot,\cdot) = \frac{1}{2}\|\cdot - \cdot\|_2^2$ and $\mathcal{Z} = \mathbb{R}^d$, then the distribution $\mathbb{Q}_{x,\epsilon}$ becomes a Gaussian distribution $\mathcal{N}(x, \epsilon I_d)$. When the transport cost $c(\cdot,\cdot)$ is decomposable in each coordinate, we can

apply the acceptance-rejection method [5] to generate samples in each coordinate independently, the complexity of which only increases linearly in the data dimension. When the transport cost $c(x,y)=\frac{1}{q}\|x-y\|_p^q$, the complexity of sampling based on Lagenvin Monte Carlo method for obtaining a τ-close sample point is $\mathcal{O}(d/\tau)$. See the detailed algorithm of sampling in Appendix EC.7.2. ♣

4.1.2. Bisection Search Bisection search requires an oracle that produces an objective estimator of (D). We describe it in Algorithm 2. For a fixed multiplier λ, it solves problem (10) using Algorithm 1 and then estimates the corresponding objective value. It has m repetitions, whose value will be determined later in Section 4.2 to achieve the best sample complexity.

Algorithm 2 Evaluating the objective value of (D)

Require: Fixed multiplier λ, error tolerance δ, batch size m.
1: **for** $j=1,2,\ldots,m$ **do**
2: Obtain a δ-optimal solution $\widehat{\theta}_j$ of problem (10) using Algorithm 1.
3: Estimate the objective in (11) using RT-MLMC estimator in (13b), denoted as $\widehat{F}(\widehat{\theta}_j;\lambda)$.
4: **end for**
Output $\widehat{F}^*(\lambda):=\lambda\overline{\rho}+\min_{j\in[m]}\ \widehat{F}(\widehat{\theta}_j;\lambda)$.

Below, we present the RT-MLMC-based objective estimator of (11) that is used in Step 3 of Algorithm 2. For notational simplicity, we define

$$A^\ell(\theta,\zeta^\ell)=U_{1:2^\ell}(\theta,\zeta^\ell)-\frac{1}{2}U_{1:2^{\ell-1}}(\theta,\zeta^\ell)-\frac{1}{2}U_{2^{\ell-1}+1:2^\ell}(\theta,\zeta^\ell).$$

It is worth noting that [59, Theorem 4.1] proposed and analyzed the SG estimator for estimating the objective value of CSO problems. Its sample complexity is $\mathcal{O}(\delta^{-3})$, while our proposed RT-MLMC-based estimator has improved sample complexity $\widetilde{\mathcal{O}}(\delta^{-2})$ (see a formal statement in Proposition 1).
– *Randomized Truncation MLMC (RT-MLMC) Estimator:* Sample n_L° random levels $\iota_1,\ldots,\iota_{n_L^\circ}$ i.i.d. from a distribution $\mathbb{P}(\iota=\ell)=\frac{2^{-\ell}}{2-2^{-L}},\ell=0,1,\ldots,L$. Construct

$$\widehat{F}^{\text{RT-MLMC}}(\theta)=\frac{1}{n_L^\circ}\sum_{i=1}^{n_L^\circ}\frac{1}{\mathbb{P}(\iota=\iota_i)}A^{\iota_i}(\theta,\zeta^{\iota_i}). \tag{14}$$

Given an inexact objective oracle of (D), we use bisection search to find a near-optimal multiplier in (D); see Algorithm 3.

Another approach to solving (D) is to jointly optimize (λ,θ). However, as mentioned in [79], the variance of the gradient estimate of the objective function with respect to λ becomes unstable when λ is small. Consequently, in our algorithm, we have developed a bisection method to update λ instead.

We also remark that, as a practical alternative, one can solve (10) using Algorithm 1 alone and tune the hyperparameter λ, as tuning the radius $\overline{\rho}$ is equivalent to tuning the Lagrangian multiplier λ in (10). This corresponds to the Sinkhorn robust learning problem with a soft Sinkhorn constraint.

4.2. Convergence Analysis

In this subsection, we analyze the convergence properties of the proposed algorithms. We begin with the following standard assumptions on the loss function f_θ:

ASSUMPTION 2. (I) *(Convexity): The loss function $f_\theta(z)$ is convex in θ.*
(II) *(Boundedness): The loss function $f_\theta(z)$ satisfies $0\le f_\theta(z)\le B$ for any $\theta\in\Theta$ and $z\in\mathcal{Z}$.*
(III) *(Lipschitz Continuity): For fixed z and θ_1,θ_2, it holds that $|f_{\theta_1}(z)-f_{\theta_2}(z)|\le L_f\|\theta_1-\theta_2\|_2$.*

In the following analysis, the sample complexity of an algorithm is defined as the number of samples (x,z), where $x\sim\widehat{\mathbb{P}}$ and $z\sim\mathbb{Q}_x$, in total used by the algorithm, and the storage cost is defined the number of samples stored in the memory.

Algorithm 3 Bisection search for finding the optimal multiplier of (D)

Require: Interval $[\lambda_l, \lambda_u]$, maximum outer iterations T_{out}

$\lambda^{(0)} = \lambda_l$, $\chi_l^{(0)} = \lambda_l, \chi_u^{(0)} = \lambda_u$.
for $t = 1, \ldots, T_{\text{out}}$ **do**
  Update $\beta_l^{(t)} = \frac{1}{3}\left[2\chi_l^{(t-1)} + \chi_u^{(t-1)}\right]$ and $\beta_u^{(t)} = \frac{1}{3}\left[\chi_l^{(t-1)} + 2\chi_u^{(t-1)}\right]$.
  Query Algorithm 2 to obtain objective estimators of $\beta_l^{(t)}, \beta_u^{(t)}$, denoted as $\widehat{F}^*(\beta_l^{(t)}), \widehat{F}^*(\beta_u^{(t)})$, respectively.
  if $\widehat{F}^*(\beta_l^{(t)}) \le \widehat{F}^*(\beta_u^{(t)})$ **then**
    Update $(\chi_l^{(t)}, \chi_u^{(t)}) = (\chi_l^{(t-1)}, \beta_u^{(t)})$.
    if $\widehat{F}^*(\beta_l^{(t)}) \le \widehat{F}^*(\lambda^{(t-1)})$, update $\lambda^{(t)} = \beta_l^{(t)}$; **else** update $\lambda^{(t)} = \lambda^{(t-1)}$.
  else
    Update $(\chi_l^{(t)}, \chi_u^{(t)}) = (\beta_l^{(t)}, \chi_u^{(t-1)})$.
    if $\widehat{F}^*(\beta_u^{(t)}) \le \widehat{F}^*(\lambda^{(t-1)})$, update $\lambda^{(t)} = \beta_u^{(t)}$; **else** update $\lambda^{(t)} = \lambda^{(t-1)}$.
  end if
end for
Output $\lambda^{(T_{out})}$.

4.2.1. Complexity of Biased Stochastic Mirror Descent In this part, we discuss the complexity of Algorithm 1. We say θ is a δ-optimal solution if $\mathbb{E}[F(\theta)] - F(\theta^*) \le \delta$, where θ^* is the optimal solution of (10). By properly tuning hyper-parameters to balance the trade-off between bias and second-order moment of the gradient estimate, we establish its performance guarantees in Theorem 2. The explicit constants and detailed proof can be found in Appendix EC.7.

THEOREM 2. *Under Assumptions 2(I), 2(II), and 2(III), with properly chosen hyper-parameters as in Table 2, the following results hold:*

(I) *The sample complexity of Algorithm 1 using SG scheme (13a) to obtain a δ-optimal solution is $\mathcal{O}(\delta^{-3})$, with storage cost $\mathcal{O}(\delta^{-1})$.*

(II) *The sample complexity of Algorithm 1 using RT-MLMC scheme (13b) to obtain a δ-optimal solution is $\widetilde{\mathcal{O}}(\delta^{-2})$, with storage cost $\widetilde{\mathcal{O}}(1)$.*

Table 2 Hyper-parameters, sample complexity (Samp.), and storage cost (Stor.) of Algorithm 1.

Estimators	Hyper-parameters	Samp./Stor.
SG	$L = \mathcal{O}(\log \frac{1}{\delta}), \quad T_{\text{in}} = \mathcal{O}(\delta^{-2})$ $n_L^{\circ} = \mathcal{O}(1), \quad \gamma = \mathcal{O}(\delta)$	Samp. $= \mathcal{O}(T_{\text{in}} n_L^{\circ} 2^L) = \mathcal{O}(\delta^{-3})$ Stor. $= \mathcal{O}(n_L^{\circ} 2^L) = \mathcal{O}(\delta^{-1})$
RT-MLMC	$L = \mathcal{O}(\log \frac{1}{\delta}), \quad T_{\text{in}} = \widetilde{\mathcal{O}}(\delta^{-2})$ $n_L^{\circ} = \mathcal{O}(1), \quad \gamma = \widetilde{\mathcal{O}}(\delta)$	Samp. $= \mathcal{O}(T_{\text{in}}(n_L^{\circ} L)) = \widetilde{\mathcal{O}}(\delta^{-2})$ Stor. $= \mathcal{O}(n_L^{\circ} L) = \widetilde{\mathcal{O}}(1)$

Theorem 2 indicates that the sample complexity of BSMD using SG scheme for solving (10) is $\mathcal{O}(\delta^{-3})$, and can be further improved to $\widetilde{\mathcal{O}}(\delta^{-2})$ using the RT-MLMC gradient estimator. This complexity is near-optimal, which matches the information-theoretic lower bound for general convex optimization problems [82]. Besides, using RT-MLMC gradient estimator instead of SG estimator achieves cheaper storage cost of $\widetilde{\mathcal{O}}(1)$, which is nearly error tolerance-independent.

We remark that the problem (10) can be viewed as a special conditional stochastic optimization (CSO) problem [59, 61, 60]. In contrast to the previous state-of-the-art [60], which focused solely on unconstrained optimization where stochastic gradient descent is applied, we consider a constrained optimization on θ where stochastic mirror descent is applied. Furthermore, our approach achieves

the near-optimal sample complexity for both smooth and nonsmooth loss functions, distinguishing it from [61, Theorem 3.2], which shows sample complexity of $O(\delta^{-4})$ for nonsmooth loss functions. Our result can be easily extended to general CSO problems with general convex loss functions which, to the best of our knowledge, is the first algorithm for such problems with provably near-optimal sample complexity.

Remark 8 (Comparison with Biased Sample Average Approximation). Another way to solve (10) is to approximate the objective using finite samples for both expectations. This leads to a biased sample estimate, called Biased Sample Average Approximation (BSAA). Applying [59, Corollary 4.2], it can be shown that the total sample complexity and storage cost for BSAA are both $\widetilde{\mathcal{O}}(\delta^{-3})$ for convex, bounded and Lipschitz loss functions (see formal statement in Appendix EC.7.3). Our proposed BSMD with the RT-MLMC-based gradient estimator has smaller sample complexity and storage cost. Also, the BSAA method still requires computing the optimal solution of the approximated optimization problem as the output. Hence, it typically takes considerably less time and memory to run the BSMD step rather than solving for the BSAA formulation. ♣

4.2.2. Complexity of Bisection Search We first provide the complexity analysis for Algorithm 2, which produces an estimator of the objective value of the outer minimization in (D).

Proposition 1. *Let $\eta \in (0,1)$ and set the batch size $m = \lceil \log_2 \frac{2}{\eta} \rceil$. Assume Assumptions 2(I), 2(II), and 2(III) hold, and we choose hyper-parameters in Step 3 in Algorithm 2 as*

$$L = \mathcal{O}(\log \frac{1}{\delta}), \quad n_L^{\circ} = \widetilde{\mathcal{O}}\left(\frac{1}{\delta^2} \log \frac{1}{\alpha}\right).$$

Then the output in Algorithm 2 satisfies $|\widehat{F}^*(\lambda) - F^*(\lambda)| \le \delta$ with probability at least $1-\eta$. In addition,

(I) *The sample complexity of Algorithm 2 using SG gradient estimator (13a) in the BSMD step and RT-MLMC objective estimator (14) is $\mathcal{O}\left(\delta^{-3} \cdot \text{polylog}\frac{1}{\eta}\right)$, with storage cost $\widetilde{\mathcal{O}}\left(\delta^{-2} \cdot \text{polylog}\frac{1}{\eta}\right)$.*

(II) *The sample complexity of Algorithm 2 using RT-MLMC gradient estimator (13b) in the BSMD step and RT-MLMC objective estimator (14) is $\widetilde{\mathcal{O}}\left(\delta^{-2} \cdot \text{polylog}\frac{1}{\eta}\right)$, with storage cost $\widetilde{\mathcal{O}}\left(\delta^{-2} \cdot \text{polylog}\frac{1}{\eta}\right)$.*

Next, we provide the convergence analysis for Algorithm 3.

Theorem 3. *Let $\eta \in (0,1)$. Assume Assumptions 2(I), 2(II), and 2(III) hold and $0 < \lambda_l \le \lambda^* \le \lambda_u < \infty$. Specify hyper-parameters in Algorithm 3 as*

$$T_{out} = \left\lceil \log_{3/2} \frac{4L_\lambda(\lambda_u - \lambda_l)}{\delta} \right\rceil, \quad \eta' = \frac{\eta}{1+2T_{out}}, \quad L_\lambda = \overline{\rho} + \frac{B}{\lambda_l}\left[1 + e^{B/(\lambda_l \epsilon)}\right].$$

Suppose there exists an oracle $\widehat{F}^*$ such that for any $\lambda > 0$, it gives estimation of the optimal value in (10) up to accuracy $\delta/4$ with probability at least $1-\eta'$, then with probability at least $1-\eta$, Algorithm 3 finds the optimal multiplier up to accuracy δ (i.e., it finds λ such that $F^*(\lambda) - \min_{\lambda_l \le \lambda \le \lambda_u} F^*(\lambda) \le \delta$) by calling the inexact oracle $\widehat{F}^*$ for $\mathcal{O}(\log \frac{1}{\delta})$ times.

In practical implementations, λ_l and λ_u can be found numerically. For examples in Section 5.1 and 5.2, we set $\lambda_l = 0.5$ and $\lambda_u = 50$. Combining Proposition 1 and Theorem 3, the overall sample complexity for obtaining a δ-optimal solution of (D) with probability at least $1-\eta$ is $\widetilde{\mathcal{O}}\left(\delta^{-2} \cdot \text{polylog}\frac{1}{\eta}\right)$, with storage cost $\widetilde{\mathcal{O}}\left(\delta^{-2} \cdot \text{polylog}\frac{1}{\eta}\right)$.

Remark 9 (Comparison with Empirical Risk Minimization). The optimal complexity for obtaining a δ-optimal solution from the empirical risk minimization (ERM) formulation $\inf_{\theta \in \Theta} \mathbb{E}_{x \sim \widehat{\mathbb{P}}}[f_\theta(x)]$ with a convex loss function $f_\theta(z)$ (regardless of the smoothness assumption) is $\mathcal{O}(\delta^{-2})$ [82]. Therefore, the complexity for solving the Sinkhorn DRO model matches that for solving the ERM counterpart up to a near-constant factor. ♣

REMARK 10 (COMPARISON WITH WASSERESTEIN DRO). Recall from Table 1 that Wasserstein DRO is tractable for a restricted family of loss functions (Table 1). Specially, Wasserstein DRO with $\widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^{n}\delta_{\hat{x}_i}$ can be formulated as a minimax problem

$$\min_{\theta\in\Theta,\lambda\geq 0}\ \max_{z_i\in\mathbb{R}^d, i\in[n]}\ \lambda\rho + \frac{1}{n}\sum_{i=1}^{n}\big[f_\theta(z_i) - \lambda c(\hat{x}_i, z_i)\big].$$

When $f_\theta(z)$ is not concave in z, the above problem generally reduces to the convex-non-concave saddle point problem, whose global optimality is difficult to obtain. Even when the Wasserstein DRO formulation is tractable, its complexity generally has non-negligible dependence on the sample size n (see [48, Remark 9] and references therein for more discussions). In contrast, the complexity for solving the Sinkhorn DRO formulation is sample size-independent. ♣

5. Applications

In this section, we apply our methodology to three applications: the newsvendor model, mean-risk portfolio optimization, and adversarial classification. We compare our model with three benchmarks: (i) the classical sample average approximation (SAA) model; (ii) the Wasserstein DRO model; and (iii) the KL-divergence DRO model. We choose the transport cost $c(\cdot,\cdot) = \|\cdot - \cdot\|_1$ for 1-Wasserstein or 1-Sinkhorn DRO model, and $c(\cdot,\cdot) = \frac{1}{2}\|\cdot - \cdot\|_2^2$ for 2-Wasserstein or 2-Sinkhorn DRO model. Throughout this section, we take the reference measure ν in the Sinkhorn distance to be the Lebesgue measure. The hyper-parameters are selected using 5-fold cross-validation. To generate the results, we run each experiment repeatedly with 200 independent trials. Further implementation details and additional experiments are included in Appendices EC.1 and EC.2, respectively.

5.1. Newsvendor Model

Consider the following distributionally robust newsvendor model:

$$\min_{\theta}\ \max_{\mathbb{P}\in\mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})}\ \mathbb{E}_{z\sim\mathbb{P}}\big[k\theta - u\min(\theta, z)\big],$$

where the random variable z stands for the random demand, whose empirical distribution $\widehat{\mathbb{P}}$ consists of n independent samples from the underlying data distribution; the decision variable θ represents the inventory level; and $k = 5, u = 7$ are constants corresponding to overage and underage costs, respectively.

In this experiment, we examine the performance of DRO models for various sample sizes $n \in \{10, 30, 100\}$ and under three different types of data distribution: (i) the exponential distribution with rate parameter 1, (ii) the gamma distribution with shape parameter 2 and scale parameter 1.5, (iii) the equiprobable mixture of two truncated normal distributions $\mathcal{N}(\mu = 1, \sigma = 1, a = 0, b = 10)$ and $\mathcal{N}(\mu = 6, \sigma = 1, a = 0, b = 10)$. In particular, we do not report the performance for 1-Wasserstein DRO model in this example, because it is identical to the SAA approach [78, Remark 6.7]. In addition, since 2-Wasserstein DRO is computationally intractable for this example, we solve the corresponding formulation by discretizing the support of the distributions.

We measure the out-of-sample performance of a solution θ based on training dataset $\mathcal{D}$ using the percentage of improvement (a.k.a., coefficient of prescriptiveness) in [11]:

$$\text{Prescriptiveness}(\theta) = 1 - \frac{J(\theta) - J^*}{J(\theta^{\text{SAA}}_{\mathcal{D}}) - J^*}, \tag{15}$$

where J^* denotes the true optimal value when the true distribution is known exactly, $\theta^{\text{SAA}}_{\mathcal{D}}$ denotes the decision from the SAA approach with dataset $\mathcal{D}$, and $J(\theta)$ denotes the expected loss of the solution θ under the true distribution, estimated through an SAA objective value with 10^5 testing samples. Thus, the higher this coefficient is, the better the solution's out-of-sample performance.

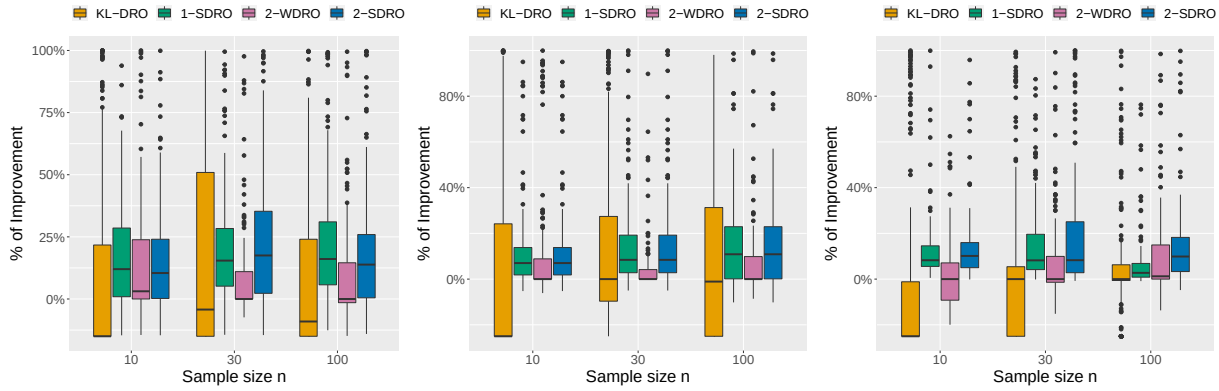

Figure 2 Out-of-sample performance for the newsvendor model with parameters $s \in \{0.25, 0.5, 0.75, 1, 2, 4\}$ and fixed sample size $n = 20$. The y-axis corresponds to the coefficent of prescriptiveness defined in (15). For figures from left to right, we specify the data distribution as exponential distribution, gamma distribution, and equiprobable mixture of two truncated normal distributions, respectively.

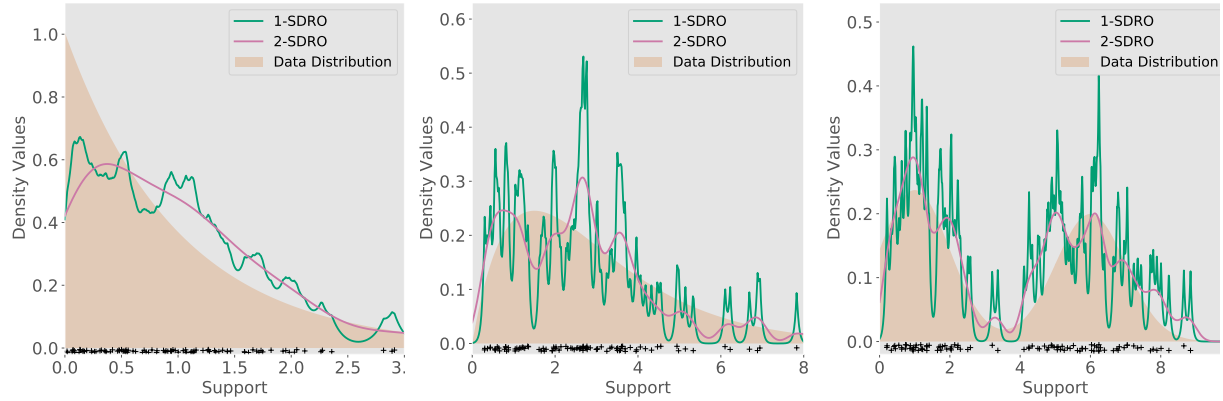

Figure 3 Plots for the density of worst-case distributions generated by the 1-SDRO or 2-SDRO model. The black dots in the bottom of figures indicate the sample points from the training dataset of sample size $n = 100$. For figures from left to right, we specify the data distribution as exponential distribution, gamma distribution, and equiprobable mixture of two truncated normal distributions, respectively.

We report the box-plots of the coefficients of prescriptiveness in Fig. 2 using 200 independent trials. We find that either 1-SDRO or 2-SDRO model achieve the best out-of-sample performance over all sample sizes and data distributions listed, as it consistently scores higher than other benchmarks in the box plots. In contrast, the KL-DRO model does not achieve satisfactory performance, and sometimes even underperforms the SAA model. While the 2-WDRO model demonstrates some improvement over the SAA model, the 2-SDRO model shows more clear improvement. Additionally, we plot the density of worst-case distributions for 1-SDRO or 2-SDRO model in Fig. 3. When specifying the data distribution as exponential, gamma, or Gaussian mixture, the corresponding worst-case distributions capture the shape of the ground truth distribution reasonably well, which partly explains why the Sinkhorn DRO model achieves superior performance when the data distribution is absolutely continuous.

5.2. Mean-risk Portfolio Optimization

Consider the following distributionally robust mean-risk portfolio optimization problem

$$\begin{aligned} \min_{\theta} \max_{\mathbb{P} \in \mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})} \quad & \mathbb{E}_{z\sim\mathbb{P}}[-\theta^{\mathrm{T}} z] + \varrho \cdot \mathbb{P}\text{-CVaR}_{\alpha}(-\theta^{\mathrm{T}} z) \\ \text{s.t.} \quad & \theta \in \Theta = \{\theta \in \mathbb{R}^d_+ : \theta^{\mathrm{T}} 1 = 1\}, \end{aligned}$$

where the random vector $z \in \mathbb{R}^d$ stands for the returns of assets; the decision variable $\theta \in \Theta$ represents the portfolio strategy that invests a certain percentage θ_i of the available capital in the i-th asset; and the term $\mathbb{P}$-CVaR$_\alpha(-\theta^{\mathrm{T}} z)$ quantifies conditional value-at-risk [98], i.e., the average of the $\alpha \times 100\%$ worst portfolio losses under the distribution $\mathbb{P}$. We follow a similar setup as in [78]. Specifically, we set $\alpha = 0.2, \varrho = 10$. The underlying true random return can be decomposed into a systematic risk factor $\psi \in \mathbb{R}$ and idiosyncratic risk factors $\epsilon \in \mathbb{R}^d$:

$$z_i = \psi + \epsilon_i, \quad i = 1, 2, \ldots, d,$$

where $\psi \sim \mathcal{N}(0, 0.02)$ and $\epsilon_i \sim \mathcal{N}(i \times 0.03, i \times 0.025)$. When solving the Sinkhorn DRO formulation, we take the Bregman divergence D_ω as the KL-divergence when performing BSMD algorithm in Algorithm 1, allowing for efficient implementation [82]. We quantify the performance of a given solution using the same criterion defined in Section 5.1 and generate box plots using 200 independent trials. Fig. 4a) reports the scenario where the data dimension $d = 30$ is fixed and sample size $n \in \{30, 50, 100, 150, 200, 400\}$, and Fig. 4b) reports the scenario where the sample size $n = 100$ is fixed and the number of assets $d \in \{5, 10, 20, 40, 80, 100\}$. We find that the KL-DRO model does not have competitive performance compared to other DRO models, especially as the data dimension d increases. This is because the ambiguity set of KL-DRO model only takes into account those distributions sharing the same support as the nominal distribution, which seems to be restrictive, especially for high-dimensional settings. Moreover, while 1-WDRO or 2-WDRO model has better out-of-sample performance than the SAA model, the corresponding 1-SDRO or 2-SDRO model has clearer improvements, as it consistently scores higher in the box plots.

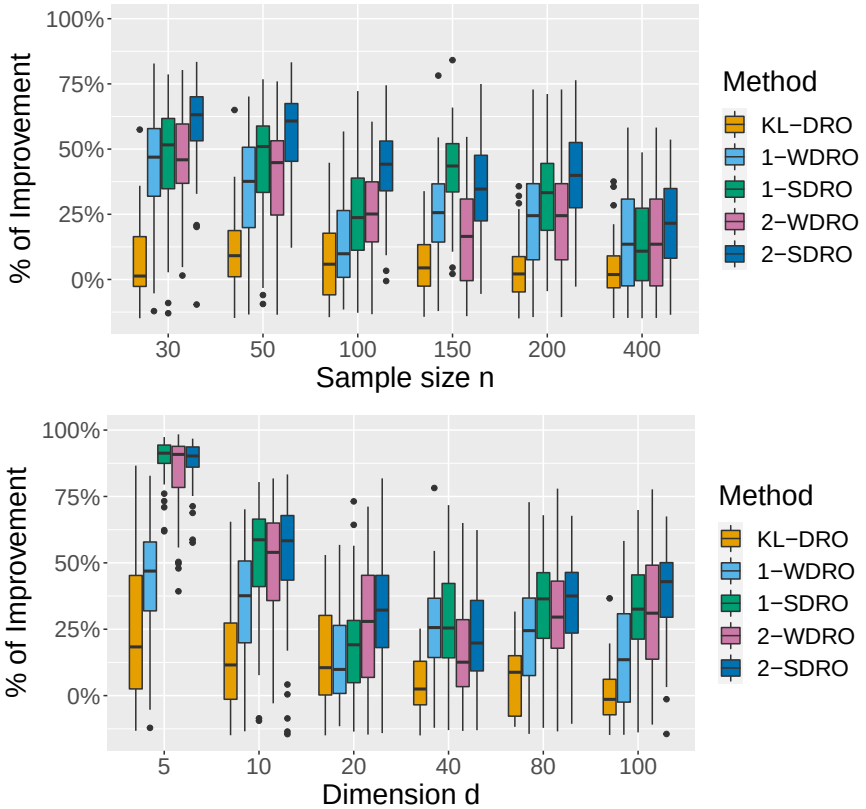

Figure 4 Out-of-sample performance for the portfolio optimization problem. Upper: fixing data dimension $d = 30$ and varying sample size $n \in \{30, 50, 100, 150, 200, 400\}$. Bottom: fixing sample size $n = 100$ and varying data dimension $d \in \{5, 10, 20, 40, 80, 100\}$.

5.3. Adversarial Multi-class Logistic Regression

The adversarial attack is an emerging topic in artificial intelligence: small perturbations to the data-generating distribution can cause well-trained machine learning models to produce unexpectedly inaccurate predictions [55]. For real applications involving high-stake environments, such as self-driving and automated detection of tumors, it is critical to deploy models which are robust to potential adversarial attacks. Many approaches for adversarial attack and defense have been proposed in literature [86, 87, 88, 99, 109, 58, 76, 116]. Among those approaches, DRO is a promising training procedure that guarantees certifiable adversarial robustness [109]. In this subsection, we consider an adversarial training model for multi-class logistic regression. Given a feature vector $x \in \mathbb{R}^d$ and its label $y \in [C]$, we denote $\boldsymbol{y} \in \{0,1\}^C$ as the corresponding one-hot label vector, and define the following negative likelihood loss:

$$h_B(x, \boldsymbol{y}) = -\boldsymbol{y}^{\mathrm{T}} B^{\mathrm{T}} x + \log\left(1^{\mathrm{T}} e^{B^{\mathrm{T}} x}\right),$$

where $B := [w_1, \ldots, w_K]$ stands for the linear classifier. Let $\widehat{\mathbb{P}}$ be the empirical distribution from training samples. Since the testing samples may have slightly different data distributions than the training samples, the DRO model aims to solve the following optimization problem to mitigate the impact of adversarial attacks:

$$\min_{B} \max_{\mathbb{P} \in \mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})} \mathbb{E}_{(x,\boldsymbol{y})\sim\mathbb{P}}\big[h_B(x, \boldsymbol{y})\big].$$

It is assumed that only the feature vector x has uncertainty but not the label $\boldsymbol{y}$.

We conduct experiments on four large-scale datasets: CIFAR10 [65], CIFAR100 [65], STL10 [30], and tinyImageNet [115]. We pre-process these datasets using the ResNet-18 network [57] pre-trained on the ImageNet dataset to extract linear features. Since this network has learned a rich set of hierarchical features from the large and diverse ImageNet dataset, it typically extracts useful features for other image datasets. We then add various perturbations to the testing datasets, such as *ℓ_1-norm and ℓ_2-norm adversarial projected gradient method attacks [76]*, *white Laplace noise*, and *white Gaussian noise*, to evaluate the performance of the DRO models. See the detailed procedure for generating adversarial perturbations and statistics on pre-processed datasets in Appendix EC.1.3. We use the mis-classification rate on testing dataset to quantify the performance for the obtained classifers.

For baseline DRO models, we solve their penalized formulations and tune the Lagrangian multiplier to obtain the best performance. However, solving the penalized 2-WDRO model is tractable only when the multiplier is sufficiently large [109], and all its maximization subproblems involve concave objective functions, which may not be the case for practical experiments. Additionally, the penalized 1-WDRO model is a convex-non-concave minimax game in high dimensions, making it intractable to solve into global optimality, with the exception that when the number of classes $C = 2$, it reduces to the tractable distributionally robust logistic regression model [104]. To approximately solve these WDRO formulations, we try gradient descent ascent heuristics, which are inspired from [109, Algorithm 1] (see Algorithm 4 in Appendix EC.1.3), but they are computationally inefficient and do not yield satisfactory classifiers. To make a fair comparison, we solve the soft formulation of KL-DRO, 1-SDRO, or 2-SDRO in this subsection and tune the hyper-parameter λ.

Figure 5 presents the classification results for different types of adversarial attacks with varying levels of perturbations on the datasets. We observe that as the level of perturbations on the testing samples increases, all methods tend to perform worse. However, both the 1-SDRO and 2-SDRO models show a slower trend of increasing error rates than other benchmarks across all types of adversarial attacks and all datasets. This suggests that SDRO models can be a suitable choice for adversarial robust training.

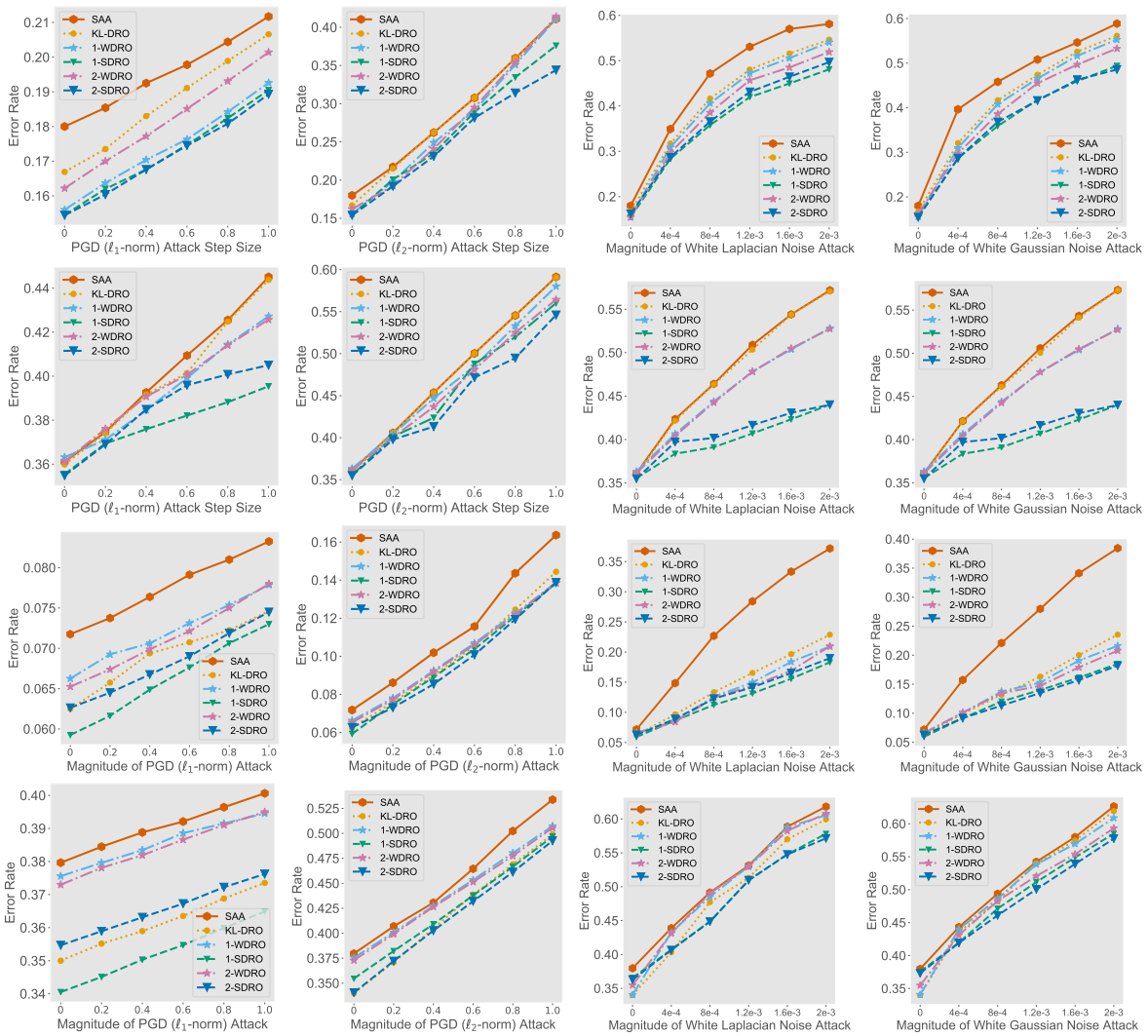

Figure 5 Results of adversarial training on various image datasets with different types of adversarial attack. From left to right, the figures correspond to (a) ℓ_1-norm PGD attack; (b) ℓ_2-norm PGD attack; (c) white Laplacian noise attack; and (d) white Gaussian noise attack. From top to bottom, the figures correspond to (a) CIFAR10 dataset; (b) CIFAR100 dataset; (c) STL10 dataset; and (d) tinyImageNet dataset.

6. Concluding Remarks

In this paper, we investigated a new distributionally robust optimization framework based on the Sinkhorn distance. By developing a strong dual reformulation and a biased stochastic mirror descent algorithm, we have shown that the resulting problem is tractable under mild assumptions, which significantly spans the tractability of Wasserstein DRO. Analysis of the worst-case distribution indicates that Sinkhorn DRO hedges a more reasonable set of adverse scenarios and is thus less conservative than Wasserstein DRO. Extensive numerical experiments demonstrated that Sinkhorn DRO is a promising candidate for modeling distributional ambiguities in decision-making under uncertainty.

In the meantime, several topics worth investigating are left for future work. For example, one important research question is the statistical performance guarantees under suitable choices of hyper-parameters. Exploring and discovering the benefits of Sinkhorn DRO in other applications may also be of future interest.

Acknowledgement

The work of Jie Wang and Yao Xie is partially supported by an NSF CAREER CCF-1650913, and NSF DMS-2134037, CMMI-2015787, CMMI-2112533, DMS-1938106, and DMS-1830210.

Notes

[1] Here the references essentially assume the numerical part of the probability vector is supported on the whole Euclidean space, such as the numerical features is supported on the entire Euclidean space.

[2] Sinha et al. [109] approximately solves the Wasserstein DRO by penalizing the Wasserstein ball constraint with fixed Lagrangian multiplier λ^*. Here the assumption of loss function holds for $\widehat{\mathbb{P}}$-almost every x.

[3] We say a transport cost $c(\cdot,\cdot)$ is strongly convex if $c(x,y) = u(x-y)$ for a strongly convex function $u(\cdot)$.

[4] We say that $f(\delta) = \mathcal{O}(g(\delta))$ if there exists a real constant $c > 0$ (which is independent of δ) and there exists $\delta_0 > 0$ such that $f(\delta) \le cg(\delta)$ for every $\delta \le \delta_0$. When $f(\delta) = \mathcal{O}\left(g(\delta) \cdot \text{polylog}\frac{1}{\delta}\right)$, we write $f(\delta) = \widetilde{\mathcal{O}}(g(\delta))$ for simplicity.

References

[1] Abdullah MA, Ren H, Ammar HB, Milenkovic V, Luo R, Zhang M, Wang J (2019) Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196* .

[2] Agrawal S, Ding Y, Saberi A, Ye Y (2012) Price of correlations in stochastic optimization. *Operations Research* 60(1):150–162.

[3] Altschuler J, Weed J, Rigollet P (2017) Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in Neural Information Processing Systems*, 1961–1971.

[4] ApS M (2021) Mosek modeling cookbook 3.2.3. https://docs.mosek.com/modeling-cookbook/index.html#.

[5] Asmussen S, Glynn PW (2007) *Stochastic simulation: algorithms and analysis*, volume 57 (Springer Science & Business Media).

[6] Azizian W, Iutzeler F, Malick J (2022) Regularization for wasserstein distributionally robust optimization. *arXiv preprint arXiv:2205.08826* .

[7] Bacharach M (1965) Estimating nonnegative matrices from marginal data. *International Economic Review* 6(3):294–310.

[8] Bai Y, Wu X, Ozgur A (2020) Information constrained optimal transport: From talagrand, to marton, to cover. *2020 IEEE International Symposium on Information Theory (ISIT)*, 2210–2215.

[9] Bayraksan G, Love DK (2015) Data-driven stochastic programming using phi-divergences. *The Operations Research Revolution*, 1–19 (INFORMS).

[10] Ben-Tal A, den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2):341–357.

[11] Bertsimas D, Kallus N (2020) From predictive to prescriptive analytics. *Management Science* 66(3):1025–1044.

[12] Bertsimas D, Natarajan K, Teo CP (2006) Persistence in discrete optimization under data uncertainty. *Mathematical programming* 108(2):251–274.

[13] Bertsimas D, Sim M, Zhang M (2019) Adaptive distributionally robust optimization. *Management Science* 65(2):604–618.

[14] Blanchet J, Chen L, Zhou XY (2022) Distributionally robust mean-variance portfolio selection with wasserstein distances. *Management Science* 68(9):6382–6410.

[15] Blanchet J, Glynn PW, Yan J, Zhou Z (2019) Multivariate distributionally robust convex regression under absolute error loss. *Advances in Neural Information Processing Systems*, volume 32, 11817–11826.

[16] Blanchet J, Kang Y (2020) Semi-supervised learning based on distributionally robust optimization. *Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods* 5:1–33.

[17] Blanchet J, Kang Y, Murthy K (2019) Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability* 56(3):830–857.

[18] Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research* 44(2):565–600.

[19] Blanchet J, Murthy K, Nguyen VA (2021) Statistical analysis of wasserstein distributionally robust estimators. *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, 227–254 (INFORMS).

[20] Blanchet J, Murthy K, Si N (2022) Confidence regions in wasserstein distributionally robust estimation. *Biometrika* 109(2):295–315.

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

[22] Blanchet JH, Glynn PW (2015) Unbiased monte carlo for optimization and functions of expectations via multi-level randomization. *2015 Winter Simulation Conference (WSC)*, 3656–3667.

[23] Chen R, Hao B, Paschalidis I (2021) Distributionally robust multiclass classification and applications in deep cnn image classifiers. *arXiv preprint arXiv:2109.12772* .

[24] Chen R, Paschalidis IC (2019) Selecting optimal decisions via distributionally robust nearest-neighbor regression. *Advances in Neural Information Processing Systems*.

[25] Chen T, Liu S, Chang S, Cheng Y, Amini L, Wang Z (2020) Adversarial robustness: From self-supervised pre-training to fine-tuning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 699–708.

[26] Chen Y, Sun H, Xu H (2020) Decomposition and discrete approximation methods for solving two-stage distributionally robust optimization problems. *Computational Optimization and Applications* 78(1):205–238.

[27] Chen Z, Kuhn D, Wiesemann W (2022) Data-driven chance constrained programs over wasserstein balls. *Operations Research* .

[28] Chen Z, Sim M, Xu H (2019) Distributionally robust optimization with infinitely constrained ambiguity sets. *Operations Research* 67(5):1328–1344.

[29] Cherukuri A, Cortés J (2019) Cooperative data-driven distributionally robust optimization. *IEEE Transactions on Automatic Control* 65(10):4400–4407.

[30] Coates A, Ng AY (2011) Analysis of large-scale visual recognition. *Advances in neural information processing systems* 24:873–881.

[31] Cohen MB, Lee YT, Miller G, Pachocki J, Sidford A (2016) Geometric median in nearly linear time. *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, 9–21.

[32] Courty N, Flamary R, Habrard A, Rakotomamonjy A (2017) Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*.

[33] Courty N, Flamary R, Tuia D (2014) Domain adaptation with regularized optimal transport. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 274–289.

[34] Courty N, Flamary R, Tuia D, Rakotomamonjy A (2016) Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(9):1853–1865.

[35] Cover TM, Thomas JA (2006) *Elements of Information Theory* (Wiley-Interscience).

[36] Cuturi M (2013) Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, volume 26, 2292–2300.

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

[38] Deming WE, Stephan FF (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics* 11(4):427–444.

[39] Derman E, Mannor S (2020) Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894* .

[40] Doan XV, Natarajan K (2012) On the complexity of nonoverlapping multivariate marginal bounds for probabilistic combinatorial optimization problems. *Operations research* 60(1):138–149.

[41] Duchi JC, Glynn PW, Namkoong H (2021) Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research* 0(0).

[42] Eckstein S, Kupper M, Pohl M (2020) Robust risk aggregation with neural networks. *Mathematical Finance* 30(4):1229–1272.

[43] Esfahani PM, Kuhn D (2018) Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 171(1):115–166.

[44] Feng Y, Schlögl E (2018) Model risk measurement under wasserstein distance. *arXiv preprint arXiv:1809.03641* .

[45] Fréchet M (1960) Sur les tableaux dont les marges et des bornes sont données. *Revue de l'Institut international de statistique* 10–32.

[46] Gao R (2022) Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research* .

[47] Gao R, Chen X, Kleywegt AJ (2022) Wasserstein distributionally robust optimization and variation regularization. *Operations Research* .

[48] Gao R, Kleywegt A (2022) Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research* .

[49] Gao R, Kleywegt AJ (2017) Data-driven robust optimization with known marginal distributions. *Working paper. Available at https://faculty.mccombs.utexas.edu/rui.gao/copula.pdf* .

[50] Gao R, Kleywegt AJ (2017) Distributionally robust stochastic optimization with dependence structure. *arXiv preprint arXiv:1701.04200* .

[51] Genevay A, Cuturi M, Peyré G, Bach F (2016) Stochastic optimization for large-scale optimal transport. *Advances in Neural Information Processing Systems*, volume 29.

[52] Genevay A, Peyre G, Cuturi M (2018) Learning generative models with sinkhorn divergences. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, 1608–1617 (PMLR).

[53] Ghadimi S, Lan G, Zhang H (2016) Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming* 155(1):267–305.

[54] Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. *Operations Research* 58(4-part-1):902–917.

[55] Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* .

[56] Härdle W (1990) *Applied nonparametric regression* (Cambridge university press).

[57] He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

[58] He W, Wei J, Chen X, Carlini N, Song D (2017) Adversarial example defense: Ensembles of weak defenses are not strong. *WOOT*, 15–15.

[59] Hu Y, Chen X, He N (2020) Sample complexity of sample average approximation for conditional stochastic optimization. *SIAM Journal on Optimization* 30(3):2103–2133.

[60] Hu Y, Chen X, He N (2021) On the bias-variance-cost tradeoff of stochastic optimization. *Advances in Neural Information Processing Systems*.

[61] Hu Y, Zhang S, Chen X, He N (2020) Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. *Advances in Neural Information Processing Systems*, volume 33, 2759–2770.

[62] Hu Z, Hong LJ (2012) Kullback-leibler divergence constrained distributionally robust optimization. *Optimization Online preprint Optimization Online:2012/11/3677* .

[63] Huang M, Ma S, Lai L (2021) A riemannian block coordinate descent method for computing the projection robust wasserstein distance. *Proceedings of the 38th International Conference on Machine Learning*, 4446–4455.

[64] Kallenberg O, Kallenberg O (1997) *Foundations of modern probability*, volume 2 (Springer).

[65] Krizhevsky A, Hinton G (2009) Learning multiple layers of features from tiny images. Technical report, Citeseer.

[66] Kruithof J (1937) Telefoonverkeersrekening. *De Ingenieur* 52:15–25.

[67] Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Operations Research & Management Science in the Age of Analytics*, 130–166 (INFORMS).

[68] Levy D, Carmon Y, Duchi JC, Sidford A (2020) Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems* 33:8847–8860.

[69] Li J, Lin S, Blanchet J, Nguyen VA (2022) Tikhonov regularization is optimal transport robust under martingale constraints. *arXiv preprint arXiv:2210.01413* .

[70] Lin CJ (accessed on 2023-04-24) Regression datasets - libsvm tools. https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html.

[71] Lin T, Fan C, Ho N, Cuturi M, Jordan M (2020) Projection robust wasserstein distance and riemannian optimization. *Advances in Neural Information Processing Systems*, volume 33, 9383–9397.

[72] Lin T, Ho N, Jordan MI (2022) On the efficiency of entropic regularized algorithms for optimal transport. *Journal of Machine Learning Research* 23(137):1–42.

[73] Liu Y, Yuan X, Zhang J (2021) Discrete approximation scheme in distributionally robust optimization. *Numer Math Theory Methods Appl* 14(2):285–320.

[74] Luise G, Rudi A, Pontil M, Ciliberto C (2018) Differential properties of sinkhorn approximation for learning with wasserstein distance. *Advances in Neural Information Processing Systems*.

[75] Luo F, Mehrotra S (2019) 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.

[76] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2019) Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* .

[77] Mensch A, Peyré G (2020) Online sinkhorn: Optimal transport distances from sample streams. *Advances in Neural Information Processing Systems* 33:1657–1667.

[78] Mohajerin Esfahani P, Kuhn D (2017) Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming* 171(1):115–166.

[79] Namkoong H, Duchi JC (2016) Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in Neural Information Processing Systems*, volume 29, 2208–2216.

[80] Natarajan K, Song M, Teo CP (2009) Persistency model and its applications in choice modeling. *Management Science* 55(3):453–469.

[81] Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization* 19(4):1574–1609.

[82] Nemirovsky A, Yudin D (1983) Problem complexity and method efficiency in optimization. *John Wiley & Sons* .

[83] Nesterov Y, Nemirovskii A (1994) *Interior-point polynomial algorithms in convex programming* (SIAM).

[84] Nguyen VA, Si N, Blanchet J (2020) Robust bayesian classification using an optimistic score ratio. *International Conference on Machine Learning*, 7327–7337.

[85] Nguyen VA, Zhang F, Blanchet J, Delage E, Ye Y (2021) Robustifying conditional portfolio decisions via optimal transport. *arXiv preprint arXiv:2103.16451* .

[86] Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A (2017) Practical black-box attacks against machine learning. *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 506–519.

[87] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A (2016) The limitations of deep learning in adversarial settings. *2016 IEEE European symposium on security and privacy (EuroS&P)*, 372–387 (IEEE).

[88] Papernot N, McDaniel P, Wu X, Jha S, Swami A (2016) Distillation as a defense to adversarial perturbations against deep neural networks. *2016 IEEE symposium on security and privacy (SP)*, 582–597 (IEEE).

[89] Patrini G, van den Berg R, Forre P, Carioni M, Bhargav S, Welling M, Genewein T, Nielsen F (2020) Sinkhorn autoencoders. *Uncertainty in Artificial Intelligence*, 733–743.

[90] Petzka H, Fischer A, Lukovnikov D (2018) On the regularization of wasserstein GANs. *International Conference on Learning Representations*.

[91] Peyre G, Cuturi M (2019) Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning* 11(5-6):355–607.

[92] Pflug G, Wozabal D (2007) Ambiguity in portfolio selection. *Quantitative Finance* 7(4):435–442.

[93] Pichler A, Shapiro A (2021) Mathematical foundations of distributionally robust multistage optimization. *SIAM Journal on Optimization* 31(4):3044–3067.

[94] Popescu I (2005) A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operations Research* 30(3):632–657.

[95] Qi Q, Lyu J, Bai EW, Yang T, et al. (2022) Stochastic constrained dro with a complexity independent of sample size. *arXiv preprint arXiv:2210.05740* .

[96] Rahimian H, Mehrotra S (2019) Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659* .

[97] Rakhlin A, Shamir O, Sridharan K (2012) Making gradient descent optimal for strongly convex stochastic optimization. *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 1571–1578.

[98] Rockafellar RT, Uryasev S, et al. (1999) Optimization of conditional value-at-risk. *Journal of risk* 2:21–42.

[99] Rozsa A, Gunther M, Boult TE (2016) Towards robust deep neural networks with bang. *arXiv preprint arXiv:1612.00138* .

[100] Scarf H (1957) A min-max solution of an inventory problem. *Studies in the mathematical theory of inventory and production* .

[101] Selvi A, Belbasi MR, Haugh MB, Wiesemann W (2022) Wasserstein logistic regression with mixed features. *Advances in Neural Information Processing Systems*.

[102] Shafieezadeh-Abadeh S, Aolaritei L, Dörfler F, Kuhn D (2023) New perspectives on regularization and computation in optimal transport-based distributionally robust optimization. *arXiv preprint arXiv:2303.03900* .

[103] Shafieezadeh-Abadeh S, Kuhn D, Esfahani PM (2019) Regularization via mass transportation. *Journal of Machine Learning Research* 20(103):1–68.

[104] Shafieezadeh Abadeh S, Mohajerin Esfahani PM, Kuhn D (2015) Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, volume 28.

[105] Shapiro A, Dentcheva D, Ruszczynski A (2021) *Lectures on stochastic programming: modeling and theory* (SIAM).

[106] Shapiro A, Zhou E, Lin Y (2021) Bayesian distributionally robust optimization. *arXiv preprint arXiv:2112.08625* .

[107] Singh D, Zhang S (2021) Distributionally robust profit opportunities. *Operations Research Letters* 49(1):121–128.

[108] Singh D, Zhang S (2022) Tight bounds for a class of data-driven distributionally robust risk measures. *Applied Mathematics & Optimization* 85(1):1–41.

[109] Sinha A, Namkoong H, Duchi J (2018) Certifiable distributional robustness with principled adversarial training. *International Conference on Learning Representations*.

[110] Sinkhorn R (1964) A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics* 35(2):876–879.

[111] Smirnova E, Dohmatob E, Mary J (2019) Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708* .

[112] Song J, Zhao C, He N (2022) Efficient wasserstein and sinkhorn policy optimization. URL https://openreview.net/forum?id=Mlwe37htstv.

[113] Staib M, Jegelka S (2019) Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems* 32:9134–9144.

[114] Taskesen B, Nguyen VA, Kuhn D, Blanchet J (2020) A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530* .

[115] TinyImageNet (2014) TinyImageNet Visual Recognition Challenge. https://tiny-imagenet.herokuapp.com/.

[116] Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P (2017) Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204* .

[117] Van Parys BP, Goulart PJ, Kuhn D (2015) Generalized gauss inequalities via semidefinite programming. *Mathematical Programming* 156(1-2):271–302.

[118] Vandenberghe L, Boyd S (1995) Semidefinite programming. *SIAM review* 38(1):49–95.

[119] Wang C, Gao R, Qiu F, Wang J, Xin L (2018) Risk-based distributionally robust optimal power flow with dynamic line rating. *IEEE Transactions on Power Systems* 33(6):6074–6086.

[120] Wang J, Gao R, Xie Y (2021) Two-sample test using projected wasserstein distance. *2021 IEEE International Symposium on Information Theory (ISIT)*.

[121] Wang J, Gao R, Xie Y (2022) Two-sample test with kernel projected wasserstein distance. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, 8022–8055 (PMLR).

[122] Wang J, Gao R, Zha H (2022) Reliable off-policy evaluation for reinforcement learning. *Operations Research* .

[123] Wang J, Jia Z, Yin H, Yang S (2021) Small-sample inferred adaptive recoding for batched network coding. *2021 IEEE International Symposium on Information Theory (ISIT)*.

[124] Wang Z, Glynn PW, Ye Y (2015) Likelihood robust optimization for data-driven problems. *Computational Management Science* 13(2):241–261.

[125] Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. *Operations Research* 62(6):1358–1376.

[126] Wozabal D (2012) A framework for optimization under ambiguity. *Annals of Operations Research* 193(1):21–47.

[127] Xie W (2019) On distributionally robust chance constrained programs with wasserstein distance. *Mathematical Programming* 186(1):115–155.

[128] Yang I (2017) A convex optimization approach to distributionally robust markov decision processes with wasserstein distance. *IEEE control systems letters* 1(1):164–169.

[129] Yang I (2020) Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Transactions on Automatic Control* 66(8):3863–3870.

[130] Yu Y, Lin T, Mazumdar EV, Jordan M (2022) Fast distributionally robust learning with variance-reduced min-max optimization. *International Conference on Artificial Intelligence and Statistics*, 1219–1250.

[131] Yule GU (1912) On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society* 75(6):579–652.

[132] Zhao C, Guan Y (2018) Data-driven risk-averse stochastic optimization with wasserstein metric. *Operations Research Letters* 46(2):262–267.

[133] Zhu J, Jitkrittum W, Diehl M, Schölkopf B (2021) Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 280–288.

[134] Zymler S, Kuhn D, Rustem B (2013) Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming* 137(1):167–198.

Supplementary for *"Sinkhorn Distributionally Robust Optimization"*

EC.1. Detailed Experiment Setup

The experiments were conducted on a MacBook Pro laptop with 32GB of memory running Python 3.7. Unless stated otherwise, we solved the SAA, Wasserstein DRO, and KL-divergence DRO baseline models exactly using the interior point method-based solver Mosek [4]. Optimization hyperparameters, such as step size, maximum iterations, and number of levels, were tuned to minimize training error after 10 outer iterations. We use RT-MLMC gradient estimator to solve the Sinkhorn DRO model. We employed the *warm starting* strategy during the iterative procedure: we set the initial guess of parameter θ at the beginning of outer iteration as the one obtained from the SAA approach. At other outer iterations, the initial guess of parameter θ is set to be the final obtained solution θ at the last outer iteration. The following subsections outline some special reformulations, optimization algorithms used to solve the baseline models together with the running time of all approaches.

EC.1.1. Setup for Newsvendor Problem and Running Time

To solve the 2-Wasserstein DRO model with radius ρ, we approximate the support of worst-case distribution using discrete grid points. Denote by $\mathcal{D}_n = \{x_1, \ldots, x_n\}$ the set of observed n samples and $\mathcal{G}_{200-n}$ the set of $200-n$ points evenly supported on the interval $[0, 10]$. Then the support of worst-case distribution is restricted to $\mathcal{D}_n \cup \mathcal{G}_{200-n} := \{\hat{z}_1, \ldots, \hat{z}_{200}\}$. The corresponding 2-Wasserstein DRO problem has the following linear programming reformulation:

$$\begin{aligned} \min_{\theta,\lambda,s} \quad & \lambda\rho + \frac{1}{n}\sum_{i=1}^{n} s_i \\ \text{s.t.} \quad & k\theta - u\min(\theta, \hat{z}_j) - \lambda(x_i - \hat{z}_j)^2 \le s_i, \quad \forall i \in [n], \forall j \in [200]. \end{aligned}$$

The computational time for the newsvendor problem in Section 5.1 is reported in Table EC.1. We observe that the training time of 2-Wasserstein DRO model increases quickly as the sample size increases, while the training time of other DRO models increases mildly in the training sample size.

Table EC.1 Average computational time (in seconds) per problem instance for the newsvendor problem.

Model	Exponential			Gamma			Gaussian Mixture		
	$n=10$	$n=30$	$n=100$	$n=10$	$n=30$	$n=100$	$n=10$	$n=30$	$n=100$
SAA	0.017	0.017	0.018	0.019	0.019	0.019	0.023	0.024	0.024
KL-DRO	0.027	0.029	0.040	0.027	0.028	0.039	0.027	0.028	0.038
1-SDRO	0.110	0.124	0.161	0.105	0.119	0.162	0.105	0.119	0.157
2-WDRO	0.123	0.358	1.307	0.128	0.354	1.337	0.134	0.402	1.428
2-SDRO	0.061	0.069	0.106	0.100	0.121	0.161	0.095	0.115	0.154

EC.1.2. Setup for Mean-risk Portfolio Optimization and Running Time

From [78, Eq. (27)] we can see that the 1-Wasserstein DRO formulation with radius ρ for the portfolio optimization problem becomes

$$\begin{aligned} \min_{\theta,\tau,\lambda,s} \quad & \lambda\rho + \frac{1}{n}\sum_{i=1}^{n} s_i \\ \text{s.t.} \quad & \theta \in \Theta, \quad b_j\tau + a_j\langle\theta, \hat{z}_i\rangle \le s_i, i \in [n], j \in [H], \\ & \|a_j\theta\|_2 \le \lambda, j \in [H]. \end{aligned}$$

Also, we argue that the 2-Wasserstein DRO formulation with radius ρ for the portfolio optimization problem has a finite convex reformulation:

$$\begin{aligned}&\inf_{\theta\in\Theta,\tau}\ \sup_{\mathbb{P}:\ W_2(\mathbb{P},\widehat{\mathbb{P}}_n)\le\rho}\ \mathbb{E}_{\mathbb{P}}\big[\max_{j\in[H]} a_j\langle\theta,z\rangle+b_j\tau\big]\\ =&\inf_{\theta\in\Theta,\tau,\lambda\ge 0}\ \left\{\lambda\rho^2+\frac{1}{n}\sum_{i=1}^n\sup_{s_i}\ \left\{\max_{j\in[H]} a_j\langle\theta,s_i\rangle+b_j\tau-\lambda\|s_i-\hat{z}_i\|_2^2\right\}\right\}.\end{aligned}$$

In particular, the inner subproblem has the following reformulation:

$$\begin{aligned}&\sup_{s_i}\ \left\{\max_{j\in[H]} a_j\langle\theta,s_i\rangle+b_j\tau-\lambda\|s_i-\hat{z}_i\|_2^2\right\}\\ =&\max_{j\in[H]}\ b_j\tau+\sup_{s_i}\left\{a_j\langle\theta,s_i\rangle-\lambda\|s_i-\hat{z}_i\|_2^2\right\}\\ =&\max_{j\in[H]}\ b_j\tau+\frac{a_j^2}{4\lambda}\|\theta\|_2^2+a_j\langle\theta,\hat{z}_i\rangle.\end{aligned}$$

Hence, the 2-Wasserstein DRO can be reformulated as

$$\begin{aligned}\min_{\theta,\tau,\lambda,s}\quad&\lambda\rho^2+\frac{1}{n}\sum_{i=1}^n s_i\\ \text{s.t.}\quad&\theta\in\Theta,\quad b_j\tau+a_j\langle\theta,\hat{z}_i\rangle+\frac{a_j^2}{4\lambda}\|\theta\|_2^2\le s_i,\quad i\in[n],j\in[H].\end{aligned}$$

The computational time for the portfolio optimization problem in Section 5.2 is reported in Table EC.2. We observe that when the data dimension is fixed and the sample size varies from 30 to 400, the computational time for all approaches does not differ too much. When the data dimension increases and the sample size is fixed, the computational time for 1-SDRO or 2-SDRO model increases linearly while other DRO models increase mildly. One possible explanation is that in this example, other DRO models have tractable finite-dimensional conic programming formulations so that off-the-shelf software can solve them efficiently. In contrast, Sinkhorn DRO models do not have special reformulation, but they can still be solved in a reasonable amount of time.

EC.1.3. Setup for Adversarial Multi-class Logistic Regression and Running Time

The procedure for generating various adversarial perturbations is reported in the following:

(I) For a given classifer B and data sample $(x,\boldsymbol{y})$, the ℓ_p-norm ($p\in\{1,2\}$) adversarial attack based on projected gradient method [76] iterates as follows: $x_0\leftarrow x$ and

$$\begin{cases}\Delta x^{k+1}\leftarrow\arg\max\limits_{\|\eta\|_p\le\xi}\ \left\{\nabla_x h_B(x^k,\boldsymbol{y})^{\mathrm{T}}\eta\right\},\\ x^{k+1}\leftarrow\mathrm{Proj}_{\{x':\ \|x-x'\|_p\le\xi\}}\ \left\{x^k+\dfrac{\alpha}{\sqrt{k+1}}\Delta x^{k+1}\right\}.\end{cases}$$

We perform the gradient update above for 15 steps. Also, we specify the initial learning rate $\alpha=1$ and vary the radius of attack $\xi\in\{0,0.2,0.4,0.6,0.8,1\}$.

(II) For a given feature vector x, the perturbed feature using white Laplacian noise becomes $x+\xi\|x\|_2\cdot\zeta$, where the random vector ζ follows the isotropic Laplace distribution with zero mean and unit variance, and we vary the ratio $\xi\in\{0,4\cdot10^{-4},8\cdot10^{-4},1.2\cdot10^{-3},1.6\cdot10^{-3},2\cdot10^{-3}\}$. Similarly, the perturbed feature using white Gaussian noise becomes $x+\xi\|x\|_2\cdot\zeta$, with ζ being the isotropic Gaussian distribution with zero mean and unit variance. Our experiment setup that adds white noise is inspired from [23, Section 4].

Table EC.2 Average computational time (in seconds) per problem instance for portfolio optimization problem.

(n, d) Values	SAA	KL-DRO	1-WDRO	1-SDRO	2-WDRO	2-SDRO
$(30, 30)$	0.013	0.038	0.018	0.125	0.015	0.090
$(50, 30)$	0.014	0.042	0.020	0.163	0.016	0.110
$(100, 30)$	0.017	0.065	0.024	0.167	0.021	0.144
$(150, 30)$	0.019	0.084	0.029	0.158	0.027	0.152
$(200, 30)$	0.023	0.115	0.035	0.151	0.032	0.150
$(400, 30)$	0.045	0.136	0.061	0.167	0.056	0.142
$(100, 5)$	0.014	0.043	0.017	0.082	0.015	0.111
$(100, 10)$	0.014	0.045	0.018	0.109	0.015	0.128
$(100, 20)$	0.014	0.048	0.021	0.138	0.017	0.187
$(100, 40)$	0.017	0.068	0.027	0.179	0.022	0.269
$(100, 80)$	0.021	0.103	0.052	0.357	0.044	0.462
$(100, 100)$	0.023	0.116	0.070	0.387	0.059	0.478

In this example, we use stochastic gradient methods to solve the SAA formulation and all penalized DRO formulations. Inspired from classical adversarial training approaches that usually require fine-tuning of hyper-parameters via grid search [25], we fine-tune the hyper-parameters (i.e., Lagrangian multiplier λ or regularization value ϵ) of DRO formulations from the following grid of values and report the optimal performance:

- KL-DRO/1-WDRO/2-WDRO: $\lambda \in \{0.05, 0.1, 0.5, 1, 3, 5, 10, 15, 20, 25, 30, 100\}$.
- 1-SDRO/2-SDRO: $\lambda \in \{1, 5, 10\}$ and $\epsilon \in \{0.1, 0.5, 1\}$.

We terminate the training of SAA or DRO models when the number of epoches, i.e., the number of times for processes each training sample, exceeds 30. In is worth mentioning that the Wasserstein DRO model with a fixed Lagrangian multiplier λ using samples $\{x_i, \boldsymbol{y}_i\}_{i=1}^n$ can be reformulated as

$$\min_B \frac{1}{n} \sum_{i=1}^n \left[\max_{x \in \mathbb{R}^d} \left\{ h_B(x, \boldsymbol{y}_i) - \lambda c(x_i, x) \right\} \right]. \tag{EC.1}$$

To approximately solve such a convex-non-concave problem, we implement a gradient descent ascent heuristic outlined in Algorithm 4. We report basic statistics of classification datasets in this example in Table EC.3. Besides, the computational time is reported in Table EC.4. The results indicate that Sinkhorn DRO models have shorter computational time than Wasserstein DRO models in general.

Table EC.3 Basic statistics of adversarial multi-class logistic regression datasets.

	CIFAR10	CIFAR100	STL10	tinyImageNet
Image Size (before pre-processing)	3072	3072	27648	12288
Feature Dimension (after pre-processing)	512	512	512	512
# of classes	10	100	10	200
Training Size	50000	50000	5000	90000
Testing Size	10000	10000	8000	10000

Table EC.4 Average computational time (in seconds) per problem instance for adversarial multi-class logistic regression problem

Dataset	SAA	KL-DRO	1-WDRO	1-SDRO	2-WDRO	2-SDRO
CIFAR10	10.55	12.11	55.10	59.52	53.04	47.26
CIFAR100	15.64	16.01	125.31	114.99	128.20	106.62
STL10	1.37	1.59	7.36	7.34	7.96	6.99
tinyImageNet	45.54	44.50	325.25	227.91	347.16	197.55

Algorithm 4 Heuristic Gradient Descent Ascent algorithm (inspired from [109, Algorithm 1]) for solving (EC.1). We use default values $\alpha = \texttt{1e-2}, m = 100, n_{\text{critic}} = 15$.

Require: Learning rate α, batch size m, number of inner iterations n_{critic}, number of outer iterations T_{out}, initial guess B.

for $t_{\text{out}} = 0, 1, \ldots, T_{\text{out}} - 1$ **do**
  Sample a subset of indices $\{n_i\}_{i=1}^m$ from $[n]$.
  Initialize $\hat{x}_{n_i} \leftarrow x_{n_i}, i \in [m]$.
  for $t = 0, 1, \ldots, n_{\text{critic}} - 1$ **do**
    Compute $g_{n_i} \leftarrow \nabla_{z_{n_i}} \left[h_B(\hat{x}_{n_i}, \boldsymbol{y}_{n_i}) - \lambda c(\hat{x}_{n_i}, x_{n_i}) \right]$ for $i \in [m]$.
    Update $\hat{x}_{n_i} \leftarrow \hat{x}_{n_i} + \alpha \text{Adam}(\hat{x}_{n_i}, g_{n_i})$ for $i \in [m]$.
  end for
  Compute $g_B \leftarrow \nabla_B \left[\frac{1}{m} \sum_{i=1}^m h_B(\hat{x}_{n_i}) \right]$.
  Update $B \leftarrow B - \alpha \text{Adam}(B, g_B)$.
end for
Output B.

EC.2. Additional Validation Experiments

EC.2.1. Comparison of Optimization Algorithms: Distributionally Robust Linear Regression

To examine the performance of different gradient estimators, we study the problem of distributionally robust linear regression (see the setup in Example 2). We take the nominal distribution $\widehat{\mathbb{P}}$ as the empirical one based on samples $\{(a_i, b_i)\}_{i=1}^n$. As a consequence, the inner objective function in (10) has the closed form expression:

$$F(\theta) = \frac{1}{n}\sum_{i=1}^{n}(a_i^{\mathrm{T}}\theta - b_i)^2 + \frac{\frac{1}{n}\sum_{i=1}^{n}(a_i^{\mathrm{T}}\theta - b_i)^2}{\frac{1}{2}\lambda\|\theta\|_2^{-2} - 1} - \frac{\lambda\epsilon}{2}\log\det\left(I - \frac{\theta\theta^{\mathrm{T}}}{\frac{1}{2}\lambda}\right), \quad \text{if } \|\theta\|_2^2 < \frac{\lambda}{2},$$

and otherwise $F(\theta) = \infty$. We take the constraint set $\Theta = \{\theta : \ \|\theta\|_2^2 \le 0.999 \cdot \frac{\lambda}{2}\}$. Similar to the setup in [69, Section 5.1], we examine the performance using three LIBSVM regression real world datasets [70]: housing, mg, and mpg.

The quality of proposed gradient estimators is examined in a single BSMD step with specified hyper-parameters $(\lambda, \epsilon) = (10^3, 10^{-1})$. For baseline comparison, we also study the performance of the standard MLMC scheme. It has been shown in [60] that this gradient estimator achieves the same sample complexity rate as RT-MLMC estimator for convex smooth CSO problem. However, this estimator will lead to sub-optimal complexity rate compared with RT-MLMC when the inner objective function is no longer smooth. We verify its performance in a toy experiment with a convex nonsmooth loss function in the next subsection.

MLMC Estimator: Fix some number $N > 0$. For $\ell = 0, 1, \ldots, L$, get $n_\ell := \lceil 2^{-\ell}N \rceil$ random variables $\{A^\ell(\theta, \zeta_i^\ell)\}_{i=1}^{n_\ell}$, where $\{\zeta_i^\ell\}_i$ are i.i.d. copies of ζ^ℓ. Construct

$$\widehat{F}^{\text{MLMC}}(\theta) = \sum_{\ell=0}^{L}\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}A^\ell(\theta, \zeta_i^\ell). \tag{EC.2}$$

Besides, we also compare two unbiased gradient estimators in literature [60]. However, the variance of them are unbounded, so that there is no convergence analysis for those two methods.

RU-MLMC Estimator: at point θ, we sample a random level ι, following a distribution $\mathbb{P}(\iota = \ell) = q_\ell, \ell = 0, 1, \ldots, \infty$, where the probability mass value $q_\ell \propto 2^{-\ell}$. Then construct

$$v^{\text{RU-MLMC}}(\theta) := \frac{1}{q_\iota}G^\iota(\theta, \zeta^\iota). \tag{EC.3}$$

RR-MLMC Estimator: at point θ, we sample a random level L, following a distribution $\mathbb{P}(\iota = \ell) = q_\ell, \ell = 0, 1, \ldots, \infty$, where the probability mass value $q_\ell \propto 2^{-\ell}$. Then construct

$$v^{\text{RR-MLMC}}(\theta) := \sum_{\ell=0}^{L} p_\ell G^\ell(x, \zeta^\ell), \tag{EC.4}$$

where $p_\ell := \frac{1}{1-\sum_{\ell'=0}^{\ell-1} q_{\ell'}}$ and $\sum_{\ell'=0}^{-1} q_{\ell'} = 0$.

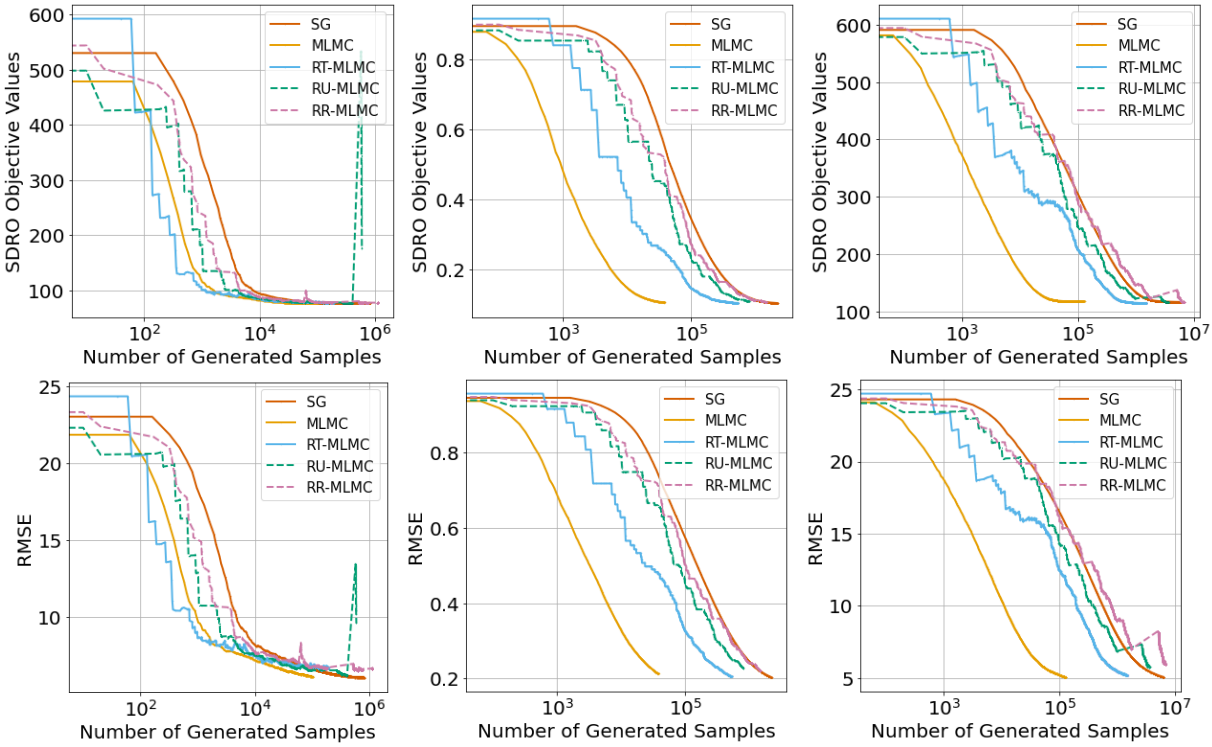

Figure EC.1 Comparison results of SG, MLMC, RT-MLMC, RU-MLMC, and RR-MLMC on robust linear regression problem. From left to right, the figures correspond to three different regression datasets: (a) housing; (b) mg; and (c) mpg. From top to bottom, the figures correspond to plots of (a) Sinkhorn DRO objective values; and (b) RMSE of obtained solutions.

For a given solution θ, we quantify its performance using the corresponding Sinkhorn DRO objective value. Besides, we report its root-mean-square error (RMSE) on training data. Thus, the smaller those two performance criteria are, the smaller the solution's optimization performance has. Figure EC.1 shows the performance of various gradient estimators in terms of the number of generated samples based on these criteria. The results demonstrate that the SG scheme does not perform competitively, as expected from our theoretical analysis that shows SG has the worst complexity order. In contrast, using other four types of MLMC methods, we can obtain optimal solutions with small sample complexity. While the RU-MLMC and RR-MLMC schemes exhibit competitive performance, the optimization procedure shows some oscillations. One possible explanation is that the variance values of those gradient estimators are unbounded, making these two approaches unstable.

EC.2.2. Comparison of Optimization Algorithms: Distributionally Robust Portfolio Optimization

In this subsection, we validate the competitive performance of RT-MLMC gradient estimator on the case where the loss is convex and nonsmooth, and we try to solve the 2-SDRO formulation. We consider the portfolio optimization problem, and specify different choices of the problem parameters (sample size n and data dimension d) as $(50,50),(100,100),(400,400)$. We quantify the performance of obtained solution using the Sinkhorn DRO objective value. Since in this problem setup no analytical expression of the objective value is available, we estimate the objective value using (14) with hyper-parameters $L=8$ and $n_L^\circ=10^3$. Figure EC.2 shows the performance in terms of the number of generated samples based on this criterion. The results demonstrate that even for nonsmooth loss function, those listed MLMC-based gradient estimators have better performance than the SG scheme. Besides, the proposed RT-MLMC and standard MLMC schemes have comparable performance, and in some cases MLMC scheme even has better performance. It is an open question that whether the MLMC scheme will have the same order of sample complexity as the RT-MLMC scheme for convex nonsmooth optimization, which can be a topic for future study.

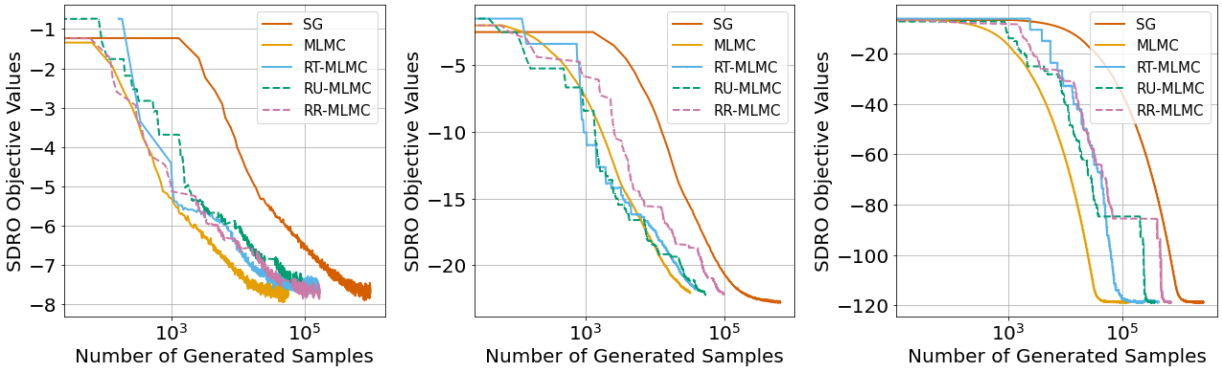

Figure EC.2 Comparison results of SG, MLMC, RT-MLMC, RU-MLMC, and RR-MLMC on robust portfolio optimization problem. From left to right, the figures correspond to three different specifications of $(n,d)\in\{(50,50),(100,100),(400,400)\}$.

EC.2.3. Sensitivity of Regularization Parameters

In this subsection, we validate the impact of regularization parameter ϵ on the performance of the Sinkhorn DRO model in two numerical examples: newsvendor and portfolio optimization problem. We examine the performance of 1-SDRO or 2-SDRO models for different regularization parameters chosen from the candidate set $\mathcal{A}$:

$$\mathcal{A} = \begin{cases} \{10^{-3}, 5\cdot 10^{-3}, 10^{-2}, 5\cdot 10^{-2}, 10^{-1}, 5\cdot 10^{-1}, 10^{0}\}, & \text{for newsvendor problem,} \\ \{5\cdot 10^{-2}, 10^{-1}, 5\cdot 10^{-1}, 10^{0}, 3\cdot 10^{0}, 5\cdot 10^{0}, 10^{1}\}, & \text{for portfolio optimization.} \end{cases}$$

For each fixed regularization parameter ϵ, we tune the corresponding Sinkhorn DRO radius $\overline{\rho}$ by cross validation. We quantify the performance of a given solution θ obtained from DRO models using the *performance gap* metric $\frac{J(\theta)-J^*}{1+|J^*|}$, where notations J^* and $J(\theta)$ are defined at the beginning of Section 5. Hence, the smaller the metric is, the better the given decision has.

Fig. EC.3 shows box plots on the performance of Sinkhorn DRO models across different choices of regularization values with different data distributions on the newsvendor problem. We can see as the regularization value increases, the performance of Sinkhorn DRO models generally improves first and then degrades.

Fig. EC.4 shows performance on the portfolio optimization problem with different choices of problem parameters (n, d), where n denotes the sample size and d denotes the data dimension. Compared with the newsvendor problem, we find a more clear trend that the model performance improves and then degrades as the regularization value increases. In this special example, we also find 2-SDRO model has more stable and satisfactory performance compared with 1-SDRO model.

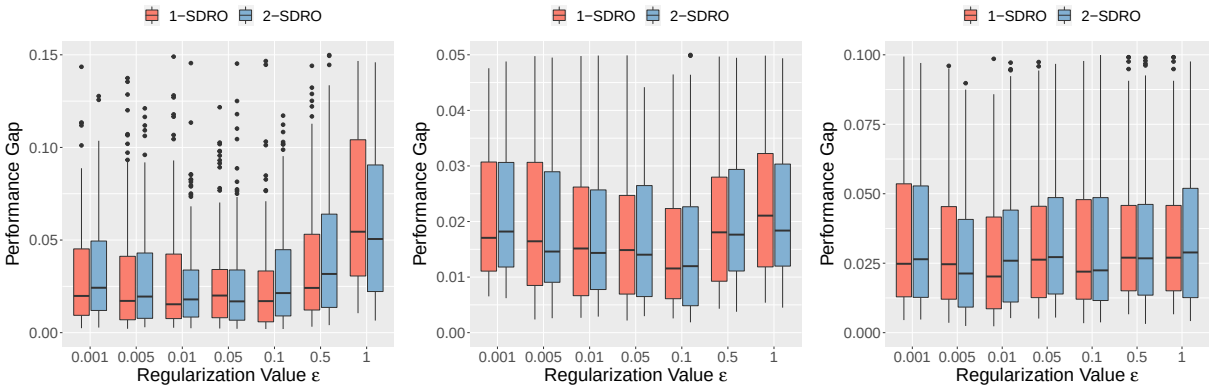

Figure EC.3 Performance of Sinkhorn DRO models for newsvendor problem versus different choices of regularization values ϵ. For figures from left to right, we specify the data distribution as exponential distribution, gamma distribution, and equiprobable mixture of two truncated normal distributions, respectively.

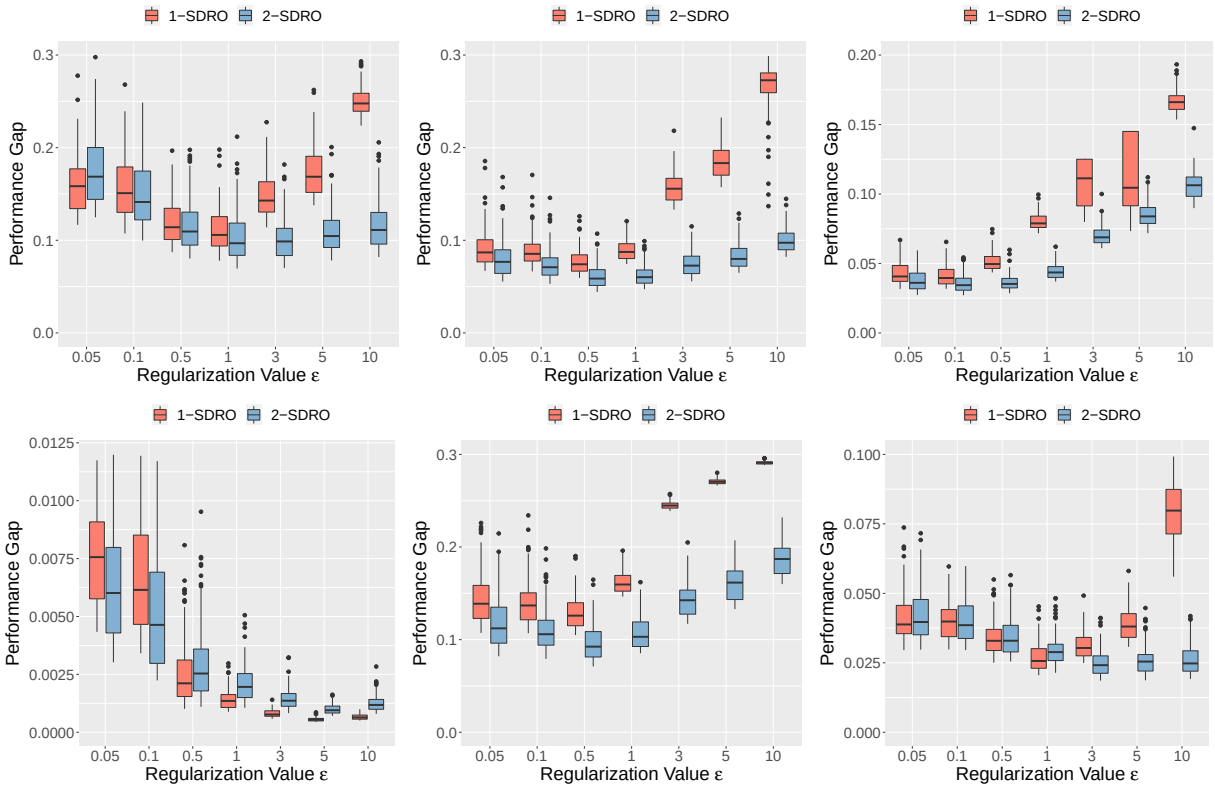

Figure EC.4 Performance of Sinkhorn DRO models for portfolio problem versus different choices of regularization values ϵ. For those fix figures from left to right, from top to bottom, we specify the problem parameters (sample size n and data dimension d) as $(30, 30), (100, 30), (400, 30), (100, 5), (100, 20), (100, 100)$, respectively.

EC.3. Sufficient Condition for Condition 1

PROPOSITION EC.1. *Condition 1 holds if there exists $p \geq 1$ so that the following conditions are satisfied:*

(I) *For any $x, y, z \in \mathcal{Z}$, $c(x,y) \geq 0$, and $(c(x,y))^{1/p} \leq (c(x,z))^{1/p} + (c(z,y))^{1/p}$.*

(II) *The nominal distribution $\widehat{\mathbb{P}}$ has a finite mean, denoted as $\overline{x}$. Moreover, $\nu\{z:\ 0 \leq c(\overline{x}, z) < \infty\} = 1$ and $\Pr_{x\sim\widehat{\mathbb{P}}}\{c(x,\overline{x}) < \infty\} = 1$.*

(III) *Assumption 1(III) holds, and there exists $\lambda > 0$ such that*

$$\mathbb{E}_{z\sim\nu}\left[e^{f(z)/(\lambda\epsilon)} e^{-2^{1-p}c(\overline{x},z)/\epsilon}\right] < \infty.$$

We make some remarks for the sufficient conditions listed above. The first condition can be satisfied by taking the transport cost as the p-th power of the metric defined on $\mathcal{Z}$ for any $p \geq 1$. The second condition requires the nominal distribution $\widehat{\mathbb{P}}$ is finite almost surely, e.g., it can be a subguassian distribution with respect to the transport cost c. We first present an useful technical lemma before showing the proof of Proposition EC.1.

LEMMA EC.1. *Under the first condition of Proposition EC.1, for any $x \in \mathcal{Z}$, it holds that*

$$\mathbb{E}_{z\sim\nu}\left[\int e^{-c(x,z)/\epsilon}\right] \geq e^{-2^{p-1}c(x,\overline{x})/\epsilon}\mathbb{E}_{z\sim\nu}\left[e^{-2^{p-1}c(\overline{x},z)/\epsilon}\right].$$

Proof of Lemma EC.1. Based on the inequality $(a+b)^p \leq 2^{p-1}(a^p + b^p)$, we can see that

$$c(x,z) \leq (c(y,z)^{1/p} + c(z,y)^{1/p})^p \leq 2^{p-1}(c(y,z) + c(z,y)), \quad \forall x, y, z \in \mathcal{Z}.$$

Since $c(x,z) \leq 2^{p-1}(c(\overline{x},z) + c(x,\overline{x}))$, we can see that

$$\mathbb{E}_{z\sim\nu}\left[\int e^{-c(x,z)/\epsilon}\right] \geq \exp\left(-2^{p-1}c(x,\overline{x})/\epsilon\right)\mathbb{E}_{z\sim\nu}\left[e^{-2^{p-1}c(\overline{x},z)/\epsilon}\right].$$

The proof is completed. □

Proof of Proposition EC.1. One can see that for any $x \in \operatorname{supp}\widehat{\mathbb{P}}$, it holds that

$$\begin{aligned}
&\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right] \\
=&\mathbb{E}_{z\sim\nu}\left[e^{f(z)/(\lambda\epsilon)}\frac{e^{-c(x,z)/\epsilon}}{\mathbb{E}_{u\sim\nu}\left[e^{-c(x,u)/\epsilon}\right]}\right] \\
\leq&\mathbb{E}_{z\sim\nu}\left[e^{f(z)/(\lambda\epsilon)}\frac{e^{-c(x,z)/\epsilon}}{\mathbb{E}_{u\sim\nu}\left[e^{-2^{p-1}c(\overline{x},u)/\epsilon}\right]}\right] \\
\leq&\mathbb{E}_{z\sim\nu}\left[e^{f(z)/(\lambda\epsilon)}\frac{e^{-2^{1-p}c(\overline{x},z)/\epsilon}e^{c(x,\overline{x})/\epsilon}}{\mathbb{E}_{u\sim\nu}\left[e^{-2^{p-1}c(\overline{x},u)/\epsilon}\right]}\right] \\
=&\frac{e^{c(x,\overline{x})(1+2^{p-1})/\epsilon}}{\mathbb{E}_{u\sim\nu}\left[e^{-2^{p-1}c(\overline{x},u)/\epsilon}\right]}\mathbb{E}_{z\sim\nu}\left[e^{f(z)/(\lambda\epsilon)}e^{-2^{1-p}c(\overline{x},z)/\epsilon}\right],
\end{aligned}$$

where the first inequality is based on the lower bound in Lemma EC.1, the second inequality is based on the triangular inequality $c(x,z) \geq 2^{1-p}c(\overline{x},z) - c(x,\overline{x})$. Note that almost surely for all $x \in \operatorname{supp}\widehat{\mathbb{P}}$, $c(x,\overline{x}) < \infty$. Moreover,

$$0 < \mathbb{E}_{z\sim\nu}\left[e^{-2^{p-1}c(\overline{x},z)/\epsilon}\right] \leq \mathbb{E}_{z\sim\nu}\left[e^{-c(\overline{x},z)/\epsilon}\right] < \infty,$$

where the lower bound is because $c(\overline{x}, z) < \infty$ almost surely for all z, the upper bound is because $c(\overline{x}, z) \geq 0$ almost surely for all z. Based on these observations, we have that

$$\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right] \leq \frac{e^{c(x,\overline{x})(1+2^{p-1})/\epsilon}}{\mathbb{E}_{z\sim\nu}\left[e^{-2^{p-1}c(\overline{x},z)/\epsilon}\right]}\mathbb{E}_{z\sim\nu}\left[e^{f(z)/(\lambda\epsilon)}e^{-2^{1-p}c(\overline{x},z)/\epsilon}\right] < \infty$$

almost surely for all $x \sim \widehat{\mathbb{P}}$. □

EC.4. Proofs of Technical Results in Section 3.2 and 3.3

Proof of Remark 3. Recall the dual objective function in (1) is

$$v(\lambda;\epsilon)=\lambda\rho+\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\lambda\epsilon\,\log\mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))/(\lambda\epsilon)}\big]\Big].$$

We take limit for the second term in $v(\lambda;\epsilon)$ to obtain:

$$\begin{aligned}
&\lim_{\epsilon\to 0}\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\lambda\epsilon\,\log\mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))/(\lambda\epsilon)}\big]\Big]\\
=&\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lim_{\beta\to\infty}\frac{\lambda}{\beta}\log\mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))\beta/\lambda}\big]\right]\\
=&\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lim_{\beta\to\infty}\lambda\nabla_\beta\log\mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))\beta/\lambda}\big]\right]\\
=&\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lim_{\beta\to\infty}\frac{\mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))\beta/\lambda}\big[f(z)-\lambda c(x,z)\big]\big]}{\mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))\beta/\lambda}\big]}\right]\\
=&\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\sup_{z\in\operatorname{supp}\nu}\big\{f(z)-\lambda c(x,z)\big\}\Big].
\end{aligned}$$

Particularly, when $\operatorname{supp}\nu=\mathcal{Z}$, it holds that

$$\sup_{z\in\operatorname{supp}\nu}\big\{f(z)-\lambda c(x,z)\big\}=\sup_{z\in\mathcal{Z}}\big\{f(z)-\lambda c(x,z)\big\}$$

and in this case the dual objective function of the Sinkhorn DRO problem converges into that of the Wasserstein DRO problem. □

Proof of Example 2. In this example, the dual objective becomes

$$V_{\mathrm{D}}=\inf_{\lambda\geq 0}\left\{\lambda\overline{\rho}+\mathbb{E}_{(a,b)\sim\widehat{\mathbb{P}}}\left[\lambda\epsilon\log\mathbb{E}_{a'\sim\mathcal{N}(a,\epsilon I_d)}\left[\exp\left(\frac{(\theta^{\mathrm{T}}a'-b)^2}{\lambda\epsilon}\right)\right]\right]\right\}. \tag{EC.5}$$

Specially, for any $a\in\mathbb{R}^d, b\in\mathbb{R}, \theta\in\mathbb{R}^d$, it holds that

$$\begin{aligned}
&\lambda\epsilon\log\left(\mathbb{E}_{a'\sim\mathcal{N}(a,\epsilon I_d)}\,\exp\left(\frac{(\theta^{\mathrm{T}}a'-b)^2}{\lambda\epsilon}\right)\right)\\
=&\lambda\epsilon\log\left(\mathbb{E}_{\Delta_a\sim\mathcal{N}(0,I_d)}\,\exp\left(\frac{\big[(\theta^{\mathrm{T}}a-b)+(\sqrt{\epsilon}\theta)^{\mathrm{T}}\Delta_a\big]^2}{\lambda\epsilon}\right)\right)\\
=&(\theta^{\mathrm{T}}a-b)^2+\lambda\epsilon\log\left(\underbrace{\mathbb{E}_{\Delta_a\sim\mathcal{N}(0,I_d)}\,\exp\left(\frac{\epsilon(\theta^{\mathrm{T}}\Delta_a)^2-2(b-\theta^{\mathrm{T}}a)\sqrt{\epsilon}\theta^{\mathrm{T}}\Delta_a}{\lambda\epsilon}\right)}_{\text{(I)}}\right).
\end{aligned}$$

Note that the term (I) can be simplified using integral of exponential functions:

$$\begin{aligned}
\text{(I)}&=(2\pi)^{-d/2}\int_{\mathbb{R}^d}\exp\left(-\frac{1}{2}\Delta_a^{\mathrm{T}}\Delta_a+\frac{(\theta^{\mathrm{T}}\Delta_a)^2}{\lambda}-2\frac{(b-\theta^{\mathrm{T}}a)\theta^{\mathrm{T}}}{\lambda\sqrt{\epsilon}}\Delta_a\right)\mathrm{d}\Delta_a\\
&=(2\pi)^{-d/2}\int_{\mathbb{R}^d}\exp\left(-\frac{1}{2}\Delta_a^{\mathrm{T}}A\Delta_a+J^{\mathrm{T}}\Delta_a\right)\mathrm{d}\Delta_a,
\end{aligned}$$

where the matrix $A = I - \frac{2\theta\theta^{\mathrm{T}}}{\lambda}$ and vector $J = 2\frac{(\theta^{\mathrm{T}}a-b)\theta}{\lambda\sqrt{\epsilon}}$. As a consequence, when $\|\theta\|_2^2 < \frac{\lambda}{2}$, it holds that

$$\begin{aligned}\text{(I)} &= \det(A)^{-1/2}\exp\left(\frac{1}{2}J^{\mathrm{T}}A^{-1}J\right)\\ &= \det\left(I-\frac{2\theta\theta^{\mathrm{T}}}{\lambda}\right)^{-1/2}\exp\left(2\frac{(\theta^{\mathrm{T}}a-b)^2}{\lambda^2\epsilon}\theta^{\mathrm{T}}A^{-1}\theta\right).\end{aligned}$$

Finally, we arrive at

$$\begin{aligned}&\lambda\epsilon\log\left(\mathbb{E}_{a'\sim\mathcal{N}(a,\epsilon I_d)}\ \exp\left(\frac{(\theta^{\mathrm{T}}a'-b)^2}{\lambda\epsilon}\right)\right)\\ =&(\theta^{\mathrm{T}}a-b)^2+\frac{(\theta^{\mathrm{T}}a-b)^2}{\frac{1}{2}\lambda\|\theta\|_2^{-2}-1}-\frac{\lambda\epsilon}{2}\log\det\left(I-\frac{2\theta\theta^{\mathrm{T}}}{\lambda}\right),\quad \text{if } \|\theta\|_2^2<\frac{\lambda}{2}.\end{aligned}$$

Substituting this expression into (EC.5) gives the desired result. □

Proof of Corollary 1. We now introduce the epi-graphical variables $s_i, i = 1,\ldots,n$ to reformulate V_{D} as

$$V_{\mathrm{D}} = \left\{\begin{array}{ll}\inf\limits_{\lambda\geq 0, s_i} & \lambda\overline{\rho}+\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}s_i\\ \text{s.t.} & \lambda\epsilon\log\mathbb{E}_{z\sim\mathbb{Q}_{\hat{x}_i,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\leq s_i,\forall i\end{array}\right.$$

For fixed i, the i-th constraint can be reformulated as

$$\begin{aligned}&\left\{\exp\left(\frac{s_i}{\lambda\epsilon}\right)\geq\mathbb{E}_{z\sim\mathbb{Q}_{\hat{x}_i,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\right\}\\ =&\left\{1\geq\mathbb{E}_{z\sim\mathbb{Q}_{\hat{x}_i,\epsilon}}\left[e^{(f(z)-s_i)/(\lambda\epsilon)}\right]\right\}\\ =&\left\{\lambda\epsilon\geq\mathbb{E}_{z\sim\mathbb{Q}_{\hat{x}_i,\epsilon}}\left[\lambda\epsilon e^{(f(z)-s_i)/(\lambda\epsilon)}\right]\right\}\\ =&\left\{\lambda\epsilon\geq\sum_{\ell=1}^{L_{\max}}\mathbb{Q}_{\hat{x}_i,\epsilon}(z_\ell)a_{i,\ell}\right\}\bigcap\left\{a_{i,\ell}\geq\lambda\epsilon\exp\left(\frac{f(z_\ell)-s_i}{\lambda\epsilon}\right),\forall\ell\right\},\end{aligned}$$

where the second constraint set can be formulated as

$$(\lambda\epsilon, a_{i,\ell}, f(z_\ell)-s_i)\in\mathcal{K}_{\exp}.$$

Substituting this expression into V_{D} completes the proof. □

EC.5. Proofs of Technical Results in Section 3.4

We rely on the following technical lemma to derive our strong duality result.

LEMMA EC.2. *For fixed τ and a reference measure $\nu \in \mathcal{M}(\mathcal{Z})$, consider the optimization problem*

$$v(\tau) = \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{z\sim\mathbb{P}} \left[f(z) - \tau \log\left(\frac{\mathrm{d}\mathbb{P}(z)}{\mathrm{d}\nu(z)} \right) \right] \right\}. \tag{EC.6}$$

Suppose there exists a probability measure $\mathbb{Q} \in \mathcal{P}(\mathcal{Z})$ such that $\mathbb{Q} \ll \nu$.

(I) *When $\tau = 0$,*

$$v(0) = \operatorname*{ess\,sup}_{\nu}(f) \triangleq \inf\{t \in \mathbb{R} : \ \nu\{f(z) > t\} = 0\}.$$

(II) *When $\tau > 0$ and*

$$\mathbb{E}_{z\sim\nu}\left[e^{f(z)/\tau}\right] < \infty,$$

it holds that

$$v(\tau) = \tau \log\left(\mathbb{E}_{z\sim\nu}\left[e^{f(z)/\tau}\right]\right),$$

and $\lim_{\tau\downarrow 0} v(\tau) = v(0)$. The optimal solution in (EC.6) has the expression

$$\mathrm{d}\mathbb{P}(z) = \frac{e^{f(z)/\tau}}{\mathbb{E}_{u\sim\nu}\left[e^{f(u)/\tau}\right]}\,\mathrm{d}\nu(z).$$

(III) *When $\tau > 0$ and*

$$\mathbb{E}_{z\sim\nu}\left[e^{f(z)/\tau}\right] = \infty,$$

we have that $v(\tau) = \infty$.

Proof of Lemma EC.2. It can be shown that $v(\tau)$ can be reformulated as a regularized linear optimization problem, where the reference measure in the relative entropy regularization is a probability measure:

$$v(\tau) = \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{z\sim\mathbb{P}} \left[g(z) - \tau \log\left(\frac{\mathrm{d}\mathbb{P}(z)}{\mathrm{d}\mathbb{Q}(z)} \right) \right] \right\}, \quad \text{where } g(z) = f(z) - \tau \log\left(\frac{\mathrm{d}\mathbb{Q}(z)}{\mathrm{d}\nu(z)} \right).$$

Also, we reformulate $v(\tau)$ based on the importance sampling trick:

$$v(\tau) = \sup_{L:\ L\geq 0} \left\{ \mathbb{E}_{z\sim\mathbb{Q}}\left[g(z)L(z) - \tau L(z)\log L(z) \right] : \ \mathbb{E}_{z\sim\mathbb{Q}}[L(z)] = 1 \right\}.$$

Then the remaining part follows the discussion in [62, Section 2.1]. □

Proof of Lemma 2. Recall from (6) that

$$V = \sup_{\{\gamma_x\}_{x\in\operatorname{supp}\widehat{\mathbb{P}}}\subset\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}[f(z)] : \ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}\left[c(x,z) + \epsilon \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)} \right) \right] \leq \rho \right\}.$$

Based on the change-of-measure identity $\log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)}\right) = \log\left(\frac{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)}{\mathrm{d}\nu(z)}\right) + \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)}\right)$ and the expression of $\mathbb{Q}_{x,\epsilon}$, the constraint can be reformulated as

$$\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}\left[c(x,z) + \epsilon \log\left(\frac{e^{-c(x,z)/\epsilon}}{\int e^{-c(x,u)/\epsilon}\,\mathrm{d}\nu(u)} \right) + \epsilon \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)} \right) \right] \leq \rho.$$

Combining the first two terms within the expectation term and substituting the expression of $\overline{\rho}$, it is equivalent to

$$\epsilon \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}\left[\log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)}\right)\right] \le \overline{\rho}.$$

In summary, the primal problem (Primal) can be reformulated as a generalized KL-divergence DRO problem

$$V = \sup_{\{\gamma_x\}_{x\in\mathrm{supp}\,\widehat{\mathbb{P}}}\subset\mathcal{P}(\mathcal{Z})}\left\{\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}[f(z)]:\ \epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}\left[\log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)}\right)\right]\le\overline{\rho}\right\}. \qquad \square$$

In the remaining of this subsection, we provide the full proof of Theorem 1. We first show that the dual minimizer exists.

LEMMA EC.3 **(Existence of Dual Minimizer)**. *Suppose $\overline{\rho}>0$ and Condition 1 is satisfied, then the dual minimizer λ^* exists, which either equals to 0 or satisfies Condition 1.*

Proof of Lemma EC.3. We first show that $\lambda^*<\infty$. Denote by $v(\lambda)$ the objective function for the dual problem:

$$v(\lambda)=\lambda\overline{\rho}+\lambda\epsilon\,\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\right].$$

The integrability condition for the dominated convergence theorem is satisfied, which implies

$$\begin{aligned}
&\lim_{\lambda\to\infty}\lambda\epsilon\,\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\right]\\
=&\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lim_{\beta\to0}\frac{\epsilon}{\beta}\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{\beta f(z)/\epsilon}\right]\right]\\
=&\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lim_{\beta\to0}\epsilon\nabla_\beta\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{\beta f(z)/\epsilon}\right]\right]\\
=&\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lim_{\beta\to0}\frac{\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[f(z)e^{\beta f(z)/\epsilon}\right]}{\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{\beta f(z)/\epsilon}\right]}\right]\\
=&\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[f(z)],
\end{aligned}$$

where the first equality follows from the change-of-variable technique with $\beta=1/\lambda$, the second equality follows from the definition of derivative, the third and the last equality follows from the dominated convergence theorem. As a consequence, as long as $\overline{\rho}>0$, we have

$$\lim_{\lambda\to\infty}v(\lambda)=\infty.$$

We can take λ satisfying Condition 1 and then $v(\lambda)<\infty$. This, toegther with the fact that $v(\cdot)$ is continuous, guarantees the existence of the dual minimizer. Hence $\lambda^*<\infty$, which implies that either $\lambda^*=0$ or λ^* satisfies Condition 1. $\square$

Next, we establish first-order optimality condition for cases $\lambda^*>0$ or $\lambda^*=0$, corresponding to whether the Sinkhorn distance constraint in (Primal) is binding or not. Lemma EC.4 below presents a necessary and sufficient condition for the dual minimizer $\lambda^*=0$, corresponding to the case where the Sinkhorn distance constraint in (Primal) is not binding.

LEMMA EC.4 **(Necessary and Sufficient Condition for $\lambda^*=0$)**. *Suppose $\overline{\rho}>0$ and Condition 1 is satisfied, then the dual minimizer $\lambda^*=0$ if and only if all the following conditions hold:*

(I) $\operatorname{ess\,sup}_\nu\ f\triangleq\inf\{t:\nu\{f(z)>t\}=0\}<\infty$.

(II) $\overline{\rho}'=\overline{\rho}+\epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)]\right]\ge0$, *where* $A:=\{z:\ f(z)=\operatorname{ess\,sup}_\nu\ f\}$.

Recall that we have the convention that the dual objective evaluated at $\lambda = 0$ equals $\operatorname{ess\,sup}_\nu f$. Thus Condition (I) ensures that the dual objective function evaluated at the minimizer is finite. When the minimizer $\lambda^* = 0$, the Sinkhorn ball should be large enough to contain at least one distribution with objective value $\operatorname{ess\,sup}_\nu f$, and Condition (II) characterizes the lower bound of $\overline{\rho}$.

Proof of Lemma EC.4. Suppose the dual minimizer $\lambda^* = 0$, then taking the limit of the dual objective function gives

$$\lim_{\lambda \to 0} v(\lambda) = \mathbb{E}_{x \sim \widehat{\mathbb{P}}}\left[H^u(x)\right] < \infty,$$

where

$$H^u(x) := \inf\{t : \mathbb{Q}_{x,\epsilon}\{f(z) > t\} = 0\} \triangleq \operatorname*{ess\,sup}_{\mathbb{Q}_{x,\epsilon}} f.$$

For notational simplicity we take $H^u = \operatorname*{ess\,sup}_\nu f$. One can check that $H^u(x) \equiv H^u$ for any $x \in \operatorname{supp} \widehat{\mathbb{P}}$: for any t so that $\mathbb{Q}_{x,\epsilon}\{f(z) > t\} = 0$, we have that

$$\mathbb{E}_{z \sim \nu}\left[1\{f(z) > t\} e^{-c(x,z)/\epsilon}\right] = 0,$$

which, together with the fact that $\nu\{c(x,z) < \infty\} = 1$ for fixed x, implies

$$\mathbb{E}_{z \sim \nu}[1\{f(z) > t\}] = 0.$$

On the contrary, for any t so that $\nu\{f(z) > t\} = 0$, we have that

$$0 \le \mathbb{E}_{z \sim \nu}\left[1\{f(z) > t\} e^{-c(x,z)/\epsilon}\right] \le \mathbb{E}_{z \sim \nu}[1\{f(z) > t\}] = 0,$$

where the second inequality is because that $\nu\{c(x,z) \ge 0\} = 1$. As a consequence, $\mathbb{Q}_{x,\epsilon}\{f(z) > t\} = 0$. Hence we can assert that $H^u(x) = H^u$ for all $x \in \operatorname{supp} \widehat{\mathbb{P}}$, which implies

$$\lim_{\lambda \to 0} v(\lambda) = H^u < \infty.$$

Then we show that almost surely for all x,

$$\mathbb{E}_{z \sim \mathbb{Q}_{x,\epsilon}}[1_A(z)] > 0, \quad \text{where } A = \{z : f(z) = H^u\}.$$

Denote by D the collection of samples x so that $\mathbb{E}_{z \sim \mathbb{Q}_{x,\epsilon}}[1_A(z)] = 0$. Assume the condition above does not hold, which means that $\widehat{\mathbb{P}}\{D\} > 0$. For any $\tau > 0$ and $x \in D$, there exists $H^l(x) < H^u$ such that

$$0 < \mathfrak{h}_x := \mathbb{E}_{z \sim \mathbb{Q}_{x,\epsilon}}[1_{B(x)}(z)] \le \tau, \quad \text{where } B(x) = \{z : H^l(x) \le f(z) \le H^u\}.$$

Define $H^{\mathrm{gap}}(x) = H^u - H^l(x)$, $\mathfrak{h}_x^c = 1 - \mathfrak{h}_x$. Then we find that for $x \in D$,

$$\begin{aligned} v_x(\lambda) &= \lambda\epsilon \log\left(\mathbb{E}_{z \sim \mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)} 1_{B(x)}(z)\right] + \mathbb{E}_{z \sim \mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)} 1_{B(x)^c}(z)\right]\right) \\ &\le H^u + \lambda\epsilon \log\left(\mathfrak{h}_x + e^{-H^{\mathrm{gap}}(x)/(\lambda\epsilon)} \mathfrak{h}_x^c\right). \end{aligned}$$

Since $\widehat{\mathbb{P}}\{D\} > 0$, the dual objective function for $\lambda > 0$ is upper bounded as

$$\begin{aligned} v(\lambda) &= \lambda\overline{\rho} + \mathbb{E}_{x \sim \widehat{\mathbb{P}}}[v_x(\lambda)] \\ &\le H^u + \lambda\overline{\rho} + \lambda\epsilon \mathbb{E}_{x \sim \widehat{\mathbb{P}}}\left[\log\left(\mathfrak{h}_x + e^{-H^{\mathrm{gap}}(x)/(\lambda\epsilon)} \mathfrak{h}_x^c\right) 1_D(x)\right]. \end{aligned}$$

We can see that

$$\lim_{\lambda \to 0} \lambda\overline{\rho} + \lambda\epsilon \mathbb{E}_{x \sim \widehat{\mathbb{P}}}\left[\log\left(\mathfrak{h}_x + e^{-H^{\mathrm{gap}}(x)/(\lambda\epsilon)} \mathfrak{h}_x^c\right) 1_D(x)\right] = 0,$$

and

$$\begin{aligned}&\lim_{\lambda\to 0}\ \nabla\left[\lambda\overline{\rho}+\lambda\epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\left(\mathfrak{h}_x+e^{-H^{\mathrm{gap}}(x)/(\lambda\epsilon)}\mathfrak{h}_x^c\right)1_D(x)\right]\right]\\ =&\overline{\rho}+\epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\left(\mathfrak{h}_x\right)1_D(x)\right]\\ \leq&\overline{\rho}+\epsilon\log(\tau)\widehat{\mathbb{P}}\{D\}\leq-\overline{\rho}<0,\end{aligned}$$

where the second inequality is by taking the constant $\tau=\exp\left(-\frac{2\overline{\rho}}{\epsilon\widehat{\mathbb{P}}\{D\}}\right)$. Hence, there exists $\overline{\lambda}>0$ such that

$$v(\overline{\lambda})\leq H^u+\overline{\lambda}\overline{\rho}+\overline{\lambda}\epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\left(\mathfrak{h}_x+e^{-H^{\mathrm{gap}}(x)/(\overline{\lambda}\epsilon)}\mathfrak{h}_x^c\right)1_D(x)\right]<v(0),$$

which contradicts to the optimality of $\lambda^*=0$. As a result, almost surely for all x, we have that

$$\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)]>0.$$

To show the second condition, we re-write the dual objective function for $\lambda>0$ as

$$v(\lambda)=\lambda\overline{\rho}+\lambda\epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\left(\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)]+\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{[f(z)-H^u]/(\lambda\epsilon)}1_{A^c}(z)\right]\right)\right]+H^u.$$

The gradient of $v(\lambda)$ becomes

$$\begin{aligned}\nabla v(\lambda)=&\overline{\rho}+\epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\left(\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)]+\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{[f(z)-H^u]/(\lambda\epsilon)}1_{A^c}(z)\right]\right)\right]\\&+\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\frac{\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{[f(z)-H^u]/(\lambda\epsilon)}1_{A^c}(z)(H^u-f(z))/(\lambda)\right]}{\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)]+\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{[f(z)-H^u]/(\lambda\epsilon)}1_{A^c}(z)\right]}\right].\end{aligned}$$

We can see that $\lim_{\lambda\to\infty}\nabla v(\lambda)=\overline{\rho}$. Take

$$v_{1,x}(\lambda)=\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{[f(z)-H^u]/(\lambda\epsilon)}1_{A^c}(z)\right].$$

Then $\lim_{\lambda\to 0}v_{1,x}(\lambda)=0$ and $v_{1,x}(\lambda)\geq 0$. Take

$$v_{2,x}(\lambda)=\frac{\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{[f(z)-H^u]/(\lambda\epsilon)}1_{A^c}(z)(H^u-f(z))/(\lambda)\right]}{\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)]+\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{[f(z)-H^u]/(\lambda\epsilon)}1_{A^c}(z)\right]}.$$

Then $\lim_{\lambda\to 0}v_{2,x}(\lambda)=0$ and $v_{2,x}(\lambda)\geq 0$. It follows that

$$\lim_{\lambda\to 0}\nabla v(\lambda)=\overline{\rho}+\epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)]\right]=\overline{\rho}'.$$

Hence, if the last condition is violated, based on the mean value theorem, we can find $\overline{\lambda}>0$ so that $\nabla v(\overline{\lambda})=0$, which contradicts to the optimality of $\lambda^*=0$.

Now we show the converse direction. For any $\lambda>0$, we find that

$$\nabla v(\lambda)=\overline{\rho}+\epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\left(\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)]+v_{1,x}(\lambda)\right)\right]+\mathbb{E}_{x\sim\widehat{\mathbb{P}}}[v_{2,x}(\lambda)].$$

For fixed x, when $\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}[1_A]=1$, we can see that $v_{1,x}(\lambda)=v_{2,x}(\lambda)=0$, then

$$\overline{\rho}+\epsilon\left[\log\left(\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)]+v_{1,x}(\lambda)\right)\right]+v_{2,x}(\lambda)=\overline{\rho}>0.$$

When $\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)]\in(0,1)$, we can see that $v_{1,x}(\lambda)>0, v_{2,x}(\lambda)>0$. Then

$$\overline{\rho}+\epsilon\left[\log\left(\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)]+v_{1,x}(\lambda)\right)\right]+v_{2,x}(\lambda)>\overline{\rho}+\epsilon\log(\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)])=\overline{\rho}'\geq 0.$$

Therefore, $\nabla v(\lambda)>0$ for any $\lambda>0$. By the convexity of $v(\lambda)$, we conclude that the dual minimizer $\lambda^*=0$. □

The first-order optimality condition for $\lambda^* > 0$ is stated in Lemma 3. The following provides the proof of this technical lemma.

Proof of Lemma 3. Since $\lambda^* > 0$, based on the optimality condition of the dual problem, we have that

$$0 = \nabla_\lambda \left[\lambda\overline{\rho} + \lambda\epsilon\,\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\right]\right]\Big|_{\lambda=\lambda^*}.$$

Or equivalently, we have that

$$\overline{\rho} + \epsilon\,\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda^*\epsilon)}\right]\right] - \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\frac{\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda^*\epsilon)} f(z)\right]}{\lambda^* \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda^*\epsilon)}\right]}\right] = 0.$$

Re-arranging the term completes the proof. □

Proof of Theorem 1. Recall the feasibility result in Theorem 1(I) can be easily shown by considering the reformulation of V in Lemma 2 and the non-negativity of KL-divergence. When $\overline{\rho} = 0$, one can see that

$$\begin{aligned} V_{\mathrm{D}} &= \inf_{\lambda\geq 0}\ \left\{\lambda\epsilon\,\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\right]\right\} \\ &\leq \lim_{\lambda\to\infty}\ \lambda\epsilon\,\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\right] = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[f(z)] = V. \end{aligned}$$

Therefore, the strong duality result holds in this case. It remains to show the strong duality result for $\overline{\rho} > 0$, which can be further separated to two cases: Condition 1 holds or not.

- When Condition 1 holds, by Lemma EC.3, the dual minimizer λ^* exists. The proof for $\lambda^* > 0$ can be found in main context. When $\lambda^* = 0$, the optimality condition in Lemma EC.4 holds. We construct the primal (approximate) solution $\mathbb{P}_* = \mathrm{Proj}_{2\#}\gamma_*$, where γ_* satisfies

$$\mathrm{d}\gamma_*(x,z) = \mathrm{d}\gamma_*^x(z)\,\mathrm{d}\widehat{\mathbb{P}}(x), \quad \text{where } \mathrm{d}\gamma_*^x(y) = \begin{cases} 0, & \text{if } z\notin A, \\ \frac{e^{-c(x,z)/\epsilon}\,\mathrm{d}\nu(z)}{\mathbb{E}_{u\sim\nu}\left[e^{-c(x,u)/\epsilon}1_A\right]}, & \text{if } z\in A. \end{cases}$$

  We can verify easily that the primal solution is feasible based on the optimality condition $\overline{\rho}' \geq 0$ in Lemma EC.4. Moreover, we can check that the primal optimal value is lower bounded by the dual optimal value:

$$V \geq \mathbb{E}_{(x,z)\sim\gamma_*}[f(z)] = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_*^x}[f(z)] = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_*^x}\left[\operatorname*{ess\,sup}_{\nu} f\right] = \operatorname*{ess\,sup}_{\nu} f = V_{\mathrm{D}},$$

  where the second equality is because that $z\in A$ so that $f(z) = \operatorname*{ess\,sup}_{\nu} f$. This, together with the weak duality result, completes the proof in this part.
- When Condition 1 does not hold, we consider a sequence of real numbers $\{R_j\}_j$ such that $R_j\to\infty$ and take the objective function $f_j(z) = f(z)1\{f(z)\leq R_j\}$. Hence, there exists $\lambda > 0$ satisfying $\Pr_{x\sim\widehat{\mathbb{P}}}\left\{x:\ \mathbb{E}_{\mathbb{Q}_{x,\epsilon}}\left[e^{f_j(z)/(\lambda\epsilon)}\right] = \infty\right\} = 0$. According to the necessary condition in Lemma EC.4, the corresponding dual minimizer $\lambda_j^* > 0$ for sufficiently large index j. Then we can apply the duality result in the first part of Theorem 1(III) to show that for sufficiently large j, it holds that

$$\sup_{\mathbb{P}\in\mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})}\ \{\mathbb{E}_{z\sim\mathbb{P}}[f_j(z)]\} \geq \lambda_j^*\overline{\rho} + \lambda_j^*\epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_j(z)/(\lambda\epsilon)}\right]\right].$$

  Taking $j\to\infty$ both sides implies that $V = \infty$. □

EC.6. Preliminaries on Stochastic Mirror Descent (SMD)

In this section, we provide some preliminaries on SMD that can be useful for proving Theorem 2. Consider the minimization of the objective function $F(\theta)$ with $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$. In particular, we assume the constraint set Θ is non-empty, closed and convex. We also impose the following assumption regarding the (sub-)gradient oracles when using the SMD algorithm:

ASSUMPTION EC.1 **(Stochastic Oracles of Gradient Estimate)**. *The objective function $F(\theta)$ is convex in θ, and we have the stochastic oracle such that for given θ we can generate a stochastic vector $G(\theta,\xi)$ such that $\mathbb{E}_\xi[G(\theta,\xi)] \in \partial F(\theta)$, where $\partial F(\theta)$ is the subdifferential set of $F(\cdot)$ at θ. Also, suppose there exists a constant $M_* > 0$ so that*

$$\mathbb{E}_\xi\left[\|G(\theta,\xi)\|_*^2\right] \le M_*^2, \quad \forall \theta \in \Theta.$$

Under the above assumption, the SMD algorithm generates the following iteration:

$$\theta_{t+1} = \mathrm{Prox}_{\theta_t}\left(\gamma_t G(\theta_t,\xi^t)\right), \quad \theta_1 \in \Theta, \quad t = 1,\ldots,T-1.$$

For simplicity of discussion, we employ constant step size policy $\gamma_t := \gamma$ for $t = 1,\ldots,T-1$. The following presents convergence results of the SMD algorithm. Similar results can be found in [105, 53, 81, 97].

LEMMA EC.5 **(SMD for Nonsmooth Convex Optimization)**. *Under Assumption EC.1, let the estimation of optimal solution at the iteration j be*

$$\tilde{\theta}_{1:j} = \frac{1}{j}\sum_{t=1}^{j}\theta_t.$$

When taking constant step size

$$\gamma = \sqrt{\frac{2\kappa V(\theta_1,\theta^*)}{TM_*^2}},$$

it holds that

$$\mathbb{E}\left[F(\tilde{\theta}_{1:T}) - F(\theta^*)\right] \le M_*\sqrt{\frac{2V(\theta_1,\theta^*)}{\kappa T}}.$$

Proof of Lemma EC.5. The proof follows from [81, Section 2.3]. □

EC.7. Proofs of Technical Results in Section 4.2.1

Since any two norms on a finite-dimensional vector space are equivalent, we impose the following assumption throughout Section 4 without loss of generality:

ASSUMPTION EC.2. *There exists* $\mathfrak{c}$ *and* $\mathfrak{d}$ *such that*

$$\mathfrak{c}\|\cdot\|_2 \le \|\cdot\| \le \mathfrak{d}\|\cdot\|_2.$$

Based on Assumption EC.2, we are also able to obtain the bound regarding the dual norm $\|\cdot\|_*$:

$$\frac{1}{\mathfrak{d}}\|\cdot\|_2 \le \|\cdot\|_* \le \frac{1}{\mathfrak{c}}\|\cdot\|_2.$$

We quantify the sample complexity for generating a single gradient estimator as the (expected) number of queries for generating the random sampling parameter (x,z) with $x\sim\widehat{\mathbb{P}}, z\sim\mathbb{Q}_{x,\epsilon}$. The complexity result of proposed gradient estimators is summarized in Remark EC.1.

REMARK EC.1 (SAMPLE COMPLEXITY OF GRADIENT ESTIMATORS). The complexity for generating SG gradient estimator $v^{\text{SG}}(\theta)$ is $\mathcal{O}(n_L^{\circ}2^L)$. The complexity for generating RT-MLMC gradient estimator $v^{\text{RT-MLMC}}(\theta)$ is $\mathcal{O}(n_L^{\circ}L)$.

Proof of Remark EC.1. Since SG estimator requires generating $g^L(\theta,\zeta_i^L)$ for n_L° times, and generating a single $g^L(\theta,\zeta_i^L)$ requires generating the random sampling parameters $\{z_j^\ell\}_{j\in[2^L]}$ of size 2^L, we imply the complexity of SG estimator is $O(n_L^{\circ}2^L)$.

Define $q_\ell = \frac{2^{-\ell}}{2-2^{-L}}, \ell = 0,1,\ldots,L$. The complexity of RT-MLMC estimator can be bounded as

$$\mathcal{O}\left(n_L^{\circ}\sum_{\ell=0}^{L} q_\ell 2^\ell\right) = \mathcal{O}\left(n_L^{\circ}\sum_{\ell=0}^{L}\frac{2^{-\ell}}{2-2^{-L}}2^\ell\right) = \mathcal{O}(n_L^{\circ}L). \qquad \square$$

Next, we present some basic properties regarding the approximation function $F^\ell(\theta)$ defined in (12) in Lemma EC.6, which can be used to show Theorem 2. The proof of this technical lemma follows from [59, Lemma 3.1] and [60, Proposition 4.1].

LEMMA EC.6. (I) *Under Assumption 2(II), it holds that*

$$\left|F^\ell(\theta) - F(\theta)\right| \le \lambda\epsilon\exp\left(2B/(\lambda\epsilon)\right)\cdot 2^{-(\ell+1)}, \qquad \forall\theta\in\Theta.$$

(II) *Under Assumption 2(II) and 2(III), it holds that*

$$\left\|\nabla F^\ell(\theta) - \nabla F(\theta)\right\|_2^2 \le L_f^2\exp\left(4B/(\lambda\epsilon)\right)\cdot 2^{-\ell}, \qquad \forall\theta\in\Theta.$$

(III) *Under Assumption 2(III), it holds that*

$$\mathbb{E}\left[\left\|g^\ell(\theta,\zeta^\ell)\right\|_2^2\right] \le L_f^2, \qquad \forall\theta\in\Theta.$$

Additionally when Assumption 2(II) holds, it holds that

$$\mathbb{E}\left[\left\|G^\ell(\theta,\zeta^\ell)\right\|_2^2\right] \le L_f^2\exp\left(4B/(\lambda\epsilon)\right)\cdot 2^{-\ell}, \qquad \forall\theta\in\Theta.$$

EC.7.1. Proof of Theorem 2

We first discuss sample complexity for nonsmooth convex optimization. Suppose for a given θ, the gradient estimate of $F(\theta)$, denoted as $v(\theta)$, satisfies

$$\mathbb{E}[v(\theta)] = \nabla \overline{F}(\theta), \quad \mathbb{E}\left[\|v(\theta)\|_*^2\right] \le M_*^2.$$

Assume the bias of objective satisfies

$$\Delta_F := \sup_{\theta \in \Theta} |\overline{F}(\theta) - F(\theta)|.$$

Denote by $\overline{\theta}^*$ an global optimum of $\overline{F}$. Then we have the following result.

PROPOSITION EC.2 **(BSMD for Nonsmooth Convex Optimization)**. *When taking the step size* $\gamma = \sqrt{\frac{2\kappa V(\theta_1, \overline{\theta}^*)}{TM_*^2}}$, *it holds that*

$$\mathbb{E}\big[F(\tilde{\theta}_{1:T}) - F(\theta^*)\big] \le 2\Delta_F + M_* \sqrt{\frac{2V(\theta_1, \overline{\theta}^*)}{\kappa T}}.$$

Proof of Proposition EC.2. Note that we can establish the following error bound:

$$\begin{aligned}
\mathbb{E}\big[F(\tilde{\theta}_{1:T}) - F(\theta^*)\big] &= \mathbb{E}\big[F(\tilde{\theta}_{1:T}) - \overline{F}(\tilde{\theta}_{1:T})\big] + \mathbb{E}\big[\overline{F}(\tilde{\theta}_{1:T}) - \overline{F}(\theta^*)\big] + \mathbb{E}\big[\overline{F}(\theta^*) - F(\theta^*)\big] \\
&\le 2\Delta_F + \mathbb{E}\big[\overline{F}(\tilde{\theta}_{1:T}) - \overline{F}(\theta^*)\big] \\
&\le 2\Delta_F + \mathbb{E}\big[\overline{F}(\tilde{\theta}_{1:T}) - \overline{F}(\overline{\theta}^*)\big],
\end{aligned}$$

where the first inequality is due to the bias approximation error bound, and the second inequality is due to the sub-optimality of θ^* for the objective $\overline{F}$. According to Lemma EC.5, if we take the step size $\gamma = \sqrt{\frac{2\kappa V(\theta_1, \overline{\theta}^*)}{TM_*^2}}$, then it holds that

$$\mathbb{E}\big[\overline{F}(\tilde{\theta}_{1:T}) - \overline{F}(\overline{\theta}^*)\big] \le M_* \sqrt{\frac{2V(\theta_1, \overline{\theta}^*)}{\kappa T}}. \qquad \square$$

Now we are ready to show complexity results for SG and RT-MLMC schemes.

SG The bias and second-order moment of the gradient estimator will not depend on the batch size n_L°, whereas the sample complexity and computation cost will be proportional to it. Therefore, we set the batch size $n_L^{\circ} = 1$. According to Proposition EC.2, Lemma EC.6(I), and the first part of Lemma EC.6(III), parameters for SG scheme satisfy

$$\Delta_F := \lambda\epsilon \exp\left(2B/(\lambda\epsilon)\right) \cdot 2^{-(L+1)}, \quad M_*^2 := \mathfrak{c}^{-2} L_f^2.$$

To obtain δ-optimal solution, we set

$$2\Delta_F \le \frac{\delta}{2}, \quad M_* \sqrt{\frac{2V(\theta_1, \overline{\theta}^*)}{\kappa T_{\text{in}}}} \le \frac{\delta}{2}.$$

As a consequence, we specify the following hyper-parameters to meet the above requirements:

$$L = \left\lceil \frac{1}{\log 2} \left[\log \frac{2\lambda\epsilon \exp(2B/(\lambda\epsilon))}{\delta} \right] \right\rceil, \quad T_{\text{in}} = \left\lceil \frac{8L_f^2 V(\theta_1, \overline{\theta}^*)}{\kappa \mathfrak{c}^2 \delta^2} \right\rceil, \quad \gamma = \sqrt{\frac{2\kappa \mathfrak{c}^2 V(\theta_1, \overline{\theta}^*)}{T_{\text{in}} L_f^2}}.$$

RT-MLMC Similar as in the proof of SG estimator, we take the batch size $n_L^{\circ}=1$. Define $q_\ell=\frac{2^{-\ell}}{2-2^{-L}}, \ell=0,1,\ldots,L$. By the second part of Lemma EC.6(III) and basic calculation, we find

$$\mathbb{E}\left[\left\|v^{\mathrm{RT-MLMC}}(\theta)\right\|_2^2\right]=\sum_{\ell=0}^{L}\frac{1}{q_\ell}\mathbb{E}\left[\left\|G^\ell(\theta,\zeta_1^\ell)\right\|_2^2\right]\le 2(L+1)L_f^2\exp(4B/(\lambda\epsilon)).$$

According to Proposition EC.2 and Lemma EC.6(I), parameters for RT-MLMC scheme satisfy

$$\Delta_F:=\lambda\epsilon\exp\left(2B/(\lambda\epsilon)\right)\cdot 2^{-(L+1)},\quad M_*^2:=2\mathfrak{c}^{-2}(L+1)L_f^2\exp((4B)/(\lambda\epsilon)),$$

To obtain δ-optimal solution, we set

$$2\Delta_F\le\frac{\delta}{2},\quad M_*\sqrt{\frac{2V(\theta_1,\overline{\theta}^*)}{\kappa T_{\mathrm{in}}}}\le\frac{\delta}{2}.$$

As a consequence, we specify the following hyper-parameters to meet the above requirements:

$$L=\left\lceil\frac{1}{\log 2}\left[\log\frac{2\lambda\epsilon\exp(2B/(\lambda\epsilon))}{\delta}\right]\right\rceil,\quad T_{\mathrm{in}}=\left\lceil\frac{16(L+1)L_f^2V(\theta_1,\overline{\theta}^*)\exp((4B)/(\lambda\epsilon))}{\kappa\mathfrak{c}^2\delta^2}\right\rceil,$$
$$\gamma=\sqrt{\frac{2\kappa V(\theta_1,\overline{\theta}^*)}{T_{\mathrm{in}}M_*^2}}.$$

After specifying the desired hyper-parameters such as $L, n_L^{\circ}, \gamma, T_{\mathrm{in}}$, one can apply Remark EC.1 to calculate the sample complexity and storage cost. Further details regarding the complexity results can be found in Table 2.

EC.7.2. Sampling Algorithm in Remark 7

In this subsection, we present an algorithm that generates samples from $\mathbb{Q}_\epsilon$, where the density function

$$\frac{\mathrm{d}\mathbb{Q}_\epsilon(z)}{\mathrm{d}z}\propto\exp\big(-V_\epsilon(z)\big),\quad V_\epsilon(z):=\|z\|_p^2$$

One can use the unadjusted Langevin algorithm for sampling:

$$\mathrm{d}X_t=-\nabla V_\epsilon(X_t)\,\mathrm{d}t+\sqrt{2}\,\mathrm{d}B_t,$$

where $\{B_t\}$ is a multi-dimensional Brownian motion. As the time index $t\to\infty$, the distribution X_t will converges to a stationary distribution $\mathbb{Q}_\epsilon$ exponentially fast. Also, for practical implementation we use the discreterized version of SDE for sampling:

$$X_{k+1}=X_k-\gamma\nabla V_\epsilon(X_k)+\sqrt{2\gamma}Z_{k+1},\quad \text{where } Z_{k+1}\sim\mathcal{N}(0,I_d).\tag{EC.7}$$

In particular, the function $V_\epsilon(z)$ is continuously differentiable with

$$\nabla V_\epsilon(z)=2\|v\|_p^{2-p}\mathrm{sign}(v)|v|^{p-1}.$$

Hence, the iteration (EC.7) returns a distribution that is τ-close to $\mathbb{Q}_\epsilon$ in terms of KL-divergence distance within $\mathcal{O}(1/\tau)$ iterations.

EC.7.3. Proof of Remark 8

If employing the BSAA technique, the estimation of optimal solution of (10) is given by the optimal value of the following problem, where the objective function is a biased estimate of the objective in (10):

$$\min_{\theta\in\Theta}\left\{\hat{F}_{n,m}(\theta):=\frac{\lambda\epsilon}{n}\sum_{i=1}^{n}\log\left(\frac{1}{m}\sum_{j=1}^{m}e^{f_\theta(z_{i,j})/(\lambda\epsilon)}\right)\right\}. \tag{EC.8}$$

Here $\{x_i\}_{i=1}^n$ are samples i.i.d. generated from $\widehat{\mathbb{P}}$, and for fixed x_i, samples $\{z_{i,j}\}_{j=1}^m$ are i.i.d. generated from $\mathbb{Q}_{x_i,\epsilon}$. Leveraging existing results in [59, Corollary 4.2], we present the following sample complexity analysis of BSAA problem.

PROPOSITION EC.3 **(Sample Complexity for BSAA Problem)**. *Assume the following conditions hold:*

(I) *The constraint set Θ is bounded with diameter $D_\Theta<\infty$*

(II) *For fixed z and θ_1,θ_2, it holds that $|f_{\theta_1}(z)-f_{\theta_2}(z)|\le L_f\|\theta_1-\theta_2\|_2$.*

(III) *The loss function f satisfies $0\le f_\theta(z)\le B$ for any $\theta\in\Theta$ and $z\in\mathcal{Z}$.*

Suppose we specify parameters in (EC.8) as

$$m=\left\lceil\frac{2\lambda\epsilon e^{2B/(\lambda\epsilon)}}{\delta}\right\rceil, n=\mathcal{O}(1)\frac{B^2+4B\lambda\epsilon e^{2B/(\lambda\epsilon)}}{\delta^2}\left[d\log\left(\frac{8e^{B/(\lambda\epsilon)}L_fD_\Theta}{\epsilon}\right)+\log\left(\frac{1}{\alpha}\right)\right],$$

then with probability at least $1-\alpha$, the solution to the SAA problem (EC.8) is an δ-optimal solution of (10).

The sample complexity of BSAA problem is $mn+n=\widetilde{\mathcal{O}}(\delta^{-3})$, which is much worse than $\widetilde{\mathcal{O}}(\delta^{-2})$, i.e., the complexity of first-order method used in our paper. Hence, we conclude that it takes considerably less time to implement the BSMD step directly rather than solving the SAA problem.

Proof of Proposition EC.3. We first verify the technical assumption imposed in [59, Corollary 4.2]. Specifically, one can show that

(a) The mapping $\phi:\ [1,e^{B/(\lambda\epsilon)}]\to\mathbb{R}$ such that $\phi(x)=\lambda\epsilon\log(x)$ is $\lambda\epsilon$-Lipscthiz continuous and $\lambda\epsilon$-smooth, and the mapping $g_z(\cdot,x):\ \Theta\to\mathbb{R}$ such that $g_z(\theta,x)=e^{f_\theta(z)/(\lambda\epsilon)}$ is $e^{B/(\lambda\epsilon)}L_f/(\lambda\epsilon)$-Lipschitz continuous.

(b) The variance

$$\begin{aligned}&\max_{\theta\in\Theta}\ \mathbb{V}\mathrm{ar}_{x\sim\widehat{\mathbb{P}}}\left(\lambda\epsilon\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_\theta(z)/(\lambda\epsilon)}\right]\right)\\ \le&\max_{\theta\in\Theta}\ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left(\lambda\epsilon\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_\theta(z)/(\lambda\epsilon)}\right]\right)^2\\ \le&B^2.\end{aligned}$$

(c) The variance

$$\begin{aligned}&\max_{\theta\in\Theta,x\in\mathrm{supp}\,\widehat{\mathbb{P}}}\ \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left(e^{f_\theta(z)/(\lambda\epsilon)}-\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}e^{f_\theta(z)/(\lambda\epsilon)}\right)^2\\ \le&\max_{\theta\in\Theta,x\in\mathrm{supp}\,\widehat{\mathbb{P}}}\ \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}e^{2f_\theta(z)/(\lambda\epsilon)}\\ \le&e^{2B/(\lambda\epsilon)}.\end{aligned}$$

(d) The mapping ϕ satisfies $|\phi(\cdot)|\le B$, and the mapping $g_z(\cdot,x)$ satisfies $|g_z(\cdot,x)|\le e^{B/(\lambda\epsilon)}$. Therefore, from [59, Corollary 4.2], we know that to obtain δ-optimal solution with probability at least $1-\alpha$, sample sizes m,n need to satisfy

$$n\ge\mathcal{O}(1)\frac{B^2+4B\lambda\epsilon e^{2B/(\lambda\epsilon)}}{\delta^2}\left[d\log\left(\frac{8e^{B/(\lambda\epsilon)}L_fD_\Theta}{\epsilon}\right)+\log\left(\frac{1}{\alpha}\right)\right]$$

and

$$m \geq \frac{2\lambda\epsilon e^{2B/(\lambda\epsilon)}}{\delta}.$$

□

EC.8. Proofs of Technical Results in Section 4.2.2

We first provide two technical lemmas that can be useful to show the main results in Section 4.2.2.

LEMMA EC.7. *Under Assumption 2(II), it holds that*

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(a^\ell(\theta,\zeta^\ell)\right) &\le B^2\\ \mathbb{V}\mathrm{ar}\left(A^\ell(\theta,\zeta^\ell)\right) &\le \lambda^2\epsilon^2 e^{2B/(\lambda\epsilon)}\cdot 2^{-\ell}.\end{aligned}$$

Proof of Lemma EC.7. First, we find

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(a^\ell(\theta,\zeta^\ell)\right) &\le \mathbb{E}\left[a^\ell(\theta,\zeta^\ell)\right]^2\\ &= \mathbb{E}\left[\lambda\epsilon\log\left(\frac{1}{2^\ell}\sum_{j\in[2^\ell]}\exp\left(\frac{f_\theta(z_j^\ell)}{\lambda\epsilon}\right)\right)\right]^2\\ &\le B^2,\end{aligned}$$

where the last inequality is because $0\le f_\theta(z_j^\ell)\le B$. Next, we find

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(A^\ell(\theta,\zeta^\ell)\right) &\le \mathbb{E}\left[A^\ell(\theta,\zeta^\ell)\right]^2\\ &= \mathbb{E}\left|\frac{1}{2}\left(U_{1:2^\ell}(\theta,\zeta^\ell)-U_{1:2^{\ell-1}}(\theta,\zeta^\ell)\right)+\frac{1}{2}\left(U_{1:2^\ell}(\theta,\zeta^\ell)-U_{2^{\ell-1}+1:2^\ell}(\theta,\zeta^\ell)\right)\right|^2\\ &\le \frac{1}{2}\mathbb{E}\left|U_{1:2^\ell}(\theta,\zeta^\ell)-U_{1:2^{\ell-1}}(\theta,\zeta^\ell)\right|^2+\frac{1}{2}\mathbb{E}\left|U_{1:2^\ell}(\theta,\zeta^\ell)-U_{2^{\ell-1}+1:2^\ell}(\theta,\zeta^\ell)\right|^2\\ &\le \frac{\lambda^2\epsilon^2}{2}\mathbb{E}\left|\frac{1}{2^\ell}\sum_{j\in[2^\ell]}\exp\left(\frac{f_\theta(z_j^\ell)}{\lambda\epsilon}\right)-\frac{1}{2^{\ell-1}}\sum_{j\in[2^{\ell-1}]}\exp\left(\frac{f_\theta(z_j^\ell)}{\lambda\epsilon}\right)\right|^2\\ &\qquad+\frac{\lambda^2\epsilon^2}{2}\mathbb{E}\left|\frac{1}{2^\ell}\sum_{j\in[2^\ell]}\exp\left(\frac{f_\theta(z_j^\ell)}{\lambda\epsilon}\right)-\frac{1}{2^{\ell-1}}\sum_{j\in[2^{\ell-1}+1:2^\ell]}\exp\left(\frac{f_\theta(z_j^\ell)}{\lambda\epsilon}\right)\right|^2\\ &= \frac{\lambda^2\epsilon^2}{4}\mathbb{E}\left|\frac{1}{2^{\ell-1}}\sum_{j\in[2^{\ell-1}]}\exp\left(\frac{f_\theta(z_j^\ell)}{\lambda\epsilon}\right)-\frac{1}{2^{\ell-1}}\sum_{j\in[2^{\ell-1}+1:2^\ell]}\exp\left(\frac{f_\theta(z_j^\ell)}{\lambda\epsilon}\right)\right|^2\\ &\le \frac{\lambda^2\epsilon^2}{4}\cdot\frac{2\exp\left(2B/(\lambda\epsilon)\right)}{2^{\ell-1}}\\ &= \lambda^2\epsilon^2 e^{2B/(\lambda\epsilon)}\cdot 2^{-\ell}.\end{aligned}$$

□

LEMMA EC.8 **(Cramer's Large Deviation Theorem)**. *Let $X_1,\ldots,X_n$ be i.i.d. samples of zero-mean random variable X with finite variance σ^2. For any $\bar\epsilon>0$, there exists $\epsilon_1>0$ such that for any $\epsilon\in(0,\epsilon_1)$, it holds that*

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right|\ge\epsilon\right)\le 2\exp\left(-\frac{n\epsilon^2}{(2+\bar\epsilon)\sigma^2}\right).$$

Next, we present the complexity of estimating objective value of a feasible solution θ using MLMC-based estimators.

PROPOSITION EC.4 **(Complexity of MLMC-based Objective Estimators)**. *Assume that Assumption 2(II) holds, then with properly chosen hyper-parameters of objective estimators in (13b), the following results hold:*

(I) *The total cost of SG scheme for estimating objective value for fixed θ up to accuracy error δ, with probability at least $1-\alpha$, is $\mathcal{O}\left(\log\frac{1}{\alpha}\cdot\delta^{-3}\right)$;*

(II) *The total cost of RT-MLMC scheme for estimating objective value for fixed θ up to accuracy error δ, with probability at least $1-\alpha$, is $\widetilde{\mathcal{O}}\left(\log\frac{1}{\alpha}\cdot\delta^{-2}\right)$.*

The configuration of optimization hyper-parameters is provided in the following:

$$\begin{aligned} \textit{SG}: \quad & L=\mathcal{O}\left(\log\frac{1}{\delta}\right), n_L^{\circ}=\mathcal{O}\left(\frac{1}{\delta^2}\cdot\log\frac{1}{\alpha}\right); \\ \textit{RT-MLMC}: \quad & L=\mathcal{O}\left(\log\frac{1}{\delta}\right), n_L^{\circ}=\widetilde{\mathcal{O}}\left(\frac{1}{\delta^2}\cdot\log\frac{1}{\alpha}\right). \end{aligned}$$

Proof of Proposition EC.4. First, we pick L such that $|F^L(\theta)-F(\theta)|\le\frac{\delta}{4}$, i,e.,

$$L=\left\lceil\frac{1}{\log 2}\left[\log\frac{2\lambda\epsilon\exp(2B/(\lambda\epsilon))}{\delta}\right]\right\rceil.$$

Assume we have the estimator $V(\theta)$ such that $\mathbb{E}[V(\theta)]=F^L(\theta)$ and $\mathbb{V}\mathrm{ar}(V(\theta))<\infty$.

$$\left\{|F(\theta)-V(\theta)|>\frac{\delta}{2}\right\}\subseteq\left\{|F^L(\theta)-V(\theta)|>\frac{\delta}{4}\right\}. \tag{EC.9}$$

Then by the relation (EC.9) and Lemma EC.8, for any $\delta'>0$, there exists $\delta_1>0$ such that for any $\delta\in(0,\delta_1)$,

$$\Pr\left\{|F(\theta)-V(\theta)|>\frac{\delta}{2}\right\}\le\Pr\left\{|F^L(\theta)-V(\theta)|>\frac{\delta}{4}\right\}\le 2\exp\left(-\frac{\delta^2}{16(\delta'+2)\mathbb{V}\mathrm{ar}(V(\theta))}\right). \tag{EC.10}$$

Specially, we find $V^{\text{SG}}(\theta), V^{\text{RT-MLMC}}(\theta)$ are all unbiased estimators of $F^L(\theta)$ with

$$\begin{aligned} \mathbb{V}\mathrm{ar}\left(V^{\text{SG}}(\theta)\right)&\le\frac{1}{n_L^{\circ}}\mathbb{V}\mathrm{ar}\left(a^L(\theta,\zeta_i^L)\right)\le\frac{B^2}{n_L^{\circ}}, \\ \mathbb{V}\mathrm{ar}\left(V^{\text{RT-MLMC}}(\theta)\right)&\le\frac{1}{n_L^{\circ}}\sum_{\ell=0}^{L}\frac{1}{q_\ell}\mathbb{V}\mathrm{ar}\left(A^\ell(\theta,\zeta_i^L)\right)\le\lambda^2\epsilon^2e^{2B/(\lambda\epsilon)}\cdot(L+1)\cdot(n_L^{\circ})^{-1}. \end{aligned}$$

The concentration for SG scheme becomes

$$\Pr\left\{|F(\theta)-V^{\text{SG}}(\theta)|>\frac{\delta}{2}\right\}\le 2\exp\left(-\frac{\delta^2 n_L^{\circ}}{16(\delta'+2)B^2}\right).$$

To make the desired coverage probability, we take

$$n_L^{\circ}=\frac{16(\delta'+2)B^2}{\delta^2}\cdot\log\frac{2}{\alpha}.$$

The concentration for RT-MLMC scheme becomes

$$\Pr\left\{|F(\theta)-V^{\text{RT-MLMC}}(\theta)|>\frac{\delta}{2}\right\}\le\exp\left(-\frac{\delta^2 n_L^{\circ}}{16(\delta'+2)\lambda^2\epsilon^2e^{2B/(\lambda\epsilon)}(L+1)}\right).$$

To make the desired coverage probability, we take

$$n_L^{\circ}=\frac{16(\delta'+2)\lambda^2\epsilon^2e^{2B/(\lambda\epsilon)}(L+1)}{\delta^2}\cdot\log\frac{2}{\alpha}.$$

□

In the following, we provide the proof of Proposition 1.

Proof of Proposition 1. The goal is to choose hyper-parameters such that

$$\Pr\left\{\left|\min_{i\in[m]} V(\widehat{\theta}_i) - F(\theta^*)\right| \le \delta\right\} \ge 1-\eta.$$

On the one hand,

$$\min_{i\in[m]} V(\widehat{\theta}_i) - F(\theta^*) \le \min_{i\in[m]} F(\widehat{\theta}_i) - F(\theta^*) + \max_{i\in[m]} |V(\widehat{\theta}_i) - F(\widehat{\theta}_i)|.$$

On the other hand,

$$F(\theta^*) - \min_{i\in[m]} V(\widehat{\theta}_i) \le F(\theta^*) - \min_{i\in[m]} F(\widehat{\theta}_i) + \max_{i\in[m]} |V(\widehat{\theta}_i) - F(\widehat{\theta}_i)| \le \max_{i\in[m]} |V(\widehat{\theta}_i) - F(\widehat{\theta}_i)|.$$

Based on those two inequalities, it suffices to choose hyper-parameters such that

$$\Pr\left\{\max_{i\in[m]} \left|V(\widehat{\theta}_i) - F(\widehat{\theta}_i)\right| \le \frac{\delta}{2}\right\} \ge 1-\frac{\eta}{2} \tag{EC.11}$$

and

$$\Pr\left\{\min_{i\in[m]} F(\widehat{\theta}_i) - F(\theta^*) \le \frac{\delta}{2}\right\} \ge 1-\frac{\eta}{2}. \tag{EC.12}$$

To ensure the relation (EC.11), it suffices to apply Proposition EC.4 such that

$$\Pr\left\{\left|V(\widehat{\theta}_i) - F(\widehat{\theta}_i)\right| \le \frac{\delta}{2}\right\} \ge 1-\frac{\eta}{2m}, \quad \forall i\in[m],$$

the total cost of which is

$$m\widetilde{\mathcal{O}}\left(\log\frac{2m}{\eta}(\delta/2)^{-2}\right) = \widetilde{\mathcal{O}}\left(m\log\frac{m}{\eta}\delta^{-2}\right).$$

To ensure the relation (EC.12), it suffices to take

$$\Pr\left\{F(\widehat{\theta}_i) - F(\theta^*) \le \frac{\delta}{2}\right\} \ge 1-\left(\frac{\eta}{2}\right)^{1/m}, \quad \forall i\in[m].$$

By Markov's inequality, it suffices to ensure

$$\mathbb{E}[F(\widehat{\theta}_i) - F(\theta^*)] \le \frac{\delta}{2}\left(\frac{\eta}{2}\right)^{1/m}, \quad \forall i\in[m]. \tag{EC.13}$$

(I) When running BSMD algorithm with SG scheme, its sample complexity to ensure the relation (EC.13) is

$$m\mathcal{O}\left(\left(\frac{\delta}{2}\left(\frac{\eta}{2}\right)^{1/m}\right)^{-3}\right) = \mathcal{O}\left(m\delta^{-3}\eta^{-3/m}\right).$$

Now we specify the batch size m such that the total cost is minimized:

$$\min_m \left\{\mathcal{O}\left(m\delta^{-3}\eta^{-3/m}\right) + \widetilde{\mathcal{O}}\left(m\log\frac{m}{\eta}\delta^{-2}\right)\right\} = \min_m \mathcal{O}\left(m\delta^{-3}\eta^{-3/m}\right).$$

The objective above is lower bounded by $\mathcal{O}(\delta^{-3})$. When taking batch size $m = \lceil\log_2\frac{1}{\eta}\rceil$, the objective value becomes $\mathcal{O}(\delta^{-3}\log\frac{1}{\eta})$, justifying that $m = \log_2\frac{1}{\eta}$ is a near-optimal choice.

(II) When running the BSMD algorithm with RT-MLMC scheme, the sample complexity becomes

$$m\widetilde{\mathcal{O}}\left(\left(\frac{\delta}{2}\left(\frac{\eta}{2}\right)^{1/m}\right)^{-2}\right)=\widetilde{\mathcal{O}}\left(m\delta^{-2}\eta^{-2/m}\right).$$

Now we specify the batch size m such that the total cost is minimized:

$$\min_{m}\ \left\{\widetilde{\mathcal{O}}\left(m\delta^{-2}\eta^{-2/m}\right)+\widetilde{\mathcal{O}}\left(m\log\frac{m}{\eta}\delta^{-2}\right)\right\}$$

Similarly, the objective above is lower bounded by $\widetilde{\mathcal{O}}(\delta^{-2})$. When taking batch size $m=\lceil\log_2\frac{1}{\eta}\rceil$, the objective value becomes $\widetilde{\mathcal{O}}(\delta^{-2}\log\left(\frac{1}{\eta}\right)\log\left(\frac{1}{\eta}\log\frac{1}{\eta}\right))$, justifying that $m=\log_2\frac{1}{\eta}$ is a near-optimal choice.

□

Finally, we show the proof of Theorem 3. A key technique is the following complexity result on bisection search with inexact oracles.

LEMMA EC.9 **(Complexity for Noisy Bisection [31, Lemma 33])**. *Let $f:\mathbb{R}\to\mathbb{R}$ be a B-Lipschitz and convex function defined on the interval $[l,u]$, and $\mathcal{G}:\mathbb{R}\to\mathbb{R}$ be an oracle so that $|\mathcal{G}(y)-f(y)|\le\tilde{\delta}$ for all y. With at most*

$$1+2\left\lceil\log_{3/2}\frac{B(u-l)}{\tilde{\delta}}\right\rceil$$

calls to $\mathcal{G}$, the algorithm `OneDimMinimizer` [31, Algorithm 8] outputs y' so that

$$f(y')-\min_{y}\ f(y)\le 4\tilde{\delta}.$$

Proof of Theorem 3. Since $f_\theta(z)$ is convex in θ, one can check that the objective $F(\theta;\lambda)$ is jointly convex in (θ,λ), and therefore the objective $F^*(\lambda)$ is convex in λ. Also, by danskin's theorem, we find

$$\frac{\partial}{\partial\lambda}F^*(\lambda)=\overline{\rho}+\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\epsilon\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_{\theta^*}(z)/(\lambda\epsilon)}\right]\right]-\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\frac{\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_{\theta^*}(z)/(\lambda\epsilon)}f_{\theta^*}(z)\right]}{\lambda\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_{\theta^*}(z)/(\lambda\epsilon)}\right]}\right].$$

Since $0\le f_\theta(z)\le B$, we find

$$0\le\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\epsilon\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_{\theta^*}(z)/(\lambda\epsilon)}\right]\right]\le\frac{B}{\lambda}\le\frac{B}{\lambda_l}$$

and

$$0\le\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\frac{\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_{\theta^*}(z)/(\lambda\epsilon)}f_{\theta^*}(z)\right]}{\lambda\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_{\theta^*}(z)/(\lambda\epsilon)}\right]}\right]\le\frac{e^{B/(\lambda\epsilon)}B}{\lambda}\le\frac{e^{B/(\lambda_l\epsilon)}B}{\lambda_l}$$

Therefore, the subgradient of $F^*(\lambda)$ is bounded:

$$\left|\frac{\partial}{\partial\lambda}F^*(\lambda)\right|\le L_\lambda\triangleq\overline{\rho}+\frac{B}{\lambda_l}\left[1+e^{B/(\lambda_l\epsilon)}\right].$$

In summary, $F^*(\lambda)$ is a L_λ-Lipschitz and convex function defined on $[\lambda_l,\lambda_u]$. Applying Lemma EC.9 with $\tilde{\delta}:=\delta/4$ together with the union bound, we are able to find the optimal multiplier up to accuracy δ with probability at least $1-\eta$ by calling the oracle $\widehat{F}$ for

$$1+2\left\lceil\log_{3/2}\frac{4L_\lambda(\lambda_u-\lambda_l)}{\delta}\right\rceil$$

times. □