TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
THESIS PROPOSAL SUMMARY

Problems of decision making under uncertainty occurs frequently in real-life applications, such as industrial engineering, computer science, and management. Traditional approaches proposed to formulate, analyze, and solve these problems through the lens of optimization, in which they assume the input optimization parameters can be obtained or estimated accurately from data. However, this is not usually the case in the era of Big Data. Big data formulations usually contain a large amount of uncertain parameters such that the estimated ones may not be representative of the ground truth. Strategies solely based on naively estimated parameters may have poor out-of-sample performance. Even worse, they can be risky or unethical and lead to severe consequences. Therefore, it is important to establish reliable decision-making modeling to handle data uncertainty due to measurement error, insufficient sample size, contamination, anomalies, or model misspecification.

In this thesis proposal, we aim to propose solutions to partially address this challenge in different decision-making problems with offline, noisy, small-sample, or high-dimensional data. On the one hand, we focus on developing computationally efficient methodologies through the lens of modern optimization techniques. On the other hand, we provide strong performance guarantees of the proposed modeling leveraging tools from statistics.

In chapter 2, we consider the problem of two-sample testing: given two sets of samples, aiming to determine whether they are from the same distribution. We propose a kernel projected Wasserstein distance (KPW) to develop a new two-sample test, which operates by finding the nonlinear mapping in the data space which maximizes the distance between projected distributions. Specially,

(I) We develop a computationally efficient algorithm for evaluating the KPW using a representer theorem to reformulate the problem into a finite-dimensional optimization

problem and a block coordinate descent optimization algorithm which is guaranteed to find an ϵ-stationary point with complexity $\mathcal{O}(\epsilon^{-3})$.

(II) To quantify the false detection rate, which is essential in setting the detection threshold, we develop non-asymptotic bounds for empirical KPW distance, and therefore demonstrate our proposed two-sample test efficiently circumvents the curse of dimensionality.

In chapter 3, we consider the distributionally robust stochastic optimization (DRO) problem: aiming to find a robust optimal decision that minimizes the expected loss under the most adverse distribution within a given set of relevant distributions, called the ambiguity set. We propose a new framework for this problem by constructing the ambiguity set using Sinkhorn distance – a variant of Wasserstein distance based on entropic regularization. Specifically,

(I) We derive a strong duality reformulation for Sinkhorn DRO when the nominal distribution is any arbitrary distribution. The Sinkhorn dual objective smooths the maximization subproblem in the Wasserstein dual objective, and converges to Wasserstein dual objective as the entropic regularization parameter goes to zero.

(II) As a byproduct of our duality proof, we characterize the worst-case distribution of the Sinkhorn DRO, which is absolutely continuous with respect to some reference measure such as Lebesgue or counting measure. Compared with Wasserstein DRO, the worst-case distribution of Sinkhorn DRO is not necessarily finitely supported even when the nominal distribution is a finitely supported distribution. This indicates that Sinkhorn DRO is a more flexible modeling choice for many applications.

(III) On the algorithmic aspect, we propose and analyze an efficient stochastic mirror descent method using biased gradient oracles with bisection search for solving the Sinkhorn DRO problem. By adequately balancing the trade-off between bias and variance of stochastic gradient estimators with low computation cost, we show the

proposed algorithm achieves computation cost $\tilde{O}(\delta^{-3})$ and memory cost $\tilde{O}(\delta^{-2})$ for finding δ-optimal solution for convex loss, and the computation cost improves to $\tilde{O}(\delta^{-2})$ for convex and smooth loss. Compared with Wasserstein DRO, the dual problem of Sinkhorn DRO is computationally tractable for a broader class of loss functions, cost functions, nominal distributions, and probability support.

In chapter 4 (which is our current project), we consider the problem of two-sample feature selection: given two sets of samples in high-dimensional space, aiming to select a subset of representative features that distinguish those two groups well. We provide a new framework that seeks the most representative features such that the Maximum Mean Discrepancy (MMD) between samples, a novel statistical divergence in literature, is maximized. From the algorithmic perspective, we reformulate this problem as mixed-integer non-convex programming, and develop exact and approximation algorithms to solve the formulation. From the theoretical perspective, we characterize statistical properties on our proposed framework.

Finally, we list some ongoing topics that will be covered in this thesis.

- In chapter 5, we again consider the problem of two-sample testing by taking the special structure of data into consideration. When the high-dimensional data points are supported on a low-dimensional manifold, we will propose efficient two-sample tests using neural networks. Leverging tools from deep learning theory, we are able to provide strong statistical performance guarantees for our proposed tests.

- In chapter 6, we will explore statistical analysis and practical applications of our proposed Sinkhorn DRO framework. Experiment results in chapter 3 have partially demonstrated the superior out-of-sample performance of the Sinkhorn DRO model, but it is important to understand when will the Sinkhorn DRO model perform better than other baseline models based on rigorous theoretical analysis. Besides, traditional distributionally robust decision-making modeling was already attractive in various

real-world applications. It is of research interest to study whether the Sinkhorn DRO will bring more benefits in those tasks.

CHAPTER 2

TWO-SAMPLE TEST WITH KERNEL PROJECTED WASSERSTEIN DISTANCE

We develop a kernel projected Wasserstein distance for the two-sample test, an essential building block in statistics and machine learning: given two sets of samples, to determine whether they are from the same distribution. This method operates by finding the nonlinear mapping in the data space which maximizes the distance between projected distributions. In contrast to existing works about projected Wasserstein distance, the proposed method circumvents the curse of dimensionality more efficiently. We present practical algorithms for computing this distance function together with the non-asymptotic uncertainty quantification of empirical estimates. Numerical examples validate our theoretical results and demonstrate good performance of the proposed method.

2.1 Background

As a fundamental problem in statistical inference [170], two-sample hypothesis testing aims to determine whether two sets of samples come from the same distribution or not. This problem has broad applications in scientific discovery fields. For example, it can be applied in anomaly detection [3, 30, 134] to identify abnormal observations that follow a distinct distribution compared with typical observations. Similarly, in change-point detection [124, 165, 166], two-sample testing is essential to detect abrupt changes in streaming data. Other notable examples include model criticism [16, 39, 103], causal inference [104], and health care [136].

Parametric or low-dimensional testing scenarios have been the main focus in classical literature. When extra knowledge about the data distributions is available, one can design parametric tests, such as Hotelling's two-sample test [78], Student's t-test [121], etc. Non-parametric two-sample tests are more attractive when the exact parametric form of the data

distributions is hard to specify. It is popular to design non-parametric tests using integral probability metrics, since the evaluation of the corresponding test statistics can be obtained based on samples without knowing the densities of data distributions. Some earlier works design tests using Kolmogorov-Smirnov distance [90, 126], total variation distance [75], and Wasserstein distance [47, 130]. However, it is not proper to use these tests for high-dimensional settings since the sample complexity for estimating those distance functions based on empirical samples suffers from the curse of dimensionality.

There is a strong need for developing non-parametric tests for high-dimensional data, especially for modern applications. A notable contribution is the two-sample test based on Maximum Mean Discrepancy (MMD) [36, 73, 74]. Although the power of MMD test with the median choice of kernel bandwidth decays quickly when the dimension of distributions increases [131], this test with properly chosen bandwidth does not have the curse of dimensionality issue for low-dimensional manifold data as pointed out in [36]. Unfortunately, the MMD test with optimized bandwidth still does not demonstrate good testing power for the small-sampled case as demonstrated numerically in this paper. In addition, recent works [157, 165] leverage the idea of dimensionality reduction for dealing with high-dimensional settings, which use the projected Wasserstein distance as the test statistic, i.e., the test statistic works by finding the linear projector such that the distance between projected distributions is maximized. However, a linear projector may not serve as an optimal design for maximizing the power of tests as demonstrated numerically in Section 2.5.

In this paper, we present a new non-parametric two-sample test statistic aiming for the high-dimensional setting based on a *kernel projected Wasserstein (KPW) distance*, with a nonlinear projector based on the reproducing kernel Hilbert space (RKHS) designed to optimize the test power via maximizing the probability distance between the distributions after projection. In addition, our contributions include the following:

- We develop a computationally efficient algorithm for evaluating the KPW using a

representer theorem to reformulate the problem into a finite-dimensional optimization problem and a block coordinate descent optimization algorithm which is guaranteed to find an ϵ-stationary point with complexity $\mathcal{O}\left(\epsilon^{-3}\right)$.

- To quantify the false detection rate, which is essential in setting the detection threshold, we develop non-asymptotic bounds for empirical KPW distance based on the covering number argument.

- We present numerical experiments to validate our theoretical results as well as demonstrate the competitive performance of our proposed test using both synthetic and real data.

Related Work. It is helpful to understand the structure of high-dimension distributions by low-dimensional projections. Notable methodologies include the principal component analysis (PCA) [89], kernel PCA [138], factor analysis [45], etc. Several works leverage this idea to design tests for high-dimensional data. [111] and [165] first design tests by finding the worst-case linear projector that maximizes the distance between projected sample points in one dimension. Later [99] and [157] naturally extend this idea by developing a projector that maps sample points into d dimensional linear subspace with $d \geq 1$, called projected Wasserstein distance. Efficient optimization algorithms and statistical properties of this distance have been investigated in recent works [85, 100]. However, a linear projector cannot efficiently capture features from data with nonlinear patterns, limiting the performance of tests mentioned above for practical applications. It is therefore promising to use nonlinear dimensionality reduction for two-sample testing. Although nonlinear projectors can be obtained using neural networks [67], the sample complexity of the corresponding test statistic will have slow convergence rates since the neural network function class usually has high complexity in terms of the covering number. Recently kernel method has been demonstrated to be beneficial for understanding data [28, 79, 91, 109] because of sharp sample complexity rate, low computational cost, and flexible representation of features. This

fact motivates us to use a nonlinear projector based on kernels to design tests. Compared with the linear projector, computing the corresponding statistic and analyzing its performance is more challenging since the function space cannot be parameterized by finite-dimensional coefficients. We leverage the kernel trick to finish these two parts.

The remaining of this paper is organized as follows. Section 2.2 introduces some preliminary knowledge on two-sample testing and related probability distances, Section 2.3 outlines a practical algorithm for computing KPW distance, Section 2.4 studies the uncertainty quantification of empirical KPW distance, Section 2.5 demonstrates some numerical experiments, and Section 2.6 presents some concluding remarks.

2.2 Problem Setup

Let $x^n := \{x_i\}_{i=1}^n$ and $y^m := \{y_i\}_{i=1}^m$ be i.i.d. samples generated from distributions μ and ν supported on $\mathbb{R}^D$, respectively. Our goal is to design a two-sample test which, given samples x^n and y^m, decides to accept the null hypothesis $H_0 :\ \mu = \nu$ or reject H_0 in favor of the alternative hypothesis $H_1 :\ \mu \neq \nu$. Denote by $T :\ (x^n, y^m) \to \{t_0, t_1\}$ the two-sample test, where t_0 means we reject H_1 and t_1 means we accept H_1 and reject H_0. Define the type-I risk as the probability of rejecting hypothesis H_0 when it is true, and the type-II risk as the probability of accepting H_0 when $\mu \neq \nu$:

$$\epsilon_{n,m}^{(\mathrm{I})} = \mathbb{P}_{x^n \sim \mu, y^m \sim \nu}\bigg(T(x^n, y^m) = t_1\bigg), \quad \text{under } H_0,$$
$$\epsilon_{n,m}^{(\mathrm{II})} = \mathbb{P}_{x^n \sim \mu, y^m \sim \nu}\bigg(T(x^n, y^m) = t_0\bigg), \quad \text{under } H_1.$$

Given parameters $\alpha, \beta \in (0, \frac{1}{2})$, we aim at building a two-sample test such that, when applied to n-observation samples x^n and m-observation samples y^m, it has the type-I risk at most α (i.e., at level α) and the type-II risk at most β (i.e., of power $1 - \beta$). Moreover, we want to ensure these specifications with sample sizes n, m as small as possible.

We propose a non-parametric test by considering the probability distance functions

between two empirical distributions constructed from observed samples. Specifically, we design a test T such that the null hypothesis H_0 is rejected when

$$\mathcal{D}(\hat{\mu}_n, \hat{\nu}_m) > \chi,$$

where $\mathcal{D}(\cdot, \cdot)$ is a divergence quantifying the differences of two distributions, χ is a data-dependent threshold, and $\hat{\mu}_n$ and $\hat{\nu}_m$ are empirical distributions from n samples in μ and m samples in ν, respectively. Several existing tests can be unified into this framework by taking $\mathcal{D}(\cdot, \cdot)$ as some special probability distances, including the MMD test, total variation distance test, etc. In this paper, we will design the divergence $\mathcal{D}$ based on the Wasserstein distance, and we specify the cost function $c(x, y) = \|x - y\|_2^2$.

Definition 1 (Wasserstein Distance)**.** *Given two distributions μ and ν, the Wasserstein distance is defined as*

$$W(\mu, \nu) = \min_{\pi \in \Pi(\mu, \nu)} \int c(x, y) \, \mathrm{d}\pi(x, y),$$

where $c(\cdot, \cdot)$ denotes the cost function quantifying the distance between two points, and $\Pi(\mu, \nu)$ denotes the joint distribution with marginal distributions μ and ν.

Although Wasserstein distance has wide applications in machine learning, the finite-sample convergence rate of Wasserstein distance between empirical distributions is slow in high-dimensional settings [57]. Therefore, it is not suitable for high-dimensional two-sample tests. Instead, existing works use the projection idea to rescue this issue.

Definition 2 (Projected Wasserstein Distance)**.** *Given two distributions μ and ν, define the projected Wasserstein distance as*

$$\mathcal{P}W(\mu, \nu) = \max_{\mathcal{A}\colon \mathbb{R}^D \to \mathbb{R}^d, A^{\mathrm{T}}A = I_d} W\left(\mathcal{A}\#\mu, \mathcal{A}\#\nu\right),$$

where the operator $\#$ denotes the push-forward operator, i.e.,

$$\mathcal{A}(z) \sim \mathcal{A}\#\mu \quad \textit{for } z \sim \mu,$$

and we denote $\mathcal{A}$ as a linear operator such that $\mathcal{A}(z) = A^{\mathrm{T}} z$ with $z \in \mathbb{R}^D$ and $A \in \mathbb{R}^{D\times d}$.

This idea is demonstrated to be useful for breaking the curse of dimensionality for the original Wasserstein distance [100, 157]. However, a linear projector is not an optimal choice for dimensionality reduction. Instead, we will consider a nonlinear projector to obtain a more powerful two-sample test, and we use functions in vector-valued reproducing kernel Hilbert space (RKHS) for projection.

Definition 3 (Vector-valued RKHS)**.** *A function $K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{d\times d}$ is said to be a positive semi-definite kernel if*

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\langle \bar{y}_i, K(\bar{x}_i, \bar{x}_j)\bar{y}_j\rangle \geq 0$$

for any finite set of points $\{\bar{x}_i\}_{i=1}^N$ in $\mathbb{R}^D$ and $\{\bar{y}_i\}_{i=1}^N$ in $\mathbb{R}^d$. Given such a kernel, there exists an unique $\mathbb{R}^d$-valued Hilbert space $\mathcal{H}_K$ with the reproducing kernel K. For fixed $x \in \mathbb{R}^D$ and $y \in \mathbb{R}^d$, define the kernel section K_x with the action y as the mapping $K_x y : \mathbb{R}^D \to \mathbb{R}^d$ such that

$$(K_x y)(x') = K(x', x)y, \quad \forall x' \in \mathbb{R}^D.$$

In particular, the Hilbert space $\mathcal{H}_K$ satisfies the reproducing property:

$$\forall f \in \mathcal{H}_K, \quad \langle f, K_x y\rangle_{\mathcal{H}_K} = \langle f(x), y\rangle.$$

Definition 4 (Kernel Projected Wasserstein Distance)**.** *Consider a $\mathbb{R}^d$-valued RKHS $\mathcal{H}$ with the corresponding kernel function K. Given two distributions μ and ν, define the kernel*

projected Wasserstein (KPW) distance as

$$\mathcal{KPW}(\mu,\nu) = \max_{f\in\mathcal{F}} W\left(f\#\mu, f\#\nu\right)$$

where the function class $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$.

Remark 1. *For* $d = 1$*, when the kernel function* $K(x,y) = \langle x, y\rangle$*, the KPW distance reduces into the PW distance. However, these two distances are not the same for general* d*. Moreover, existing works [10, 29, 108, 109] consider the design of the matrix-valued kernel function for* $d > 1$ *as*

$$K(x,x') = k(x,x')\cdot P, \tag{2.1}$$

where $k(\cdot,\cdot)$ *denotes a scalar-valued kernel function and* $P \in \mathbb{R}^{d\times d}$ *is a positive semi-definite matrix that encodes the relation between the output space. Such a design reduces the computational cost for applying vector-valued RKHS.*

In this paper, we design the two-sample test as follows. We split the data points into training and testing datasets. We first use the training set to train a nonlinear projector that maps data points into $\mathbb{R}^d$-subspace, and then perform the permutation test on testing data points that are projected based on the trained projector. The detailed algorithm is presented in Algorithm 1. This test is guaranteed to exactly control the type-I error [72] because we evaluate the p-value of the test via the permutation approach. To obtain reliable two-sample tests, we also require the KPW distance satisfies the discriminative property that $\mathcal{KPW}(\mu,\nu) = 0$ if and only if $\mu = \nu$. The following proposition reveals that this property holds by considering the vector-valued RKHS satisfying the universal property, the proof of which is provided in Appendix A.3. We also study how to compute the kernel projected distance and its related statistical properties in the following sections.

Proposition 1 (Discriminative Property of KPW)**.** *Denote by* $\mathcal{C}_b(\mathcal{X})$ *the space of bounded and continuous* $\mathbb{R}^d$*-valued functions on* $\mathcal{X}$*. Assume that* $\mathcal{H}$ *is a universal vector-valued*

Algorithm 1 Permutation two-sample test using the KPW distance

Require: Level α, number of permutation times N_p, collected samples x^n and y^m.

1: Split data as $x^n = x^{\mathrm{Tr}} \cup x^{\mathrm{Te}}$ and $y^m = y^{\mathrm{Tr}} \cup y^{\mathrm{Te}}$.
2: Formulate empirical distributions $(\hat{\mu}^{\mathrm{Tr}}, \hat{\nu}^{\mathrm{Tr}})$ corresponding to $(x^{\mathrm{Tr}}, y^{\mathrm{Tr}})$.
3: Obtain f as the (approximate) optimal projector to $\mathcal{KPW}(\hat{\mu}^{\mathrm{Tr}}, \hat{\nu}^{\mathrm{Tr}})$.
4: Compute the statistic $T = W(f\#\hat{\mu}^{\mathrm{Te}}, f\#\hat{\nu}^{\mathrm{Te}})$.
5: **for** $t = 1, \ldots, N_p$ **do**
6: Shuffle $x^{\mathrm{Te}} \cup y^{\mathrm{Te}}$ to obtain $x^{\mathrm{Te}}_{(t)}$ and $y^{\mathrm{Te}}_{(t)}$.
7: Formulate empirical distributions $(\hat{\mu}^{\mathrm{Te}}_{(t)}, \hat{\nu}^{\mathrm{Te}}_{(t)})$ corresponding to $(x^{\mathrm{Te}}, y^{\mathrm{Te}})$.
8: Compute the statistic for permuted samples $T_t = W(f\#\hat{\mu}^{\mathrm{Te}}_{(t)}, f\#\hat{\nu}^{\mathrm{Te}}_{(t)})$.
9: **end for**
Return the p-value $\frac{1}{N_p}\sum_{t=1}^{N_p} 1\{T_t \geq T\}$.

RKHS so that for any $\varepsilon > 0$ and $f \in \mathcal{C}_b(\mathcal{X})$, there exists $g \in \mathcal{H}$ so that

$$\|f - g\|_\infty \triangleq \sup_{x \in \mathcal{X}} \|f(x) - g(x)\|_2 < \varepsilon.$$

Then the KPW distance $\mathcal{KPW}(\mu, \nu) = 0$ if and only if $\mu = \nu$.

2.3 Computing KPW Distance

By the definition of Wasserstein distance, computing $\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m)$ is equivalent to the following max-min problem:

$$\max_{f \in \mathcal{H}:\ \|f\|^2_{\mathcal{H}} \leq 1} \left\{ \min_{\pi \in \Gamma} \sum_{i,j} \pi_{i,j} \|f(x_i) - f(y_j)\|_2^2 \right\}, \tag{2.2}$$

where $\Gamma = \left\{ \pi \in \mathbb{R}_+^{n \times m} :\ \sum_j \pi_{i,j} = \frac{1}{n}, \sum_i \pi_{i,j} = \frac{1}{m} \right\}$.

The computation of KPW distance has numerous challenges. It is crucial to design a suitable kernel function to obtain low computational complexity and reliable testing power, which will be discussed in Section 2.5. Moreover, the function $f \in \mathcal{H}$ is a countable combination of basis functions, i.e., the problem (2.2) is an infinite-dimensional optimization. By developing the representer theorem in Theorem 1, we are able to convert this problem into a finite-dimensional problem. Finally, there is no theoretical guarantee for finding the global

optimum since it is a non-convex non-smooth optimization problem. Moreover, Sion's minimax theorem is not applicable because the problem (2.2) is not a convex programming: the inner minimization of quadratic function makes the objective in (2.2) not concave in f in general. Based on this observation, we only focus on optimization algorithms for finding a local optimum point in polynomial time.

Theorem 1 (Representer Theorem for KPW Distance)**.** *There exists an optimal solution to* (2.2) *that admits the following expression:*

$$\hat{f} = \sum_{i=1}^{n} K_{x_i} a_{x,i} - \sum_{j=1}^{m} K_{y_j} a_{y,j},$$

where $K_x(\cdot)$ denotes the kernel section and $a_{x,i}, a_{y,j} \in \mathbb{R}^d$ for $i = 1, \ldots, n, j = 1, \ldots, m$ are coefficients to be determined.

The proof of Theorem 1 is provided in Appendix A.4, in which standard representer theorem in literature [137, Theorem 1] is not applicable since the RKHS norm serves as a hard constraint instead of the regularization of the objective function. In order to express the optimal solution as the compact matrix form, define $a_x \in \mathbb{R}^{nd}$ as the concatenation of coefficients $a_{x,i}$ for $i = 1, \ldots, n$ and

$$K_z(x^n) = \begin{pmatrix} K(z, x_1) & \cdots & K(z, x_n) \end{pmatrix} \in \mathbb{R}^{d \times nd}.$$

We also define the vector a_y and matrix $K_z(y^m)$ likewise. Then we have

$$\hat{f}(z) = K_z(x^n) a_x - K_z(y^m) a_y, \quad \forall z \in \mathcal{X}.$$

Define the gram matrix $K(x^n, x^n)$ as the $n \times n$ block matrix with the (i, j)-th block being $K(x_i, x_j)$. The gram matrices $K(x^n, y^m), K(y^m, x^n)$ and $K(y^m, y^m)$ can be defined

likewise. Denote by G the concatenation of gram matrices:

$$G = \begin{pmatrix} K(x^n, x^n) & -K(x^n, y^m) \\ -K(y^m, x^n) & K(y^m, y^m) \end{pmatrix},$$

and we assume that G is positive definite. Otherwise, we add the gram matrix with a small number times identity matrix to make it invertible. Substituting the expression of $\hat{f}(z), z \in \mathcal{X}$ into (2.2), we obtain a finite-dimensional optimization problem:

$$\max_{\omega} \left\{ \min_{\pi \in \Gamma} \sum_{i,j} \pi_{i,j} c_{i,j} : \omega^{\mathrm{T}} G \omega \leq 1 \right\},$$

where $\omega = [a_x^{\mathrm{T}}, a_y^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{d(n+m)}$, $c_{i,j} = \|A_{i,j}\omega\|_2^2$, and

$$A_{i,j} = [K_{x_i}(x^n) - K_{y_j}(x^n), K_{y_j}(y^m) - K_{x_i}(y^m)].$$

Suppose that the inverse of G admits the Cholesky decomposition $G^{-1} = UU^{\mathrm{T}}$, then by the change of variable technique $s = U^{-1}\omega$, we obtain the norm-constrained optimization problem:

$$\max_{s \in \mathbb{R}^{d(n+m)}} \left\{ \min_{\pi \in \Gamma} \sum_{i,j} \pi_{i,j} c_{i,j} : s^{\mathrm{T}} s \leq 1 \right\}, \tag{2.3}$$

and we can replace the constraint $s^{\mathrm{T}}s \leq 1$ with $s^{\mathrm{T}}s = 1$ based on the fact that the norm function satisfies the linear property. In other words, the decision variable s belongs to the Euclidean ball $\mathbb{S}^{d(n+m)-1} = \{s \in \mathbb{R}^{d(n+m)} : s^{\mathrm{T}}s = 1\}$.

For the ease of optimization, we consider the entropic regularization of the problem (2.3):

$$\max_{s \in \mathbb{S}^{d(n+m)-1}} \left\{ \min_{\pi \in \Gamma} \sum_{i,j} \pi_{i,j} c_{i,j} - \eta H(\pi) \right\}, \tag{2.4}$$

in which we denote the entropy function $H(\pi) = -\sum_{i,j} \pi_{i,j}(\log \pi_{i,j} - 1)$. By the duality theory of entropic optimal transport [64] and the change-of-variable technique, (2.4) is

equivalent to the following minimization problem:

$$\min_{s\in\mathbb{S}^{d(n+m)-1},u\in\mathbb{R}^n,v\in\mathbb{R}^m} F(u,v,s), \tag{2.5}$$

where

$$\begin{aligned} c_{i,j} &= \|A_{i,j}Us\|_2^2, \\ \pi_{i,j}(u,v,s) &= \exp\left(-\frac{1}{\eta}c_{i,j}+u_i+v_j\right), \\ F(u,v,s) &= \sum_{i,j}\pi_{i,j}(u,v,s) - \frac{1}{n}\sum_{i=1}^{n}u_i - \frac{1}{m}\sum_{j=1}^{m}v_j. \end{aligned}$$

The details for this deviation is deferred in Appendix A.4. Based on this formulation, we consider a Riemannian block coordinate descent (BCD) method [77] for optimization, which updates a block of variables by minimizing the objective function with respect to that block while fixing values of other blocks:

$$u^{t+1} = \min_{u\in\mathbb{R}^n} F(u,v^t,s^t), \tag{2.6a}$$

$$v^{t+1} = \min_{v\in\mathbb{R}^m} F(u^{t+1},v,s^t), \tag{2.6b}$$

$$\zeta^{t+1} = \sum_{i,j}\nabla_s\pi_{i,j}(u^{t+1},v^{t+1},s^t), \tag{2.6c}$$

$$\xi^{t+1} = \mathcal{P}_{s^t}\big(\zeta^{t+1}\big), \tag{2.6d}$$

$$s^{t+1} = \mathrm{Retr}_{s^t}\big(-\tau\xi^{t+1}\big), \tag{2.6e}$$

where the operator $\mathcal{P}_s(\zeta)$ denotes the orthogonal projection of the vector ζ onto the tangent space of the manifold $\mathbb{S}^{d(n+m)-1}$ at s:

$$\mathcal{P}_s\big(\zeta\big) = \zeta - \langle s,\zeta\rangle s, \qquad s\in\mathbb{S}^{d(n+m)-1},$$

Algorithm 2 BCD Algorithm for Solving (2.5)

Require: Empirical distributions $\hat{\mu}_n$ and $\hat{\nu}_m$.

1: Initialize v^0, s^0
2: **for** $t = 0, 1, 2, \ldots, T-1$ **do**
3: Update u^{t+1} according to (2.6g)
4: Update v^{t+1} according to (2.6h)
5: Update the Euclidean and Riemannian gradient ζ^{t+1} and ξ^{t+1}, according to (2.6i) and (2.6d), respectively.
6: Update s^{t+1} according to (2.6e)
7: **end for**

Return $u^* = u^T, v^* = v^T, s^* = s^T$**.**

and the retraction on this manifold is defined as

$$\text{Retr}_s\big(-\tau\xi\big) = \frac{s - \tau\xi}{\|s - \tau\xi\|}, \qquad s \in \mathbb{S}^{d(n+m)-1}. \tag{2.6f}$$

Note that the update steps (2.6a) and (2.6b) have closed-form expressions:

$$u^{t+1} = u^t + \left\{ \log \frac{1/n}{\sum_j \pi_{i,j}(u^t, v^t, s^t)} \right\}_{i \in [n]}, \tag{2.6g}$$

$$v^{t+1} = v^t + \left\{ \log \frac{1/m}{\sum_i \pi_{i,j}(u^{t+1}, v^t, s^t)} \right\}_{j \in [m]}, \tag{2.6h}$$

and the Euclidean gradient ζ^{t+1} in (2.6c) can be computed using the chain rule:

$$\zeta^{t+1} = -\frac{1}{\eta} U^{\mathrm{T}} \left[\sum_{i,j} \pi_{i,j}(u^{t+1}, v^{t+1}, s^t) A_{i,j}^{\mathrm{T}} A_{i,j} \right] U s^t. \tag{2.6i}$$

The overall algorithm for solving the problem (2.5) is summarized in Algorithm 2. We provide details for efficient implementation of the proposed algorithms in Appendix A.6. We also give a brief introduction to Riemannian optimization in Appendix A.2. The following theorem gives a convergence analysis of our proposed algorithm.The proof of this result is provided in Appendix A.4, which follows similar procedure in [85]. The main difference lies in establishing the descent lemma for updating the variable s on sphere instead of Stiefel manifold. Specifically, the procedure for finding the upper bound on the cost function $c_{i,j}$,

the Lipschitz constant for $\pi_{i,j}(u, v, s)$ in s, and the Lipschitz constants of the retraction operator (2.6f) will be different.

Theorem 2 (Convergence Analysis for BCD)**.** *We say that* $(\hat{u}, \hat{v}, \hat{s})$ *is a* (ϵ_1, ϵ_2)*-stationary point of* (2.5) *if*

$$\|\mathit{Grad}_s F(\hat{u}, \hat{v}, \hat{s})\| \leq \epsilon_1,$$
$$F(\hat{u}, \hat{v}, \hat{s}) - \min_{u,v} F(u, v, \hat{s}) \leq \epsilon_2,$$

where $\mathit{Grad}_s F(u, v, s)$ *denotes the derivative of* F *with respect to* s *on the sphere* $\mathbb{S}^{d(n+m)-1}$*. Let* $\{u^t, v^t, s^t\}$ *be the sequence generated by Algorithm 2, then Algorithm 2 returns an* (ϵ_1, ϵ_2)*-stationary point in*

$$T = \mathcal{O}\left(\log(mn) \cdot \left[\frac{1}{\epsilon_2^3} + \frac{1}{\epsilon_1^2 \epsilon_2}\right]\right),$$

iterations, where the notation $O(\cdot)$ *hides constants related to the initial guess* (v^0, s^0) *and the term* $\max_{i,j} \|A_{i,j} U\|$*.*

Remark 2 (Complexity of Algorithm 2)**.** *Denote* $N = n \vee m$[1]*. Note that the iteration* (2.6g) *and* (2.6h) *can be implemented in* $O(N)$ *iterations. Second, the retraction step in* (2.6e) *requires* $O(dN)$ *arithmetic operations. Third, the computation of the Euclidean vector in* (2.6c) *can be implemented in* $O(d^3 N^3)$ *operations, and the projection step can be done in* $O(dN)$ *operations. Therefore, the number of arithmetic operations in each iteration is of* $O(d^3 N^3)$*. In summary, Algorithm 2 returns an* (ϵ_1, ϵ_2)*-stationary point in*

$$\mathcal{O}\left(d^3 N^3 \log(N) \cdot \left[\frac{1}{\epsilon_2^3} + \frac{1}{\epsilon_1^2 \epsilon_2}\right]\right)$$

arithmetic operations. Note that this computational complexity is independent of the dimension D *of samples since we only need to compute the gram matrix as an input. The*

[1]We denote $a \vee b$ for $\max\{a, b\}$ and $a \wedge b$ for $\min\{a, b\}$.

storage cost is of $\mathcal{O}(d^2N^2)$, in which the most expensive step is to store the gram matrix G.

2.4 Performance Guarantees

In this section, we build statistical properties of the empirical KPW distance, though in practice we may not succeed in finding a global optimum solution to the non-convex optimization problem (2.2). We assume the cost function for the Wasserstein distance has the form $c(x, y) = \|x - y\|_2^p$ with $p \in [1, \infty)$. Moreover, results throughout this section are based on the following assumption.

Assumption 1. *For any $x, x' \in \mathcal{X}$, the matrix-valued kernel $K(x, x')$ is symmetric and satisfies*

$$0 \preceq K(x, x') \preceq BI_d.$$

Definition 5 ((Projection) Poincare Inequality)**.** (I) *A distribution μ is said to satisfy a Poincare inequality if there exists an $M > 0$ for $X \sim \mu$ so that Var$[f(X)] \leq M\mathbb{E}[\|\nabla f(X)\|^2]$ for any f satisfying $\mathbb{E}[f(X)^2] < \infty$ and $\mathbb{E}[\|\nabla f(X)\|^2] < \infty$.*

(II) *A distribution μ is said to satisfy a projection Poincare inequality if there exists an $M > 0$ for any $f \in \mathcal{F}$ and $X \sim f\#\mu$ so that Var$[f(X)] \leq M\mathbb{E}[\|\nabla f(X)\|^2]$ for any f satisfying $\mathbb{E}[f(X)^2] < \infty$ and $\mathbb{E}[\|\nabla f(X)\|^2] < \infty$.*

Remark 3. *The Poincare inequality characterizes the relation about the variance of a function and its derivative in the spirit of the Sobolev inequality. It is a standard technical assumption for investigating the empirical convergence of Wasserstein distance [96, 100], and is satisfied for various exponential measures such as the Gaussian distribution. See [95] for more examples.*

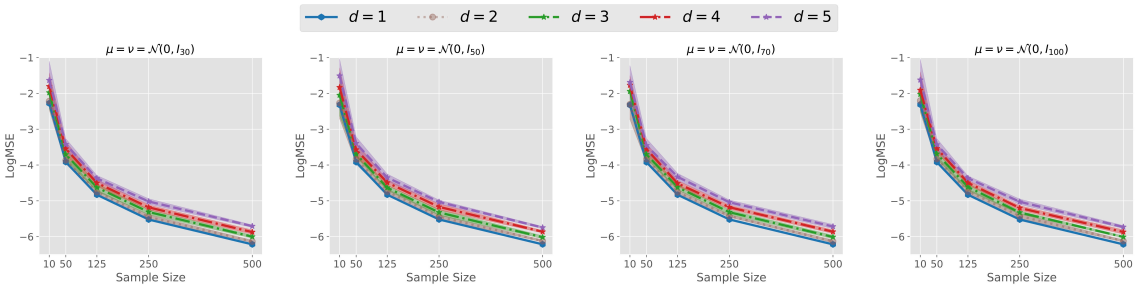

Figure 2.1: Average values of KPW distances between empirical distributions $\hat{\mu}_n$ and $\hat{\nu}_n$ as the sample size n varies. Results are averaged for 10 independent trials and the shaded areas show the corresponding error bars.

Lemma 1. *Assume that the distribution μ satisfies a projection Poincare inequality. Then*

$$\mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p}] \lesssim n^{-\frac{1}{(2p)\vee d}} (\log n)^{\zeta_{p,d}/p} + n^{-1/(2\vee p)} \sqrt{\log(n)} + n^{-1/p} \log(n),$$

where $\zeta_{p,d} = 1\{d = 2p\}$, and $\lesssim$ refers to "less than" with a constant depending only on (p, B).

Lemma 2. *Assume that the distribution μ satisfies a Poincare inequality, and any $f \in \mathcal{F}$ is L-Lipschitz. Then with probability at least $1 - \alpha$, it holds that*

$$\left| (\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p} - \mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p}] \right| \leq \max\left\{ \varrho \log(1/\alpha), \sqrt{\varrho \log(1/\alpha)} \right\} n^{-1/(2\vee p)} L^{1/p},$$

where $\varrho > 0$ is a constant that depends on M.

Proof of two lemmas above follows similar covering number arguments in [100], the details of which are deferred in Appendix A.5. The main difference is that we incorporate the reproducing property of vector-valued RKHS to give a valid bound on the covering number of the RKHS ball $\mathcal{F}$. Based on these two lemmas and the triangular inequality for Wasserstein distance, we give a finite-sample guarantee for the convergence of the KPW

distance in Theorem 3. Compared with the sample complexity of estimating Wasserstein distance, KPW distance does not suffer from the curse of dimensionality as the RKHS ball $\mathcal{F}$ has low complexity.

Theorem 3 (Finite-sample Guarantee)**.** *Suppose the target distributions $\mu = \nu$, which satisfies projection Poincare inequality and Poincare inequality. Moreover, any $f \in \mathcal{F}$ is L-Lipschitz. Take $N = n \wedge m$, then with probability at least $1 - 2\alpha$, it holds that*

$$\begin{aligned}(\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m))^{1/p} \lesssim\ & N^{-\frac{1}{(2p)\vee d}} (\log N)^{\zeta_{p,d}/p} \\ & + N^{-1/(2\vee p)} \sqrt{\log(N)} + N^{-1/p} \log(N) \\ & + \max\left\{ \varrho \log(1/\alpha), \sqrt{\varrho \log(1/\alpha)} \right\} N^{-1/(2\vee p)} L^{1/p}.\end{aligned}$$

2.4.1 Performance Guarantees for $p \in [1, 2)$

When showing concentration results for p-Wasserstein distance with $p \in [1, 2)$, however, it is not necessary to rely on the Poincare inequality assumption. The main result for this case is summarized in Theorem 4 (see details in Appendix A.5.3).

Theorem 4 (Finite-sample Guarantee)**.** *Suppose the target distributions $\mu = \nu$. Then with probability at least $1 - 2\alpha$, it holds that*

$$\begin{aligned}(\mathcal{KPW}(\hat{\mu}_n, \nu_m))^{1/p} \lesssim\ & N^{-\frac{1}{(2p)\vee d}} (\log N)^{\zeta_{p,d}/p} \\ & + N^{1/2-1/p} \sqrt{\log(N)} + N^{-1/p} \\ & + N^{1/2-1/p} \sqrt{\log \frac{2}{\alpha}}.\end{aligned}$$

where $N = n \wedge m$ and $\lesssim$ refers to "less than" with a constant depending only on (p, B).

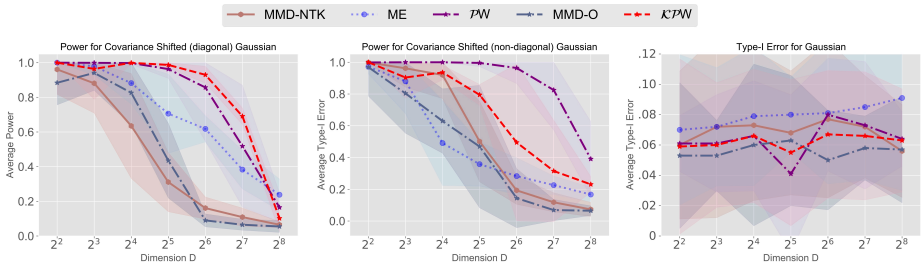

Figure 2.2: Testing results on Gaussian distributions across different choices of dimension D. Left: power for Gaussian distributions, where the shifted covariance matrix is still diagonal; Middle: power for Gaussian distributions, where the shifted covariance matrix is non-diagonal; Right: Type-I error.

Table 2.1: Average test power and standard error about detecting distribution abundance change in *MNIST* dataset across different choices of sample size.

N	MMD-NTK	MMD-O	ME	PW	KPW
200	0.639±0.029	**0.696±0.006**	0.298±0.031	0.302±0.033	0.663±0.015
250	0.763±0.010	0.781±0.002	0.472±0.017	0.369±0.030	**0.785±0.014**
300	0.813±0.016	0.869±0.002	0.630±0.025	0.524±0.023	**0.928±0.001**
400	0.881±0.013	0.956±0.003	0.779±0.020	0.591±0.044	**0.978±0.000**
500	0.950±0.002	0.988±0.000	0.927±0.006	0.782±0.040	**1.000±0.000**
Avg.	0.809	0.858	0.621	0.513	**0.870**

2.4.2 Sample Complexity

We also numerically examine the sample complexity of the empirical KPW distance $\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_n)$ with $\mu = \nu = \mathcal{N}(0, I_D)$, where $n \in \{10, 50, 125, 250, 500\}$ and $D \in \{30, 50, 70, 100\}$. Figure 2.1 reports the average distances and the shaded areas show the corresponding error bars over 10 independent trials. We defer the detailed experiment setup and the plots of the computation time in Appendix A.7.1. From the plot we can see that the empirical KPW distances decay to zero quickly when the sample size n increases. Moreover, the distances with smaller values of d have faster decaying rates. Finally, the convergence behavior of the empirical KPW distances is nearly independent of the choice of D, which alleviates the issue of the curse of dimensionality for the original Wasserstein distance. These facts confirm the finite-sample guarantee discussed in Theorem 3.

2.5 Numerical Experiments

Throughout this section, we compare the performance of tests with the following procedures. (i) PW: the projected Wasserstein test where the projector is a linear mapping [157]; (ii) MMD-O: the MMD test with a Gaussian kernel whose bandwidth is optimized [101]; (iii) MMD-NTK: the test that combines both neural networks and MMD [37]; and (iv) ME: the mean embedding test with optimized hyper-parameters [88]. Implementation details on those baseline methods are omitted in Appendix A.7.2. When dealing with synthetic datasets, we generate a single sample set as the training set to learn parameters for each method. Then we evaluate the power of tests on 100 new sample sets generated from the same distribution. When dealing with real datasets, we randomly take part of samples as the training set, and evaluate the power on 100 randomly chosen subsets from the remaining samples. The number of permutations in Algorithm 1 is set to be $N_p = 100$. We control the type-I error for all tests at $\alpha = 0.05$.

When using the KPW distance, we follow (2.1) to design kernels to decrease the computational complexity. More specifically, we choose the scalar-valued kernel $k(\cdot, \cdot)$ to be a standard Gaussian kernel with the bandwidth σ^2, and

$$P = (1 - \rho)\mathbf{1}\mathbf{1}^{\mathrm{T}} + \rho I_d, \quad \text{with } \rho \in [0, 1].$$

We use the cross-validation approach to select the hyper-parameters ρ and σ^2, the details of which are deferred in Appendix A.7.3. The dimension d is pre-specified and fixed into 3 in all experiments. We also present a study on the impact of hyper-parameters such as the projected dimension d and regularization parameter η in Appendix A.8.

2.5.1 Tests for Synthetic Datasets

We first investigate the performance when μ and ν are Gaussian distributions with diagonal covariance matrices. Specifically, we take $\mu = \mathcal{N}(0, I_D)$ and $\nu = \mathcal{N}(0, \Sigma)$ is the covariance

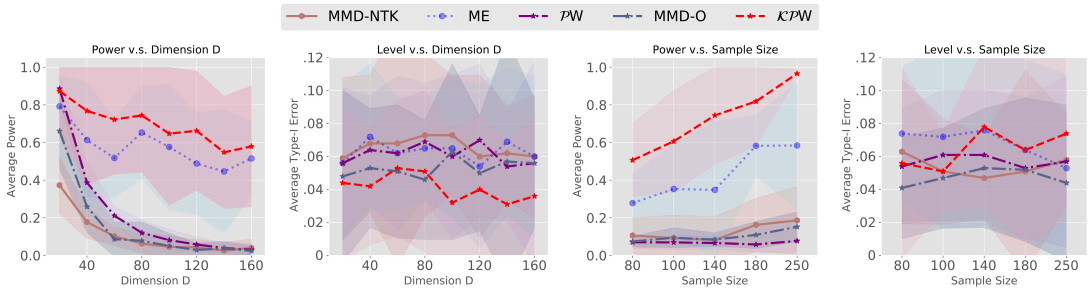

Figure 2.3: Testing results on Gaussian-mixture distributions. Left two: type-I and type-II errors across different choices of dimension D with fixed sample size $n = m = 200$; Right two: type-I and type-II errors across different choices of sample size $n = m$ with fixed dimension $D = 140$.

shifted Gaussian, where the matrix $\Sigma = \mathrm{diag}(4, 4, 4, 1, \ldots, 1)$. In other words, we only scale the first three entries of the covariance matrix to make the high-dimensional testing problem challenging to handle. Fig. 2.2 reports the type-I and type-II errors for various tests across different choices of dimension D. We observe that both PW and KPW tests perform the best, while the power for other benchmark methods degrades quickly when the dimension D increases.

Next, we examine the case where ν has a non-diagonal covariance matrix. We take $\mu = \mathcal{N}(0, I_D)$ and $\nu = \mathcal{N}(0, V\Sigma V^{\mathrm{T}})$, where V is an orthogonal matrix with $V_{i,j} = \sqrt{2/(D+1)}\sin(ij\pi/(D+1))$ and $\Sigma = \mathrm{diag}(5, 5, 5, 1, \ldots, 1)$. Testing results for various choices of dimension D is reported in the middle of Fig. 2.2. In this case, the PW test performs slightly better than the KPW test. One possible explanation is that linear mapping seems to be the optimal choice for two-sample testing with covariance shifted Gaussian distributions. It is promising to design other types of matrix-valued kernel functions to improve performances of the KPW test.

Finally, we study the case where sample points are generated from high-dimensional Gaussian mixture distributions. We take $\mu = \frac{1}{2}\mathcal{N}(0, I_D) + \frac{1}{2}\mathcal{N}(\Delta_2, I_D)$ with $\Delta_2 = (1, 1, \ldots, 1)$ and $\nu = \frac{1}{2}\mathcal{N}(0, \Sigma_1) + \frac{1}{2}\mathcal{N}(\Delta_3, \Sigma_2)$ with $\Delta_3 = (1+0.8/\sqrt{D}, \ldots, 1+0.8/\sqrt{D})$. Covariance matrix Σ_1 is defined with $\Sigma_1[1,1] = \Sigma_1[2,2] = 4, \Sigma_1[1,2] = \Sigma_1[2,1] = -0.9, \Sigma_1[i,i] = 1, 3 \leq i \leq D$, and $\Sigma_1[i,j] = 0$ for indexes elsewhere. Covariance matrix

Σ_2 is defined with $\Sigma_2[1,2] = \Sigma_2[2,1] = 0.9$, $\Sigma_2[i,i] = 1, 1 \leq i \leq D$, and $\Sigma_2[i,j] = 0$ for indexes elsewhere. Testing results (type-I and type-II errors) across different choices of dimension D for fixed sample size $n = m = 200$ is presented in the left two plots in Fig. 2.3. We also report results for increasing sample sizes $n = m$ by fixing the dimension $D = 140$ in the right two plots in Fig. 2.3. From the plot, we can see that all approaches have expected type-I error rates. Moreover, the tests based on PW and KPW distances outperform other benchmark methods, which indicates that the idea of dimension reduction is helpful for high-dimensional testing. The KPW test generally has the highest power in this case, since the nonlinear projector in the unit ball of RKHS is flexible enough to capture the differences between distributions. Other experiment details of this subsection is omitted in Appendix A.7.4.

2.5.2 Tests for MNIST handwritten digits

We now perform two-sample tests on the MNIST dataset [94]. Let p be the distribution uniformly generated from the dataset, and $q = 0.85p + 0.15p_{\text{cohort}}$, where p_{cohort} is the distribution from a class with digit 1. Both training and testing sample sizes are set to be $N \in \{200, 250, \ldots, 500\}$. Before performing two-sample tests, we pre-process this dataset by taking the sigmoid transformation of each image such that all scaled pixels are within the interval $[0, 1]$. Table 2.1 presents the testing power of various tests across different choices of N, from which we can see that the KPW test is competitive compared with other methods. We observe that performances of MMD-O in MNIST dataset are significantly better than that in synthetic datasets provided in Section 2.5.1. One possible explanation is that isotropic kernel functions will limit the power of MMD tests in some numerical examples [101, Section 3]. Average type-I error for various tests is presented in Table A.1 in Appendix A.7.5, from which we can see all tests have the type-I error close to $\alpha = 0.05$.

Table 2.2: Delay time for detecting the transition in *MSRC-12* that corresponds to four users.

User	MMD-NTK	MMD-O	ME	PW	KPW
1	36	73	82	47	**33**
2	8	7	97	9	**1**
3	15	13	27	**2**	20
4	22	83	69	16	**12**
Mean	20.25	44.0	68.8	18.50	**16.5**
Std	**12.0**	39.5	30.1	19.8	13.5

2.5.3 Human activity detection

Finally, we apply the KPW test to perform online change-point detection for human activity transition. We use a real-world dataset called the Microsoft Research Cambridge-12 (MSRC-12) Kinect gesture dataset [56]. After pre-processing, this dataset consists of actions from four people, each with 855 samples in $\mathbb{R}^{60}$, and with a change of action from *bending* to *throwing* at the time index 500. More experimental details are omitted in Appendix A.7.6. Fix the window size $W = 100$. We pre-train a nonlinear projector using the data (sample size as the window) before time index 300 and compute the null statistics for many times to obtain the true threshold such that the false alarm rate is controlled within $\alpha = 0.05$. Then we perform online change-point detection based on a sliding window that moves forward with time. We compute the detection statistic by comparing the distribution between the block of data before time 300 and the data from the sliding window. We reject the null hypothesis and claim a change is happened if the statistic is above the threshold. Table 2.2 reports the delay time for detecting the behavior transition, from which we observe that the KPW test detects the change in the shortest time.

2.6 Conclusion

We proposed the KPW distance for the task of two-sample testing, which operates by finding the nonlinear mapping in the data space to maximize the distance between projected distributions. Practical algorithms together with uncertainty quantification of empirical estimates are discussed to help with this task.

The extension of this work is as follows. First, it is promising to consider milder technical assumptions than the projected Poincare inequality when establishing performance guarantees. Second, a meaningful research question is to determine the optimal hyper-parameters for the KPW test, including the projected subspace dimension d and the matrix-valued kernel function K. Third, it is desirable to study how to systematically pick the regularization parameter η to balance the trade-off between computational efficiency and accuracy of the obtained solution.

CHAPTER 3
SINKHORN DISTRIBUTIONALLY ROBUST OPTIMIZATION

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 a general nominal distribution. Compared with Wasserstein DRO, it is computationally tractable for a larger class of loss functions, and its worst-case distribution is more reasonable. To solve the dual reformulation, we propose an efficient stochastic mirror descent algorithm using biased gradient oracles. It finds a δ-optimal solution with computation cost $\tilde{O}(\delta^{-3})$ and memory cost $\tilde{O}(\delta^{-2})$, and the computation cost further improves to $\tilde{O}(\delta^{-2})$ when the loss function is smooth. Finally, we provide various numerical examples using synthetic and real data to demonstrate its superior performance.

3.1 Introduction

Decision-making problems under uncertainty have broad applications in operations research, machine learning, engineering, and economics. When the data involves uncertainty due to measurement error, insufficient sample size, contamination, anomalies, or model misspecification, distributionally robust optimization (DRO) is a promising data-driven approach. It seeks a minimax robust optimal decision that minimizes the expected loss under the most adverse distribution within a given set of relevant distributions, called an ambiguity set. It provides a principled framework to produce a solution with more promising out-of-sample performance than the traditional sample average approximation (SAA) method for stochastic programming [142]. We refer to [128] for a recent survey on DRO.

At the core of DRO is the choice of the ambiguity set. Ideally, a good ambiguity set should induce a computationally tractable formulation while taking practical interpretability into account. And it should be rich enough to contain distributions relevant to the decision-

making while, in meantime, should exclude unnecessary distributions that may lead to overly conservative decisions. Various DRO formulations have been proposed in the literature. Among them, the ambiguity set based on Wasserstein distance has recently received much attention [22, 61, 110, 164]. The Wasserstein distance incorporates the geometry of sample space and thereby is suitable for comparing distributions with non-overlapping supports and hedging against data perturbations [61]. Nice statistical performance guarantees have been established theoretically [21, 23, 24, 59, 140] and empirically in a variety of applications in operations research [18, 38, 117, 146, 147, 160], machine learning [19, 32, 106, 116, 141, 152], stochastic control [1, 50, 150, 159, 168, 169], etc; see [93] and references therein for more discussions.

On the other hand, the current Wasserstein DRO framework is not without limitations. First, from the *computational efficiency* perspective, the tractability of Wasserstein DRO is usually available only 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. In Table 3.1, we summarize known tractable cases for solving the Wasserstein DRO $\min_{\theta\in\Theta}\max_{\mathbb{P}\in\mathcal{P}}\ \mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)]$, where the loss function $f_\theta(z)$ is convex in θ with a closed and convex feasible region Θ, and the ambiguity set is centered around a nominal distribution $\widehat{P}$ and contains distributions supported on a space $\mathcal{Z}$. A general approach to solve the Wasserstein DRO is to use a finite and discrete grid of scenarios to approximate the whole sample space and solve the formulation restricted on the discrete sample space. Specially, for 1-Wasserstein DRO, a convex reformulation is only known when the loss function can be expressed as a pointwise maximum of finitely many concave functions or generalized linear model [55, 61, 98, 139, 141, 171]; for general p-Wasserstein DRO, efficient first-order algorithms have been developed only for smooth loss functions and sufficiently small radius (or equivalently, sufficiently large Lagrangian multiplier) so that the involved subproblem becomes strongly convex [25, 148]. Second, from the *modeling* perspective, for data-driven Wasserstein DRO in which the nominal distribution is finitely

supported (usually the empirical distribution), the worst-case distribution is shown to be a discrete distribution [61] (which is unique when the regularized loss function has a unique maximizer), despite that the underlying true distribution in many practical applications may well be continuous. This raises the concern of whether Wasserstein DRO hedges the right family of distribution and whether it induces over-conservative solutions.

Table 3.1: Existing tractability result of Wasserstein DRO

Reference(s)	Loss function $f_\theta(z)$	Cost function	Nominal distribution $\widehat{P}$	Support $\mathcal{Z}$
[33, 102, 173]	General	General	General	Discrete and finite set
[148]	$z \mapsto f_\theta(z) - \lambda^* c(x, z)$ is strongly concave[1]	General	General	General
[55, 61]	Piecewise concave in z	Norm function	Empirical distribution	Polytope
[98, 139, 141, 171]	Generalized linear model in (z, θ)	Norm function	Empirical distribution	Whole Euclidean space [2]
[25]	Generalized linear model in (z, θ)	Squared norm function	General	Whole Euclidean space

To address these potential issues while maintaining the advantages of Wasserstein DRO, in this paper, we propose Sinkhorn DRO, which hedges against distributions that are close to some nominal distribution in the Sinkhorn distance [46]. The Sinkhorn distance can be viewed as a smoothed Wasserstein distance, defined as the cheapest transport cost between two distributions associated with an optimal transport problem with entropic regularization (see Definition 6 in Section 3.2). As far as we know, this paper is the first to study the DRO formulation using the Sinkhorn distance. Our main contributions are summarized as follows.

(I) We derive a strong duality reformulation for Sinkhorn DRO (Theorem 5) when the nominal distribution is any arbitrary distribution. The Sinkhorn dual objective smooths the maximization subproblem in the Wasserstein dual objective, and converges to Wasserstein dual objective as the entropic regularization parameter goes to zero (Remark 7). Moreover, 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 8).

(II) As a byproduct of our duality proof, we characterize the worst-case distribution of the Sinkhorn DRO (Remark 6), which is absolutely continuous with respect to some reference measure such as Lebesgue or counting measure. Compared with Wasserstein DRO, the worst-case distribution of Sinkhorn DRO is not necessarily finitely supported even when the nominal distribution is a finitely supported distribution. This indicates that Sinkhorn DRO is a more flexible modeling choice for many applications.

(III) On the algorithmic aspect, we propose and analyze an efficient stochastic mirror descent method using biased gradient oracles with bisection search for solving the Sinkhorn DRO problem (Section 3.4). By adequately balancing the trade-off between bias and variance of stochastic gradient estimators with low computation cost, we show the proposed algorithm achieves computation cost $\tilde{O}(\delta^{-3})$ and memory cost $\tilde{O}(\delta^{-2})$ for finding δ-optimal solution for convex loss, and the computation cost improves to $\tilde{O}(\delta^{-2})$ for convex and smooth loss.[3] Compared with Wasserstein DRO, the dual problem of Sinkhorn DRO is computationally tractable for a broader class of loss functions, cost functions, nominal distributions, and probability support.

(IV) We provide experiments (Section 3.5) to validate the performance of the proposed Sinkhorn DRO model in the context of newsvendor problem, mean-risk portfolio optimization, and multi-class classification, using both synthetic and real data sets. Numerical results demonstrate its superior out-of-sample performances and fast computational speed compared with several benchmarks including SAA, Wasserstein DRO, and KL-divergence DRO.

Finally, we remark that Blanchet and Kang [20, Section 3.2] solves the Wasserstein DRO formulation based on its log-sum-exp smooth approximation. This smoothed approximation

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

can be viewed as a special case of the dual reformulation of our Sinkhorn DRO model. The main differences between their formulation and ours are that: (i) we start with the primal form of the Sinkhorn DRO model and uncover it coincides with the smoothed approximation of the Wasserstein DRO dual formulation, while they focus on the dual formulation only. (ii) they focus on the data-driven DRO training for semi-supervised learning tasks only, while our DRO result applies to the general loss function, cost function, and nominal distribution. (iii) they optimize the smoothed objective function by simulating unbiased gradient estimators but with unbounded variance, and no convergence results are established.

Azizian et al. [7] (which is a contribution made public 234 days after we posted the first version of this paper in arXiv) present a very similar duality result shown in this paper. The main differences include the following: (i) their theoretical results rely on a Slater-like assumption which is more restrictive. (ii) they do not provide numerical algorithm to solve the Sinkhorn DRO formulation.

Related Literature

On DRO Models The construction of ambiguity sets plays a key role in the performance of DRO models. Generally, there are two ways to construct ambiguity sets in literature. First, ambiguity sets can be defined using descriptive statistics, such as the support information [14], moment conditions [15, 35, 48, 71, 135, 163, 175], shape constraints [125, 153], marginal distributions [2, 51, 58, 113]. Second, a more recently popular approach that makes full use of the available data is to consider distributions within a pre-specified statistical distance from a nominal distribution, usually chosen as the empirical distribution of samples. Commonly used statistical distances used in literature include ϕ-divergence [11, 12, 52, 84, 161], Wasserstein distance [22, 34, 61, 110, 122, 164, 167, 173], and maximum mean discrepancy [151, 174]. Our proposed Sinkhorn DRO can be viewed as a variant of Wasserstein DRO. In the literature on Wasserstein DRO, besides the computational tractability, its regularization effects and statistical inference have also be investigated.

In particular, it has been shown that Wasserstein DRO is asymptotically equivalent to a statistical learning problem with variation regularization [21, 60, 140], and when the radius is chosen properly, the worst-case loss of Wasserstein DRO serves as an upper confidence bound on the true loss [21, 23, 24, 59] . Other variants of Wasserstein DRO have been explored, by combining with other information such as moment information [63, 156] and marginal distributions [53, 62].

On Sinkhorn Distance Sinkhorn distance [46] is proposed to improve the computational complexity of Wasserstein distance by regularizing the original mass transportation problem with relative entropy penalty on the transport mapping. 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 [120]. Such an approach first aroused in the areas of economics and survey statistics [8, 49, 92, 172], and its convergence analysis is attributed to the mathematician Sinkhorn [149], which gives the name of Sinkhorn distance. Altschuler et al. [4] further design an accelerated algorithm to compute Sinkhorn distance in near-linear time. Using Sinkhorn distance other than Wasserstein distance has been demonstrated to be beneficial because of lower computational cost in various applications, including domain adaptations [41, 42, 43], generative modeling [68, 105, 118, 119], dimensionality reduction [86, 99, 157, 158], etc. To the best of our knowledge, the study of Sinkhorn distance for distributionally robust optimization is new in literature.

Solving DRO Models In the introduction, we have elaborated on the literature that propose efficient optimization algorithms for solving the Wasserstein DRO formulation [25, 33, 55, 61, 98, 102, 139, 141, 148, 171, 173]. Unfortunately, the tractability of these literature is limited to the special loss function, cost function, nominal distribution, or probability support. In addition, the algorithmic framework for ϕ-divergence DRO model is also exploited in recent literature. A natural optimization idea is to generate sample estimates of the dual formulation of ϕ-divergence DRO and then optimize the approximated

objective function [143, Section 7.5.4], called the *sample average approximation (SAA)* technique. It is worth noting that the SAA technique is not a computation- and storage-efficient choice, since it requires storing the input data for the approximated problem first, and then solving the new problem numerically. Recent literature [97, 112, 127] propose first-order methods to solve ϕ-divergence DRO formulations. In comparison with the SAA technique, the complexity of first-order methods is usually independent of the sample size of the nominal distribution to obtain a near-optimal solution. Motivated by these literature, we propose to solve the Sinkhorn DRO model by simulating stochastic gradient estimators and then establish sample-size independent convergence results.

The rest of the paper is organized as follows. In Section 3.2, we describe the main formulation for the Sinkhorn DRO model. In Section 3.3, we develop its strong dual reformulation. In Section 3.4, we propose a first-order optimization algorithm that solves the reformulation efficiently. We report several numerical results in Section 3.5, and conclude the paper in Section 3.6. All omitted proofs can be found in Appendix.

3.2 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 $\text{supp}(\mathbb{P})$ the support of $\mathbb{P}$, and write $\mathbb{P} \ll \mu$ if $\mathbb{P}$ is absolutely continuous with respect to μ. 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 $\text{Proj}_{1\#}\gamma$ and $\text{Proj}_{2\#}\gamma$ the first and the second marginal distributions of γ, respectively. For a given set A, define the characteristic function $1_A(x)$ such that $1_A(x) = 1$ when $x \in A$ and otherwise $1_A(x) = 0$, and define the indicator function $\tau_A(x)$ such that $\tau_A(x) = 0$ when $x \in A$ and otherwise $\tau_A(x) = \infty$. Define the distance between two sets

A and B in the Euclidean space as $\text{Dist}(A, B) = \sup_{x\in A} \inf_{y\in B} \|x - y\|_2$. Define the sign function $\text{sign}(\cdot)$ such that $\text{sign}(x) = 1$ if $x > 0$ and otherwise $\text{sign}(x) = -1$. For a given function $\omega : \Theta \to \mathbb{R}$, 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 6 (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*

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

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 cost function, and $H(\gamma \mid \mu \otimes \nu)$ denotes the relative entropy of γ with respect to the product measure $\mu \otimes \nu$:

$$H(\gamma \mid \mu \otimes \nu) = \int \log\left(\frac{\mathrm{d}\gamma(x, y)}{\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)}\right) \mathrm{d}\gamma(x, y),$$

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).

$\diamondsuit$

Remark 4 (Variants of Sinkhorn Distance)**.** *Sinkhorn distance in Definition 6 is based on general reference measures μ and ν. Special forms of the distance has been investigated in literature, for instance, when the reference measures μ and ν were chosen to be $\mathbb{P}, \mathbb{Q}$, i.e., marginal distributions of γ, respectively [66, Section 2]. The relative entropy regularization term can also be considered as a hard-constrained variant for the optimal transport problem,*

which has been discussed in [46, Definition 1] and [9]:

$$\mathcal{W}_R^{\mathsf{Info}}(\mathbb{P},\mathbb{Q}) \;=\; \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{\mathbb{E}_{(X,Y)\sim\gamma}[c(X,Y)] :\; H(\gamma \mid \mathbb{P}\otimes\mathbb{Q}) \le R\right\},$$

where $R \ge 0$ quantifies the upper bound for the relative entropy between distributions γ and $\mathbb{P}\otimes\mathbb{Q}$. Another variant of the optimal transport problem is to consider the negative entropy for regularization [46, Equation (2)]:

$$\mathcal{W}_\epsilon^{\mathsf{Ent}}(\mathbb{P},\mathbb{Q}) \;=\; \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \;\left\{\mathbb{E}_{(X,Y)\sim\gamma}[c(X,Y)] + \epsilon H(\gamma)\right\},$$

where $H(\gamma) = \int \log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}x\,\mathrm{d}y}\right)\mathrm{d}\gamma(x,y)$ and $\mathrm{d}x$, $\mathrm{d}y$ are Lebesgue measures if the corresponding marginal distributions are continuous, or counting measures if the marginal distributions are discrete. For given $\mathbb{P}$ and $\mathbb{Q}$, one can check the two regularized optimal transport distances above are equivalent up to a constant:

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

♣

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

$$\begin{aligned} V := &\sup_{\mathbb{P}\in\mathbb{B}_{\rho,\epsilon}(\widehat{P})} \mathbb{E}_{z\sim\mathbb{P}}[f(z)],\\ &\text{where } \mathbb{B}_{\rho,\epsilon}(\widehat{P}) = \left\{\mathbb{P} :\; \mathcal{W}_\epsilon(\widehat{P},\mathbb{P}) \le \rho\right\},\end{aligned} \qquad \texttt{(Sinkhorn Primal)}$$

where $\mathbb{B}_{\rho,\epsilon}(\widehat{P})$ is the Sinkhorn ball of the radius ρ centered at the nominal distribution $\widehat{P}$. Due to the convex entropic regularizer [44] $\mathcal{W}_\epsilon(\widehat{P},\mathbb{P})$ with respect to $\mathbb{P}$, the Sinkhorn

distance $\mathcal{W}_\epsilon(\widehat{P}, \mathbb{P})$ is convex in $\mathbb{P}$, i.e., when $\mathbb{P}_1$ and $\mathbb{P}_2$ are two probability distributions, it holds that

$$\mathcal{W}_\epsilon(\widehat{P}, \lambda\mathbb{P}_1 + (1-\lambda)\mathbb{P}_2) \leq \lambda\mathcal{W}_\epsilon(\widehat{P}, \mathbb{P}_1) + (1-\lambda)\mathcal{W}_\epsilon(\widehat{P}, \mathbb{P}_2)$$

for all $0 \leq \lambda \leq 1$. Therefore, the Sinkhorn ball is a convex set, and the problem (`Sinkhorn Primal`) is an (infinite-dimensional) convex program. We refer to Remark 6 on the discussion of the existence of worst-case distribution.

Remark 5 (Choice of Reference Measures)**.** *We discuss below the choices of the two references measures μ and ν in Definition 6.*

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 $\mathcal{W}_\epsilon(\widehat{P}, \mathbb{P})$ can be written as

$$\begin{aligned} H(\gamma \mid \mu \otimes \nu) &= \int \log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\widehat{P}(x)\,\mathrm{d}\nu(y)}\right) + \log\left(\frac{\widehat{P}(x)}{\mathrm{d}\mu(x)}\right) \mathrm{d}\gamma(x,y) \\ &= \int \log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\widehat{P}(x)\,\mathrm{d}\nu(y)}\right) \mathrm{d}\gamma(x,y) + \int \log\left(\frac{\widehat{P}(x)}{\mathrm{d}\mu(x)}\right) \mathrm{d}\widehat{P}(x). \end{aligned}$$

Therefore, any choice of the reference measure μ satisfying $\widehat{P} \ll \mu$ is equivalent up to a constant. For simplicity, in the sequel we will take $\mu = \widehat{P}$.

For the reference measure ν, observe that the worst-case solution $\mathbb{P}$ in the problem (`Sinkhorn Primal`) *should satisfy that $\mathbb{P} \ll \nu$ since otherwise the entropic regularization in Definition 6 is undefined. As a consequence, we can choose ν such that the underlying true distribution is absolutely continuous with respect to it. Typical choices include the Lebesgue measure or Gaussian measure for continuous random variables, and counting measure for discrete measures. See [123, Section 3.6] for the construction of a general reference measure.* ♣

In the following sections, we first derive the tractable formulation of the Sinkhorn DRO model and then develop an efficient first-order method to solve it. Finally, we examine its

performance by several numerical examples.

3.3 Strong Duality Reformulation

Problem (Sinkhorn 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 (Sinkhorn Primal). Our main goal is to derive the strong dual problem

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

where the dual decision variable λ corresponds to the Sinkhorn ball constraint in problem (Sinkhorn 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; and we define the constant

$$\overline{\rho} := \rho + \epsilon \int \log \left(\int e^{-c(x,z)/\epsilon} \, \mathrm{d}\nu(z) \right) \mathrm{d}\widehat{P}(x), \tag{3.1}$$

and the kernel probability distribution

$$\mathrm{d}\mathbb{Q}_{x,\epsilon}(z) := \frac{e^{-c(x,z)/\epsilon}}{\int e^{-c(x,u)/\epsilon} \, \mathrm{d}\nu(u)} \, \mathrm{d}\nu(z). \tag{3.2}$$

The rest of this section is organized as follows. In Section 3.3.1, we summarize our main results on the strong duality reformulation of Sinkhorn DRO. Next, we provide detailed discussions in Section 3.3.2. In Section 3.3.3, we provide a proof sketch of our main results.

3.3.1 Main Theorem

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

Assumption 2. (I) $\nu\{z:\ 0 \le c(x,z) < \infty\} = 1$ *for* $\widehat{P}$*-almost every* x*;*

(II) $\int e^{-c(x,z)/\epsilon}\,\mathrm{d}\nu(z) < \infty$ *for* $\widehat{P}$*-almost every* x*;*

(III) $\mathcal{Z}$ *is a measurable space, and the function* $f:\ \mathcal{Z} \to \mathbb{R}\cup\{\infty\}$ *is measurable.*

By reference [120, Proposition 4.1], the Sinkhorn distance has the equivalent formulation:

$$\mathcal{W}_\epsilon(\widehat{P},\mathbb{P}) = \min_{\gamma\in\Gamma(\widehat{P},\mathbb{P})} \int \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\mathcal{K}}(x,y)\right)\,\mathrm{d}\gamma(x,y),$$

where

$$\mathrm{d}\mathcal{K}(x,y) = e^{-c(x,y)/\epsilon}\,\mathrm{d}\widehat{P}(x)\,\mathrm{d}\nu(y).$$

Assumption 2(I) implies that $0 \le c(x,y) < \infty$ for $\widehat{P}\otimes\nu$-almost every (x,y), and therefore the reference measure $\mathcal{K}$ is well-defined. Assumption 2(II) ensures the optimal transport mapping γ^* for Sinkhorn distance $\mathcal{W}_\epsilon(\widehat{P},\mathbb{P})$ exists with density value $\frac{\mathrm{d}\gamma^*(x,y)}{\mathrm{d}\widehat{P}(x)\,\mathrm{d}\nu(y)} \propto e^{-c(x,y)/\epsilon}$. Hence, Assumption 2(I) and 2(II) together ensure the Sinkhorn distance is well-defined.

Assumption 2(III) ensures the expected loss $\mathbb{E}_{z\sim\mathbb{P}}[f(z)]$ to be well-defined and lower bounded for any distribution $\mathbb{P}$.

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

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

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

Theorem 5 (Strong Duality)**.** *Let* $\widehat{P}\in\mathcal{P}(\mathcal{Z})$*, and assume Assumption 2 holds. Then the following holds:*

(I) *The primal problem* (`Sinkhorn Primal`) *is feasible if and only if* $\overline{\rho}\ge 0$*;*

(II) *Whenever* $\overline{\rho}\ge 0$*, it holds that* $V = V_D$*.*

(III) *If, in addition, Condition 1 holds, 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 5 in Section 3.3.3 below. Therefore, we have $V = V_D$ as long as Assumption 2 holds. Along the proof we also obtain the dual reformulation on the soft distributionally robust formulation of (`Sinkhorn Primal`).

Corollary 1. *Let* $\widehat{P} \in \mathcal{P}(\mathcal{Z})$ *and* $\lambda > 0$*, and assume Assumption 2 holds. Then the primal problem*

$$V_\lambda = \sup_{\mathbb{P}} \ \left\{ \mathbb{E}_{z\sim\mathbb{P}}[f(z)] - \lambda \mathcal{W}_\epsilon(\widehat{P}, \mathbb{P}) \right\}. \tag{SDRO(λ)}$$

has the equivalent dual reformulation:

$$V_\lambda^{\texttt{Dual}} = \lambda\epsilon \int \log\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}[e^{f(z)/(\lambda\epsilon)}]\right) \mathrm{d}\widehat{P}(x) + C, \tag{SDRO(λ)-Dual}$$

where the constant

$$C = \lambda\epsilon \int \log\left(\int e^{-c(x,u)/\epsilon} \mathrm{d}\nu(u)\right) \mathrm{d}\widehat{P}(x).$$

3.3.2 Discussions

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

Remark 6 (Worst-case Distribution)**.** *Assume the optimal Lagrangian multiplier in* (`Dual`) $\lambda^* > 0$*. As we will demonstrate in the proof of Theorem 5, the worst-case distribution for* (`Sinkhorn Primal`) *maps every* $x \in \operatorname{supp} \widehat{P}$ *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 $\alpha_x := \left[\int \exp\Big(\big(f(z) - \lambda^* c(x,z)\big)/(\lambda^*\epsilon)\Big) \mathrm{d}\nu(z)\right]^{-1}$ *is a normalizing constant to ensure the conditional distribution well-defined. As such, the density of the worst-case*

distribution can be expressed as

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

from which we can see that the worst-case distribution shares the same support as the measure ν. For the case where $\lambda^* = 0$, the worst-case distribution will ensure the corresponding objective function equal the essential supremum of the loss function f. Particularly, when $\widehat{P}$ 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 [61]. This is another difference, or advantage possibly, of Sinkhorn DRO compared with Wasserstein DRO. Indeed, for many practical problems, the underlying distribution can be 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 3.5 demonstrate some empirical advantages of Sinkhorn DRO. ♣

Remark 7 (Connection with Wasserstein DRO)**.** *As the regularization parameter $\epsilon \to 0$, the dual objective of the Sinkhorn DRO converges to*

$$\lambda\rho + \int \operatorname*{ess\,sup}_{\nu}\big\{f(\cdot) - \lambda c(x,\cdot)\big\}\,\mathrm{d}\widehat{P}(x).$$

The proof is provided in Appendix B.4, which essentially follows from the fact that the log-sum-exp function is a smooth approximation of the supremum. Particularly, when $\mathrm{supp}(\nu) = \mathcal{Z}$, the dual objective of the Sinkhorn DRO converges to the dual formulation of the Wasserstein DRO problem [61, Theorem 1]. There are several advantages of Sinkhorn DRO.

(I) *As we will demonstrate in Section 3.4, Sinkhorn DRO is tractable for a large class*

of loss functions. For the empirical nominal distribution, the worst-case loss can be evaluated efficiently for any measurable loss function f. In contrast, the main computational difficulty in Wasserstein DRO is to solve the maximization problem inside the integration above. In fact, 1-Wasserstein DRO is shown to be tractable only when the loss function can be expressed as a pointwise maximum of finitely many concave functions [110, Theorem 4.2], and 2-Wasserstein DRO is shown to be tractable only when the loss function is smooth and the radius of the ambiguity set is sufficiently small [25, Theorem 3].

(II) *The strong duality of Sinkhorn DRO holds in an even more general setting. Essentially, the only requirements on the space $\mathcal{Z}$ and the nominal distribution $\widehat{P}$ are measurability. In contrast, the strong duality for Wasserstein DRO ([61, Theorem 1], [22, Theorem 1]) requires the nominal distribution $\widehat{P}$ to be a Borel probability measure and the set $\mathcal{Z}$ to be a Polish space.*

We 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 based on [61, Theorem 1 and Proposition 2], the Wasserstein DRO is finite if and only if the loss function satisfies a growth condition $f(z) \leq L_f c(z, z_0) + M, \forall z \in \mathcal{Z}$ for some constants $L_f, M > 0$ and some $z_0 \in \mathcal{Z}$. ♣

Remark 8 (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*

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

which corresponds to the dual objective function [84] for the following KL-divergence DRO

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

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

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

where $K_\epsilon(x) \propto \exp(-\|x\|_2^2/\epsilon)$ represents the Gaussian kernel. By Lemma 3 and divergence inequality [44, 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 $\mathbb{P}^0$:

$$V = \mathbb{E}_{\mathbb{P}^0}[f(z)] = \int \mathbb{E}_{\mathbb{Q}_{x,\epsilon}}[f(z)]\,\mathrm{d}\widehat{P}(x). \tag{3.3}$$

In non-parameteric statistics, it has been shown in [76, Theorem 4.2.1] that the optimal bandwidth to minimize the mean-squared-error between the estimated distribution $\mathbb{P}_0$ and the underlying true one is at rate $\epsilon = O(n^{-1/(d+4)})$. 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 3.5, we select ϵ based on cross-validation. ♣

Remark 9 (Connection with Bayesian DRO)**.** *Recently, the Bayesian DRO [144] framework proposed to solve*

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

where $\widehat{P}$ 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.,

$$\mathcal{P}_x := \{\mathbb{P} :\ D_{KL}(\mathbb{P}\|\mathbb{Q}_x) \le \eta\},$$

with $\mathbb{Q}_x$ being the parametric distribution conditioned on x. Based on [144, Section 2.1.3], a relaxation of the Bayesian DRO dual formulation given by

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

When specifying the parametric distribution $\mathbb{Q}_x$ as kernel probability distribution in (3.2) *and applying the change-of-variable technique such that λ is replaced with $\lambda\epsilon$, this relaxed formulation becomes*

$$\inf_{\lambda\geq 0}\ \left\{\lambda(\eta\epsilon)+\lambda\epsilon\int\log\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\right)\,\mathrm{d}\widehat{P}(x)\right\}.$$

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

Let us illustrate our result for a linear or quadratic loss function f, which turns out to be equivalent to a simple optimization problem.

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 cost function 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

$$V_{\mathrm{D}}=\inf_{\lambda>0}\ \left\{\lambda\overline{\rho}+\lambda\epsilon\int\Lambda_x(\lambda)\,\mathrm{d}\widehat{P}(x)\right\},$$

where

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

Therefore

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

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 performing linear regression with quadratic loss $f(z) = (a^{\mathrm{T}}\theta - b)^2$, 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 cost function as $c((a, b), (a', b')) = \|a - a'\|_2^2 + \infty|b - b'|$. In this case, the dual problem becomes*

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

In comparison with the corresponding Wasserstein DRO formulation with radius ρ (see, e.g., [23, Example 4])

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

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

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

Corollary 2 (Conic Reformulation for Finite Support)**.** *Suppose that the support contains L elements, i.e., $\mathcal{Z} = \{z_\ell\}_{\ell=1}^L$, and the nominal distribution $\widehat{P} = \frac{1}{n}\sum_{i=1}^n \delta_{\hat{x}_i}$. If Condition 1 holds and $\overline{\rho} \geq 0$, the dual problem* (`Dual`) *can be formulated as the following conic*

optimization:

$$\begin{aligned} V_{\mathrm{D}} = \min_{\substack{\lambda \geq 0, s \in \mathbb{R}^n, \\ a \in \mathbb{R}^{n \times L}}} \quad & \lambda \overline{\rho} + \frac{1}{n} \sum_{i=1}^{n} s_i \\ \textit{s.t.} \quad & \lambda \epsilon \geq \sum_{\ell=1}^{L} q_{i,\ell} a_{i,\ell}, i \in [n], \\ & (\lambda \epsilon, a_{i,\ell}, f(z_\ell) - s_i) \in \mathcal{K}_{\exp}, i \in [n], \ell \in [L]. \end{aligned} \tag{3.4}$$

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* (3.2)*, 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 (3.4) is a convex program that minimizes a linear function with respect to linear and conic constraints, which can be solved using interior point algorithms [115, 154]. We will develop an efficient first-order optimization algorithm in Section 3.4 that is able to solve a more general problem (without a finite support).

3.3.3 Proof of Theorem 5

In this subsection, we present a sketch of the proof for Theorem 5. We first show the feasibility result in Theorem 5(I). The key is based on the observation that the primal problem (`Sinkhorn Primal`) can be reformulated as a generalized KL-divergence DRO problem.

Lemma 3 (Reformulation of (`Sinkhorn Primal`))**.** *Under Assumption 2, it holds that*

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

Due to Lemma 3, Theorem 5(I) holds based on the non-negativity of KL-divergence.

Next, we develop the duality result for the primal problem V. We begin with the weak duality result in Lemma 4, which can be shown by the application of Lagrangian weak duality theorem.

Lemma 4 (Weak Duality)**.** *Assume Assumption 2 holds. Then* $V \leq V_{\mathrm{D}}$.

When Condition 1 holds, we prove the strong duality by constructing the worst-case distribution. We first show the existence of the dual minimizer (Lemma 5), and then build the corresponding first-order optimality condition (Lemma 6 and Lemma 7). Those results help us to construct a primal optimal solution for (`Sinkhorn Primal`) that shares the same optimal value as V_{D}, which completes the first part of Theorem 5(III). When Condition 1 does not hold, we construct a sequence of DRO problems with finite optimal values converging into V and consequently $V = V_{\mathrm{D}} = \infty$, which completes the second part of Theorem 5(III). Putting these two parts together imply Theorem 5(II).

Below we provide the proof of the first part of Theorem 5(III) for the case $\overline{\rho} > 0$ under Condition 1, and defer proofs of other degenerate cases to Appendix B.5. To prove the strong duality, we will construct a feasible solution of (`Sinkhorn Primal`) whose loss coincides with V_{D}. To this end, we first show that the dual minimizer exists.

Lemma 5 (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.*

We separate two cases: $\lambda^* > 0$ and $\lambda^* = 0$, corresponding to whether the Sinkhorn distance constraint in (`Sinkhorn Primal`) is binding or not.

Lemma 6 below presents a necessary and sufficient condition for the dual minimizer $\lambda^* = 0$, corresponding to the case where the Sinkhorn distance constraint in (`Sinkhorn Primal`) is not binding.

Lemma 6 (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) $\overline{\rho}' = \overline{\rho} + \epsilon \int \log\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}[1_A]\right) \mathrm{d}\widehat{P}(x) \geq 0$, *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 $\overline{\rho}$.

Lemma 7 below considers the optimality condition when the dual minimizer $\lambda^* > 0$, obtained by simply setting the derivative of the dual objective function to be zero.

Lemma 7 (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*

$$\lambda^* \left[\overline{\rho} + \epsilon \int \log \left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}} \left[e^{f(z)/(\lambda^* \epsilon)} \right] \right) \, \mathrm{d}\widehat{P}(x) \right] = \int \frac{\mathbb{E}_{\mathbb{Q}_{x,\epsilon}} \left[e^{f(z)/(\lambda^* \epsilon)} f(z) \right]}{\mathbb{E}_{\mathbb{Q}_{x,\epsilon}} \left[e^{f(z)/(\lambda^* \epsilon)} \right]} \, \mathrm{d}\widehat{P}(x). \quad (3.5)$$

Now we are ready to prove Theorem 5.

Proof. Proof of Theorem 5(III) under Condition 1 with $\overline{\rho} > 0$. The proof is separated for two cases: $\lambda^* > 0$ or $\lambda^* = 0$. For each case we prove by constructing a primal (approximate) optimal solution.

When $\lambda^* > 0$, we take the transport mapping γ_* such that

$$\frac{\mathrm{d}\gamma_*(x,z)}{\mathrm{d}\widehat{P}(x)\,\mathrm{d}\nu(z)} = \alpha_x \exp\left(\frac{1}{\lambda^* \epsilon} \phi(\lambda^*; x, z) \right), \quad \text{where } \phi(\lambda; x, z) \; = f(z) - \lambda c(x,z),$$

and $\alpha_x := \left[\int \exp\left(\frac{1}{\lambda^* \epsilon} \phi(\lambda^*; x, z) \right) \, \mathrm{d}\nu(z) \right]^{-1}$ is a normalizing constant such that

$$\mathrm{Proj}_{1\#} \gamma_* = \widehat{P}.$$

Also define the primal (approximate) optimal distribution $\mathbb{P}_* := \mathrm{Proj}_{2\#} \gamma_*$. Recall the

expression of the Sinkhorn distance in Definition 6, one can verify that

$$
\begin{aligned}
&\mathcal{W}_\epsilon(\widehat{P}, \mathbb{P}_*) \\
&= \inf_{\gamma \in \Gamma(\widehat{P}, \mathbb{P}_*)} \left\{ \mathbb{E}_\gamma \left[c(x,z) + \epsilon \log \left(\frac{\mathrm{d}\gamma(x,z)}{\mathrm{d}\widehat{P}(x)\,\mathrm{d}\nu(z)} \right) \right] \right\} \\
&\leq \mathbb{E}_{\gamma_*} \left[c(x,z) + \epsilon \log \left(\frac{\mathrm{d}\gamma_*(x,z)}{\mathrm{d}\widehat{P}(x)\,\mathrm{d}\nu(z)} \right) \right] \\
&= \mathbb{E}_{\gamma_*} \left[c(x,z) + \epsilon \log \left(\frac{\exp\left(\frac{\phi(\lambda^*;x,z)}{\lambda^*\epsilon}\right)}{\int \exp\left(\frac{\phi(\lambda^*;x,u)}{\lambda^*\epsilon}\right)\,\mathrm{d}\nu(u)} \right) \right] \\
&= \frac{1}{\lambda^*} \Bigg\{ \iint \frac{f(z) \exp\left(\frac{\phi(\lambda^*;x,z)}{\lambda^*\epsilon}\right)}{\int \exp\left(\frac{\phi(\lambda^*;x,z)}{\lambda^*\epsilon}\right)\,\mathrm{d}\nu(z)}\,\mathrm{d}\nu(z)\,\mathrm{d}\widehat{P}(x) \\
&\qquad\qquad - \lambda^*\epsilon \int \log \left(\int \exp\left(\frac{\phi(\lambda^*;x,u)}{\lambda^*\epsilon} \right)\,\mathrm{d}\nu(u) \right)\,\mathrm{d}\widehat{P}(x) \Bigg\},
\end{aligned}
$$

where the second relation is because γ_* is a feasible solution in $\Gamma(\widehat{P}, \mathbb{P}_*)$, the third and the fourth relation is by substituting the expression of γ_*. Since $\overline{\rho} > 0$ and the dual minimizer $\lambda^* > 0$, the optimality condition in (3.5) holds, which implies that $\mathcal{W}_\epsilon(\widehat{P}, \mathbb{P}_*) \leq \rho$, i.e., the distribution $\mathbb{P}_*$ is primal feasible for the problem (`Sinkhorn Primal`). Moreover, we can see that the primal optimal value is lower bounded by the dual optimal value:

$$
\begin{aligned}
V \geq \mathbb{E}_{\mathbb{P}_*}[f(z)] &= \int f(z)\,\mathrm{d}\gamma_*(x,z) \\
&= \iint f(z) \left(\frac{\mathrm{d}\gamma_*(x,z)}{\mathrm{d}\widehat{P}(x)\,\mathrm{d}\nu(z)} \right)\,\mathrm{d}\nu(z)\,\mathrm{d}\widehat{P}(x) \\
&= \iint f(z) \frac{\exp\left(\frac{\phi(\lambda^*;x,z)}{\lambda^*\epsilon}\right)}{\int \exp\left(\frac{\phi(\lambda^*;x,u)}{\lambda^*\epsilon}\right)\,\mathrm{d}\nu(u)}\,\mathrm{d}\nu(z)\,\mathrm{d}\widehat{P}(x) \\
&= \lambda^* \left[\rho + \epsilon \int \log \left(\int \exp\left[\frac{\phi(\lambda^*;x,z)}{\lambda^*\epsilon} \right]\,\mathrm{d}\nu(z) \right)\,\mathrm{d}\widehat{P}(x) \right] \\
&= V_{\mathrm{D}},
\end{aligned}
$$

where the third equality is based on the optimality condition in Lemma 7. This, together with the weak duality result, completes the proof for $\lambda^* > 0$.

When $\lambda^* = 0$, the optimality condition in Lemma 6 holds. We construct the primal (approximate) solution $\mathbb{P}_* = \mathrm{Proj}_{2\#}\gamma_*$, where γ_* satisfies

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

We can verify easily that the primal solution is feasible based on the optimality condition $\overline{\rho}' \geq 0$ in Lemma 6. Moreover, we can check that the primal optimal value is lower bounded by the dual optimal value:

$$\begin{aligned} V \geq & \int f(z)\,\mathrm{d}\gamma_*(x,z) \\ = & \iint f(z)\,\mathrm{d}\gamma_*^x(z)\,\mathrm{d}\widehat{P}(x) \\ = & \iint \operatorname*{ess\,sup}_{\nu} f\,\mathrm{d}\gamma_*^x(z)\,\mathrm{d}\widehat{P}(x) \\ = & \operatorname*{ess\,sup}_{\nu} f = V_{\mathrm{D}}, \end{aligned}$$

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 for $\lambda^* = 0$. □

3.4 Efficient First-order Algorithm for Sinkhorn Robust Optimization

Consider the Sinkhorn robust optimization problem

$$\inf_{\theta\in\Theta} \sup_{\mathbb{P}\in\mathbb{B}_{\rho,\epsilon}(\widehat{P})} \mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)]. \tag{3.6}$$

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

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

Require: Interval $[\lambda_\ell, \lambda_u]$, maximum outer iterations T_{out}, inexact objective oracle of (D) (denoted as $\widehat{F}^*(\cdot)$, constructed from Algorithm 4).

$\lambda^{(0)} \leftarrow \lambda_\ell$, $y_\ell^{(0)} \leftarrow \lambda_\ell, y_u^{(0)} \leftarrow \lambda_u$.
for $t = 1, \ldots, T_{\text{out}}$ **do**
  Update $z_\ell^{(t)} \leftarrow \frac{1}{3}\big[2y_\ell^{(t-1)} + y_u^{(t-1)}\big]$ and $z_u^{(t)} \leftarrow \frac{1}{3}\big[y_\ell^{(t-1)} + 2y_u^{(t-1)}\big]$.
  if $\widehat{F}(z_\ell^{(t)}) \le \widehat{F}(z_u^{(t)})$ **then**
    Update $(y_\ell^{(t)}, y_u^{(t)}) \leftarrow (y_\ell^{(t-1)}, z_u^{(t)})$.
    If $\widehat{F}(z_\ell^{(t)}) \le \widehat{F}(\lambda^{(t-1)})$, update $\lambda^{(t)} \leftarrow z_\ell^{(t)}$.
  else if $\widehat{F}(z_\ell^{(t)}) > \widehat{F}(z_u^{(t)})$ **then**
    Update $(y_\ell^{(t)}, y_u^{(t)}) \leftarrow (z_\ell^{(t)}, y_u^{(t-1)})$.
    If $\widehat{F}(z_u^{(t)}) \le \widehat{F}(\lambda^{(t-1)})$, update $\lambda^{(t)} \leftarrow z_u^{(t)}$.
  end if
end for
Output $\lambda^{(\text{Last})}$.

distribution $\widehat{P}$. Based on our strong dual expression (Dual), we reformulate (3.6) as

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

where the constant $\overline{\rho}$ and the distribution $\mathbb{Q}_{x,\epsilon}$ are defined in (3.1) and (3.2), 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, in this section, we present a stochastic mirror descent algorithm with bisection search to solve this problem. Throughout this section, we assume the loss function $f_\theta(z)$ is convex in θ.

3.4.1 Main Algorithm

We present several notations before outlining the main algorithm. Define the objective value of (D) as

$$F^*(\lambda) := \lambda\overline{\rho} + \inf_{\theta\in\Theta} \left\{ F(\theta;\lambda) := \mathbb{E}_{x\sim\widehat{P}}\left[\lambda\epsilon \log\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}\left[e^{f_\theta(z)/(\lambda\epsilon)}\right]\right)\right]\right\}. \tag{3.7}$$

Algorithm 4 BSMD with Sampling for finding the inexact objective oracle of (D)

Require: Batch size m, maximum inner iterations T_{in}, constant step size γ, initial guess θ_1, fixed multiplier λ.

1: **for** $i = 1, \ldots, m$ **do**
2: **for** $t = 0, 1, \ldots, T_{\text{in}} - 1$ **do** {Step 2-6: BSMD Step}
3: Formulate (biased) gradient estimate of $F(\theta_t; \lambda)$, denoted as $v(\theta_t; \lambda)$.
4: Update $\theta_{t+1} = \text{Prox}_{\theta_t}\big(\gamma v(\theta_t; \lambda)\big)$.
5: **end for**
6: Obtain estimate of optimal solution $\widehat{\theta}_i = \frac{1}{T_{\text{in}}} \sum_{t=1}^{T_{\text{in}}} \theta_t$.
7: Formulate objective estimate of $F(\widehat{\theta}_i; \lambda)$, denoted as $V(\widehat{\theta}_i; \lambda)$. {Sampling Step}
8: **end for**

Output the estimator $\min_{i \in [m]} V(\widehat{\theta}_i; \lambda)$.

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 $\|\cdot\|$, where the norm function $\|\cdot\|$ satisfies that its dual norm $\|\cdot\|_* \leq c\|\cdot\|_2$. This induces the Bregman divergence $D_\omega(\theta, \theta') : \Theta \times \Theta \to \mathbb{R}_+$:

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

Next, we define the *prox-mapping* $\text{Prox} : \mathbb{R}^{d_\theta} \to \Theta$ as

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

With these notations in hand, we present our algorithm, which consists of outer and inner iterations. At the outer iterations, we apply the bisection search algorithm to seek a near-optimal multiplier in (D) provided that an inexact objective oracle of (D) is given. The inner iterations find such an oracle based on the following steps:

BSMD Step: First, we propose a **S**tochastic **M**irror **D**escent algorithm with **B**iased gradient estimators (BSMD) to obtain a near-optimal decision of (3.7) for a given multiplier λ.

Sampling Step: Next, we present a sampling-based algorithm to estimate the objective value of (3.7) for a given decision θ and multiplier λ.

Combining these two steps gives us an inexact objective oracle of (D). We summarize the

algorithm at outer and inner iterations in Algorithm 3 and 4, respectively. In the following, we elaborate how to formulate gradient estimators in BSMD step and how to formulate objective estimators in sampling step.

An alternative approach for solving the Sinkhorn DRO formulation (D) is to update (λ, θ) jointly using stochastic gradient methods. However, as pointed out in [112], for λ of a small value, the variance of the gradient estimate of the objective function with respect to λ is unstable. Hence, we develop a bisection method to update λ at outer iterations in Algorithm 3.

Configuration of Gradient Simulation in BSMD Step

Observe that the objective function of (3.7) involves a nonlinear transformation of the expectation, thus an unbiased gradient estimate could be challenging to obtain when $\mathbb{Q}_{x,\epsilon}$ is a general probability distribution. Based on a batch of simulated samples from $\mathbb{Q}_{x,\epsilon}$, we provide stochastic gradient estimators that are possibly biased. As we will elaborate in Section 3.4.2, by properly tuning the hyper-parameters of these estimators to balance their bias and variance trade-off, the BSMD step can efficiently find a near-optimal decision of (3.7). Since the multiplier λ is fixed in (3.7), we omit the dependence of λ when defining objective or gradient terms in the remaining of this section.

Remark 10 (Sampling from $\mathbb{Q}_{x,\epsilon}$)**.** *In many cases, generating samples from $\mathbb{Q}_{x,\epsilon}$ is easy. When the cost function $c(\cdot,\cdot) = \frac{1}{2}\|\cdot - \cdot\|_2^2$ 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 cost function $c(\cdot,\cdot)$ is decomposable in each coordinate, we can apply the acceptance-rejection method [6] to generate samples in each coordinate independently, the complexity of which only increases linearly in the data dimension. When the cost function $c(x,y) = \frac{1}{q}\|x-y\|_p^q$, the complexity of sampling based on Lagenvin Monte Carlo method for obtaining a τ-close sample point is of $O(d/\tau)$. See the detailed algorithm of sampling in Appendix B.7.3.* ♣

We follow the idea of multi-level Monte-Carlo (MLMC) simulation in [82] to to tackle

the difficulty of simulating gradient estimate. First, we construct a sequence of approximation functions $\{F^\ell(\theta)\}_{\ell\geq 0}$ instead, where

$$F^\ell(\theta) = \mathbb{E}_{x^\ell}\, \mathbb{E}_{\{z_j^\ell\}_{j\in[2^\ell]}|x^\ell}\left[\lambda\epsilon \log\left(\frac{1}{2^\ell}\sum_{j\in[2^\ell]} \exp\left(\frac{f_\theta(z_j^\ell)}{\lambda\epsilon}\right)\right)\right]. \tag{3.8}$$

Here the random variable x^ℓ follows distribution $\widehat{P}$, and for fixed value 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

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

For fixed parameter θ, we define

$$g^\ell(\theta,\zeta^\ell) = \nabla_\theta U_{1:2^\ell}(\theta,\zeta^\ell), \tag{3.9}$$

$$G^\ell(\theta,\zeta^\ell) = \nabla_\theta\left[U_{1:2^\ell}(\theta,\zeta^\ell) - \frac{1}{2}U_{1:2^{\ell-1}}(\theta,\zeta^\ell) - \frac{1}{2}U_{2^{\ell-1}+1:2^\ell}(\theta,\zeta^\ell)\right]. \tag{3.10}$$

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 $\nabla F^\ell(\theta) - \nabla F^{\ell-1}(\theta)$. Since $\nabla F^\ell(\theta)$ and $\nabla F^{\ell-1}(\theta)$ are close to each other for large ℓ, stochastic estimators of them using the same random sampling parameters ζ^ℓ will be highly correlated. Consequently, the gradient estimator $G^\ell(\theta,\zeta^\ell)$ will have small variance for large ℓ, making it a suitable recipe for stochastic optimization. We list the following choices of MLMC-based gradient estimator in Step 3 of Algorithm 4 using $g^\ell(\theta,\zeta^\ell)$ and $G^\ell(\theta,\zeta^\ell)$:

Vanilla Stochastic Gradient Descent (V-SGD) Estimator: at point θ, query oracle for

n_L° times to obtain $\{g^L(\theta, \zeta_i^L)\}_{i=1}^{n_L^{\circ}}$ and construct

$$v^{\text{V-SGD}}(\theta) = \frac{1}{n_L^{\circ}} \sum_{i=1}^{n_L^{\circ}} g^L(\theta, \zeta_i^L). \tag{3.11a}$$

Vanilla MLMC Estimator (V-MLMC): at point θ, for each ℓ we query oracle for $n_\ell := \lceil 2^{-\ell} N \rceil$ times to obtain $\{G^\ell(\theta, \zeta_i^\ell)\}_{i=1}^{n_\ell}$ and construct

$$v^{\text{V-MLMC}}(\theta) = \sum_{\ell=0}^{L} \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} G^\ell(\theta, \zeta_i^\ell). \tag{3.11b}$$

Randomized Truncation MLMC (RT-MLMC) Estimator: at point θ, we sample *random levels* for n_L° times, denoted as $\iota_1, \ldots, \iota_{n_L^{\circ}}$, following distribution $Q_{\text{RT}} = \{q_\ell\}_{\ell=0}^{L}$ with $\mathbb{P}(\iota = \ell) = q_\ell$, where the probability mass value $q_\ell \propto 2^{-\ell}$. Then construct

$$v^{\text{RT-MLMC}}(\theta) = \frac{1}{n_L^{\circ}} \sum_{i=1}^{n_L^{\circ}} \frac{1}{q_{\iota_i}} G^{\iota_i}(\theta, \zeta^{\iota_i}). \tag{3.11c}$$

In the convergence analysis part in Section 3.4.2, we demonstrate V-MLMC and RT-MLMC estimators are more computationally efficient than V-SGD estimator provided that the loss function $f_\theta(z)$ is smooth in θ. Once gradient recipes $G^{\iota_i}(\theta, \zeta^{\iota_i})$ are fixed, the V-MLMC estimator is a deterministic way for gradient simulation while the RT-MLMC estimator is a randomized approach.

Configuration of Objective Simulation in Sampling Step

Similar to the gradient simulation part, we list MLMC-based sampling methods for estimating the objective value in (3.7) for fixed θ. For notation simplicity, define

$$A^\ell(\theta, \zeta^\ell) = U_{1:2^\ell}(\theta, \zeta^\ell) - \frac{1}{2} U_{1:2^{\ell-1}}(\theta, \zeta^\ell) - \frac{1}{2} U_{2^{\ell-1}+1:2^\ell}(\theta, \zeta^\ell). \tag{3.12}$$

V-MLMC Estimator: at point θ, for each ℓ we query oracle for $n_\ell := \lceil 2^{-\ell} N \rceil$ times to obtain $\{A^\ell(\theta, \zeta_i^\ell)\}_{i=1}^{n_\ell}$ and construct

$$V^{\text{V-MLMC}}(\theta) = \sum_{\ell=0}^{L} \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} A^\ell(\theta, \zeta_i^\ell). \tag{3.13a}$$

RT-MLMC Estimator: at point θ, we sample *random levels* for n_L^{o} times, denoted as $\iota_1, \ldots, \iota_{n_L^{\text{o}}}$, following distribution $Q_{\text{RT}} = \{q_\ell\}_{\ell=0}^{L}$ with $\mathbb{P}(\iota = \ell) = q_\ell$, where the probability mass value $q_\ell \propto 2^{-\ell}$. Then construct

$$V^{\text{RT-MLMC}}(\theta) = \frac{1}{n_L^{\text{o}}} \sum_{i=1}^{n_L^{\text{o}}} \frac{1}{q_{\iota_i}} A^{\iota_i}(\theta, \zeta^{\iota_i}). \tag{3.13b}$$

3.4.2 Convergence Properties

With our algorithm described, we now study its convergence properties. We begin with the following assumptions regarding the loss function f_θ:

Assumption 3. (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 $|f_{\theta_1}(z) - f_{\theta_2}(z)| \le L_f \|\theta_1 - \theta_2\|_2$.*

(IV) *(Lipschitz Smoothness): The loss function $f_\theta(z)$ is continuously differentiable and for fixed z and θ_1, θ_2, it holds that $\|\nabla f_{\theta_1}(z) - \nabla f_{\theta_2}(z)\|_2 \le S_f \|\theta_1 - \theta_2\|_2$.*

From Section 3.4.1 and 3.4.1, we can see the computational and storage bottleneck of our algorithm is on the random sampling parameter $\zeta^\ell = (x^\ell, \{z_j^\ell\}_{j \in [2^\ell]})$. To this end, we quantify the computation cost of our algorithm as *the (expected) number of queries for generating the random sampling parameter*, and the memory cost as *the stored (expected)*

number of random sampling parameter in data buffer .

BSMD Step at Inner Iterations

In this part, we discuss the BSMD step (see Step 2-6 in Algorithm 4) for estimating an optimal solution of (3.7). By Corollary 1, the formulation (3.7) corresponds to the Sinkhorn robust learning problem with a softened Sinkhorn ball constraint. To quantify the quality of the obtained solution, we say a given random vector θ is a δ-*optimal solution* if $\mathbb{E}[F(\theta)] - F(\theta^*) \leq \delta$, where θ^* is the optimal solution of (3.7).

The BSMD step iteratively obtains a stochastic gradient estimate (not necessarily unbiased) of the objective function and then performs a proximal gradient update. By properly tuning hyper-parameters to balance the trade-off between bias and variance of gradient estimate, we establish performance guarantees for our proposed BSMD step in Theorem 6. A detailed proof and formal statement can be found in Appendix B.7.

Theorem 6 (Complexity Analysis of BSMD Step)**.** *Under Assumption 3*(I)*, 3*(II)*, and 3*(III)*, with properly chosen hyper-parameters of gradient estimators in* (3.11)*, the following results hold:*

(I) *When the loss function $f_\theta(z)$ is nonsmooth in θ, the computation cost of V-SGD scheme for finding δ-optimal solution is of $O(\delta^{-3})$, with memory cost $O(1/\delta)$.*

(II) *Additionally assume Assumption 3*(IV) *holds, then the computation cost of V-SGD scheme for finding δ-optimal solution is of $O(\delta^{-3})$, with memory cost $O(1/\delta)$; the computation cost of V-MLMC scheme is of $O(\delta^{-2})$, with memory cost $\tilde{O}(1/\delta)$; and the computation cost of RT-MLMC scheme is of $O(\delta^{-2})$, with memory cost $\tilde{O}(1)$.*

The configuration of optimization hyper-parameters is provided in Table 3.2.

Theorem 6 indicates that the computation cost of BSDM algorithm for solving (3.7) is of $O(\delta^{-3})$. When the additional smoothness assumption of loss function holds, the computation

Table 3.2: Configuration of optimization hyper-parameters together with the computational/memory cost for obtaining δ-optimal solution of (3.7) in Theorem 6. Here "Comp." and "Memo." are the abbreviations of "Computation" and "Memory", respectively.

Smooth?	V-SGD		V-MLMC		RT-MLMC	
	Parameter	Comp./Memo.	Parameter	Comp./Memo.	Parameter	Comp./Memo.
No	$L = O(\log \frac{1}{\delta})$ $T_{\text{in}} = O(1/\delta^2)$ $n_L^{\text{o}} = O(1)$ $\gamma = O(\delta)$	Comp. $= O(T n_L^{\text{o}} 2^L)$ $= O(1/\delta^3)$ Memo. $= O(n_L^{\text{o}} 2^L)$ $= O(1/\delta)$	N/A	N/A	N/A	N/A
Yes	$L = O(\log \frac{1}{\delta})$ $T_{\text{in}} = O(1/\delta^2)$ $n_L^{\text{o}} = O(1)$ $\gamma = O(\delta)$	Comp. $= O(T n_L^{\text{o}} 2^L)$ $= O(1/\delta^3)$ Memo. $= O(n_L^{\text{o}} 2^L)$ $= O(1/\delta)$	$L = O(\log \frac{1}{\delta})$ $T_{\text{in}} = O(1/\delta)$ $N = O(1/\delta)$ $\gamma = O(1)$	Comp. $= O(T(NL + 2^L))$ $= \tilde{O}(1/\delta^2)$ Memo. $= O(NL + 2^L)$ $= \tilde{O}(1/\delta)$	$L = O(\log \frac{1}{\delta})$ $T_{\text{in}} = \tilde{O}(1/\delta^2)$ $n_L^{\text{o}} = O(1)$ $\gamma = O(\delta)$	Comp. $= O(T(n_L^{\text{o}} L))$ $= \tilde{O}(1/\delta^2)$ Memo. $= O(n_L^{\text{o}} L)$ $= \tilde{O}(1)$

cost further reduces to $\tilde{O}(\delta^{-2})$ using V-MLMC or RT-MLMC gradient estimators. This complexity is near-optimal for solving general convex and smooth optimization problems [17]. It is an open question whether the BSMD step with V-SGD gradient estimator is optimal for solving (3.7) with convex and nonsmooth loss functions. In comparison with V-MLMC scheme, the RT-MLMC scheme has the same order of computation cost but achieves cheaper memory cost $\tilde{O}(1)$, which is nearly error tolerance-independent.

Moreover, we note that the problem (3.7) can be viewed as a special case of the conditional stochastic optimization (CSO) studied in [81, 82, 83]. As far as we have known, the stochastic approximation-based idea proposed in [82] is the most efficient algorithm to solve the generic CSO problem. Although we follow the similar idea to design BSMD step, the difference is that we consider a more practical constrained optimization scenario, while Hu et al. [82] focused on unconstrained optimization only.

Remark 11 (Comparison with Biased Sample Average Approximation Approach)**.** *Another way for solving the formulation* (3.7) *is to formulate sample estimates of the inner expectation and then optimize the biased sample estimate of the objective function instead, called the biased sample average approximation (BSAA) technique. Applying [81, Corollary 4.2], it can be shown that the total computation cost and memory cost for BSAA formulation of achieving δ-optimal solution are both of $\tilde{O}(\delta^{-3})$ for Lipschitz continuous loss functions (See*

Table 3.3: Configuration of optimization hyper-parameters together with the computational/memory cost for estimating optimal value in (3.7) in Theorem 6. Here the "cost in Step 7" refers to both the computation and memory cost when implementing V-MLMC or RT-MLMC sampling method.

V-MLMC		**RT-MLMC**	
Parameter	Cost in Step 7	Parameter	Cost in Step 7
$L = O(\log \frac{1}{\delta})$	$O(NL + 2^L)$	$L = O(\log \frac{1}{\delta})$	$O(n_L^{\circ} L)$
$N = \tilde{O}\left(\frac{1}{\delta^2} \log \frac{1}{\alpha}\right)$	$=\tilde{O}(\delta^{-2})$	$n_L^{\circ} = \tilde{O}\left(\frac{1}{\delta^2} \log \frac{1}{\alpha}\right)$	$=\tilde{O}(\delta^{-2})$

formal statement in Appendix B.7.4). The storage complexity for BSAA is always worse compared with the proposed BSMD approach which is at most in the order of δ^{-1}*. The computation complexity for BSAA formulation is worse compared with the proposed BSMD approach when the loss functions are also Lipschitz smooth. Also, the BSAA method still requires a solution of the approximated optimization problem. Hence, it typically takes considerably less time and memory to run the BSMD step rather than solving for the BSAA formulation.* ♣

Sampling Step at Inner Iterations

After a near-optimal solution of (3.7) is obtained, we estimate its objective value in sampling step. With properly chosen hyper-parameters, the output of Algorithm 4 gives an estimation of the optimal value in (3.7) with negligible error with high probability. The complexity analysis of Algorithm 4 is provided in Theorem 7.

Theorem 7 (Complexity for Estimating Optimal Value in (3.7))**.** *Fix an error probability* $\eta \in (0, 1)$ *and specify* $m = \lceil \log_2 \frac{2}{\eta} \rceil$*. Assume that Assumption 3*(I)*, 3*(II)*, and 3*(III) *hold, then with properly chosen hyper-parameters, the output in Algorithm 4 satisfies*

$$Pr\left\{ \left| \min_{i \in [m]} V(\widehat{\theta}_i) - F(\theta^*) \right| \leq 3\delta \right\} \geq 1 - \eta.$$

In addition,

(I) *The computation cost of Algorithm 4 with V-SGD scheme for BSMD step and V-MLMC (or RT-MLMC) objective estimator is of* $O\left(\delta^{-3} \cdot \textit{polylog}\frac{1}{\eta}\right)$*, with memory cost* $\tilde{O}\left(\delta^{-2} \cdot \textit{polylog}\frac{1}{\eta}\right)$.

(II) *Additionally assume Assumption 3*(IV) *holds, then the computation cost of Algorithm 4 with V-MLMC (or RT-MLMC) scheme for BSMD step and V-MLMC (or RT-MLMC) objective estimator is of* $\tilde{O}\left(\delta^{-2} \cdot \textit{polylog}\frac{1}{\eta}\right)$*, with memory cost* $\tilde{O}\left(\delta^{-2} \cdot \textit{polylog}\frac{1}{\eta}\right)$.

Optimization hyper-parameters in BSMD step in Algorithm 4 follow the discussion in Section 3.4.2. The configuration of optimization hyper-parameters in sampling step is provided in Table 3.3.

Hu et al. [81, Theorem 4.1] proposed and analyzed the V-SGD estimator for estimating the objective value of generic CSO problem. Specially, the complexity of this estimator for estimating the optimal value in (3.7) with accuracy error δ is of $O(\delta^{-3})$, while our proposed V-MLMC and RT-MLMC estimators have improved sample complexity $\tilde{O}(\delta^{-2})$.

Bisection Search at Outer Iterations

In Algorithm 3, we propose a bisection search algorithm for solving (D) by iteratively querying the oracle for estimating the optimal value of (3.7). It can be shown that under mild assumptions, we can solve the constrained DRO formulation (D) to accuracy δ by computing $O(\delta)$-accurate optimal values of (3.7) for $O(\log \frac{1}{\delta})$ times.

Theorem 8 (Complexity of Bisection Search)**.** *Fix an error probability* $\eta \in (0, 1)$*. Assume that Assumption 3*(I) *and 3*(II) *hold, and one can pick* λ_ℓ, λ_u *such that the Lagrangian multiplier* λ^* *in* (D) *satisfies* $0 < \lambda_\ell \leq \lambda^* \leq \lambda_u < \infty$*. Specify hyper-parameters in Algorithm 3 as*

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

Suppose there exists an oracle $\widehat{F}$ such that for any $\lambda > 0$, it gives estimation of the optimal value in (3.7) *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 by calling the inexact oracle $\widehat{F}$ for $O(\log\frac{1}{\delta})$ times.*

In particular, Algorithm 4 presents a way for constructing the oracle required by Algorithm 3. Combining Theorem 7 and 8, the overall computation cost for obtaining a δ-optimal solution of (D) with probability at least $1-\eta$ is of $\tilde{O}\left(\delta^{-3}\cdot \text{polylog}\frac{1}{\eta}\right)$, with memory cost $\tilde{O}\left(\delta^{-2}\cdot \text{polylog}\frac{1}{\eta}\right)$. Additionally, when the smoothness condition Assumption 3(IV) holds, the computation cost reduces to $\tilde{O}\left(\delta^{-2}\cdot \text{polylog}\frac{1}{\eta}\right)$.

Remark 12 (Comparison with Wasserestein DRO)**.** *In comparison with our proposed algorithm for Sinkhorn DRO, one should note that Wasserstein DRO is not always tractable. Especially, the Wasserstein robust optimization problem with nominal distribution $\widehat{P} = \frac{1}{n}\sum_{i=1}^{n}\delta_{\hat{x}_i}$ corresponding to* (D) *can be formulated as the minimax problem*

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

However, when $f_\theta(z)$ is not concave in z, the above problem generally reduces to the convex-non-concave minimax learning problem and it is difficult to obtain the global optimum. See Table 3.1 for detailed summary on the tractability of Wasserstein DRO formulation under some special cases. Even when the Wasserstein DRO formulation is tractable, its complexity usually has non-negligible dependence on the sample size n (see [61, Remark 9] and references therein for more discussions). In contrast, the complexity for solving the Sinkhorn DRO formulation is sample size independent. ♣

Also, in many scenarios one need to tune the Sinkhorn radius $\overline{\rho}$ for problem (D) to achieve satisfactory out-of-sample performance. It will often make sense to directly tune the Lagrangian multiplier λ in (3.7) rather than a target radius $\overline{\rho}$. In other words, it is more computationally efficient to solve the subproblem (3.7) with a tuned Lagrangian multiplier

λ directly instead of problem (D) with a tuned Sinkhorn radius $\overline{\rho}$.

3.5 Applications

In this section, we apply our methodology on three applications: the newsvendor model, mean-risk portfolio optimization, and multi-class classification. We examine the performance of the (`Sinkhorn Primal`) model by comparing it with four benchmarks: (i) the classical sample average approximation (SAA) model; (ii) the Wasserstein DRO model; and (iii) the KL-divergence DRO model. We choose the cost function $c(\cdot,\cdot) = \|\cdot - \cdot\|_1^1$ for 1-Wasserstein or 1-Sinkhorn DRO model, and $c(\cdot,\cdot) = \frac{1}{2}\|\cdot - \cdot\|^2$ for 2-Wasserstein or 2-Sinkhorn DRO model. Unless otherwise specified, we take the reference measure ν for the Sinkhorn distance is chosen to be the Lebesgue measure. For each of the three applications, with n training samples, we select hyper-parameters of DRO models using the K-fold cross-validation method with $K = 5$. We run the repeated experiments for 200 independent trials.

In Section 3.5.1 and Section 3.5.2, we measure the out-of-sample performance of a solution θ based on training dataset $\mathcal{D}$ using the coefficient of prescriptiveness in [13]:

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

where J^* denotes the true optimal value when the true distribution is known exactly, $\theta_{\mathcal{D}}^{\text{SAA}}$ denotes the decision from 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 out-of-sample performance the solution has. Further details and additional experiments are included in Appendix B.2 and B.3, respectively.

3.5.1 Newsvendor Model

We consider the following distributionally robust newsvendor model:

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

where the random variable z stands for the random demand; its empirical distribution $\widehat{P}$ consists of n independent samples from the data distribution $\mathbb{P}_*$; 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 size $n \in \{10, 30, 100\}$ 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 this model shares the same formulation as the SAA approach [110, Remark 6.7]. Since 2-Wasserstein DRO is computationally intractable in this example, we solve the corresponding formulation by approximating the support of distribution using discrete grid points.

Table 3.4: 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	0.118	0.131	0.185	0.119	0.133	0.187	0.122	0.132	0.181
2-WDRO	0.123	0.358	1.307	0.128	0.354	1.337	0.134	0.402	1.428
2-SDRO	0.070	0.077	0.120	0.117	0.132	0.177	0.108	0.127	0.178

We report the box plots for the percentage of improvement across different DRO ap-

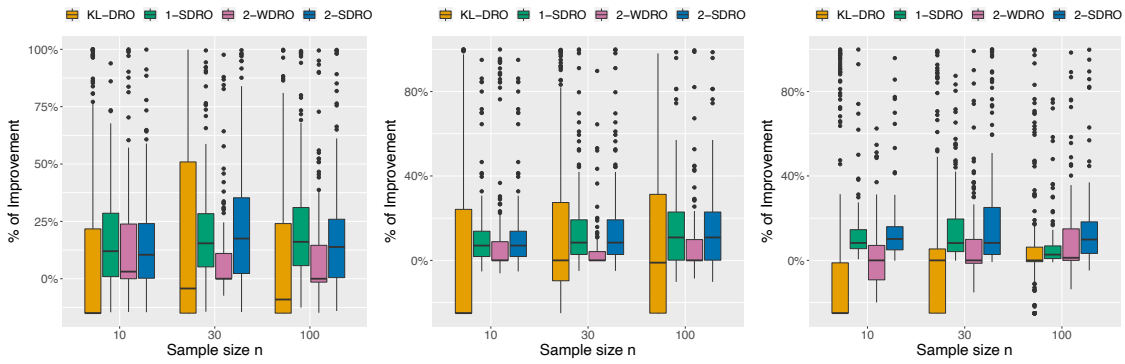

Figure 3.1: Out-of-sample performances for the newsvendor model with parameters $s \in \{0.25, 0.5, 0.75, 1, 2, 4\}$ and the fixed sample size $n = 20$. For figures from left to right, we specify the data distribution as exponential distribution, gamma distribution, and equiprobable mixture of two truncated normal distributions, respectively.

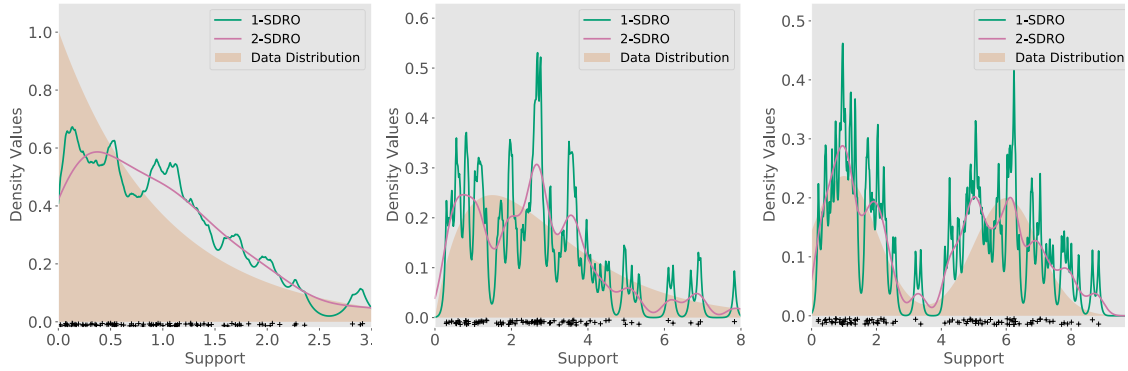

Figure 3.2: Plots for the density of worst-case distributions generated by the 1-SDRO or 2-SDRO model. In all figures we fix the sample size $n = 100$. For figures from left to right, we specify the data distribution as exponential distribution, gamma distribution, and equiprobable mixture of two truncated normal distributions, respectively.

proaches in Fig. 3.1. We find that Sinkhorn DRO has the best out-of-sample performance in all figures. We report the computational time for various approaches in Table 3.4. 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. Finally, we plot the density of worst-case distributions for 1-SDRO or 2-SDRO model in Fig. 3.2. When specifying the data distribution as exponential, gamma, or Gaussian mixture, one can check the corresponding worst-case distributions capture the shape of the ground truth distribution reasonably well. Since the worst-case distributions from Sinkhorn DRO models are more reasonable, the corresponding decisions are less

conservative compared to those from Wasserstein DRO models.

3.5.2 Mean-risk Portfolio Optimization

We consider the following distributionally robust mean-risk portfolio optimization problem

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

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 $\mathbb{P}\text{-CVaR}_{\alpha}(-\theta^{\mathrm{T}} z)$ quantifies conditional value-at-risk [133], i.e., the average of the $\alpha \times 100\%$ worst portfolio losses under the distribution $\mathbb{P}$. We follow a similar setup as in [110]. Specifically, we set $\alpha = 0.2, \varrho = 10$. The random asset $z \sim \mathbb{P}_*$ 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 $\epsilon_i \sim \mathcal{N}(i \times 0.03, i \times 0.025)$. We solve this problem by taking the Bregman divergence D_ω as the KL-divergence, so that the proximal gradient update can be implemented efficiently. Fig. 3.3a) reports the scenario where the data dimension $d = 30$ is fixed and sample size $n \in \{30, 50, 100, 150, 200, 400\}$, and Fig. 3.3b) reports the scenario where the sample size $n = 100$ is fixed and the number of assets $d \in \{5, 10, 20, 40, 80, 100\}$. Those box plots are generated from 200 independent trials, from which we can see that for all problem instances, the 1-SDRO or 2-SDRO model outperforms other DRO baselines.

Table 3.5 reports the average computational time for various DRO models. We observe that when data dimension is fixed and sample size varies from 30 to 400, the computational time for all approaches does not differ too much. When the data dimension increases and sample size is fixed, the computational time for 1-SDRO or 2-SDRO model increases linearly while other DRO models increases mildly. One possible explaination is that in this example,

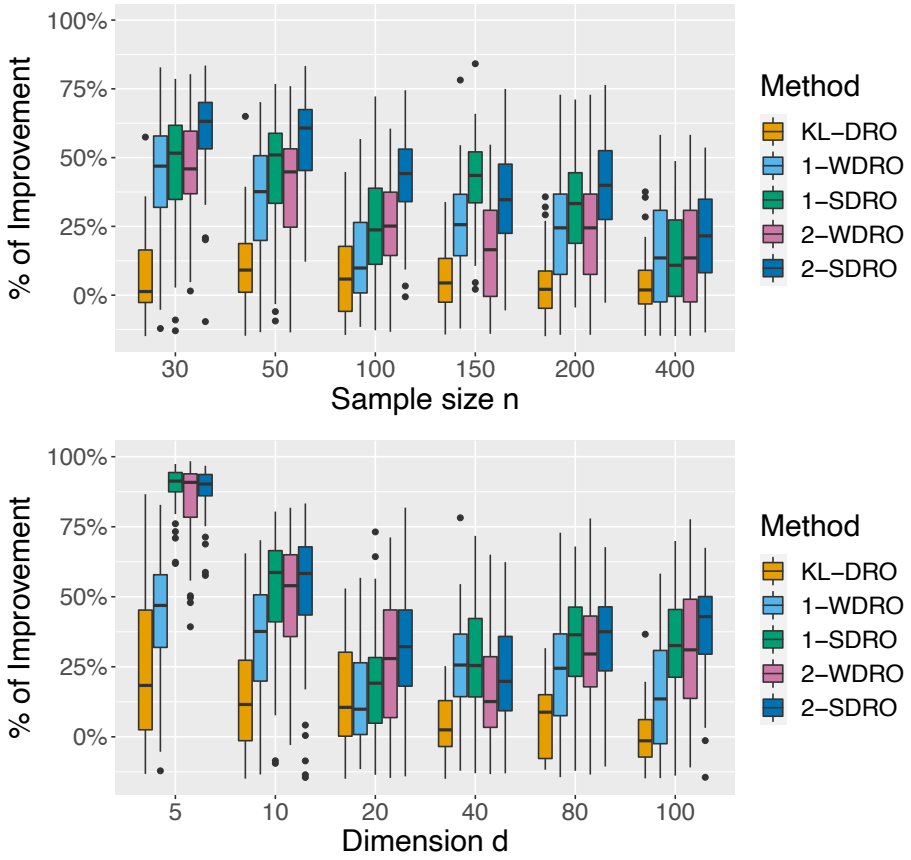

Figure 3.3: Box plot for the the portfolio optimization problem, where we try 200 independent trials for each problem instance. The x-axis indicates the number of observed samples n or the data dimension d, and the y-axis indicates the percentage of improvement in comparison with the SAA baseline. Upper: $d = 30$ and $n \in \{30, 50, 100, 150, 200, 400\}$. Bottom: $n = 100$ and $d \in \{5, 10, 20, 40, 80, 100\}$.

other DRO models have tractable finite-dimensional conic programming formulations so that off-the-shelf softwares can solve them efficiently. In contrast, Sinkhorn DRO models do not have special reformulation, but they can still be solved in reasonable amount of time.

3.5.3 Linear Classification Incorporating Structural Information

Finally, we investigate the multi-class logistic classification to illustrate the computational benefits of Sinkhorn DRO. Given a feature vector $x \in \mathbb{R}^d$ and its label $y \in [C]$, we denote $\vec{y} \in \{0, 1\}^C$ as the corresponding one-hot label vector, and define the following negative

Table 3.5: 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	0.161	0.015	0.148
$(50, 30)$	0.014	0.042	0.020	0.209	0.016	0.175
$(100, 30)$	0.017	0.065	0.024	0.204	0.021	0.261
$(150, 30)$	0.019	0.084	0.029	0.203	0.027	0.258
$(200, 30)$	0.023	0.115	0.035	0.201	0.032	0.263
$(400, 30)$	0.045	0.136	0.061	0.205	0.056	0.257
$(100, 5)$	0.014	0.043	0.017	0.108	0.015	0.175
$(100, 10)$	0.014	0.045	0.018	0.133	0.015	0.229
$(100, 20)$	0.014	0.048	0.021	0.169	0.017	0.295
$(100, 40)$	0.017	0.068	0.027	0.233	0.022	0.432
$(100, 80)$	0.021	0.103	0.052	0.420	0.044	0.758
$(100, 100)$	0.023	0.116	0.070	0.500	0.059	0.836

likelihood loss:

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

where $B := [w_1, \ldots, w_K]$ stands for the linear classifier. Then the DRO model aims to solve the following formulation:

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

Assume only the feature vector x has uncertainty but not the label y. Also, assume the feature vector x lie in a subset of Euclidean space Ξ so that its infinity norm is bounded by 1. We specify the measure ν as Lebesgue measure supported on Ξ.

This experiment is conducted using 6 real datasets from LIBSVM website. For the data preprocessing step, we learn the linear embedding of feature vectors using neural networks such that feature vectors have bounded infinity norm and the mis-classification risk for SAA approach is controlled within 50%. Detailed statistics on pre-processed datasets can be found in Table B.1 in Appendix B.2. We use the testing error for new observations to quantify the performance for the obtained classifers.

Classification results for these different approaches are reported in Table 3.6, where

the first number of each entry represents the average classification error, and the second number of entry represents the half-length of the 95% confidence interval. We can see that in all problem instances SDRO models outperform the corresponding WDRO models, and either 1-SDRO or 2-SDRO model has the best classification performance. The average computational time for various approaches is reported in Table 3.7, from which we can see that Sinkhorn DRO models has shorter computational time than Wasserstein DRO models. In this example, Wasserstein DRO models can be reformulated as convex non-concave minimax problems. We try gradient descent ascent heuristics (see Algorithm 5 in Appendix B.2) to approximately solve those formulations, but they are not computationally efficient to obtain satisfactory classifiers. For small boxes highlighted in Table 3.6, we find the Wasserstein DRO models in some cases even have worse performance than the traditional SAA model.

Table 3.6: Classification results on real datasets. Each experiment is repeated for 200 independent trials, and 95% confidence intervals of classification errors for worse-case subgroup are reported for different approaches.

Dataset	SAA	KL-DRO	1-WDRO	1-SDRO	2-WDRO	2-SDRO
MNIST	$.075 \pm .002$	$.067 \pm .002$	$.037 \pm .003$	$\mathbf{.035 \pm .002}$	$.047 \pm .003$	$\mathbf{.041 \pm .002}$
IRIS	$.396 \pm .024$	$.351 \pm .015$	$.321 \pm .021$	$\mathbf{.308 \pm .021}$	$.378 \pm .023$	$\mathbf{.342 \pm .022}$
wine	$.089 \pm .010$	$.086 \pm .010$	$.082 \pm .005$	$\mathbf{.077 \pm .005}$	$.078 \pm .005$	$\mathbf{.076 \pm .005}$
vowel	$.481 \pm .012$	$.478 \pm .011$	$\boxed{.492 \pm .012}$	$\mathbf{.456 \pm .011}$	$.476 \pm .012$	$\mathbf{.443 \pm .012}$
vehicle	$.379 \pm .007$	$.368 \pm .007$	$\boxed{.481 \pm .014}$	$\mathbf{.343 \pm .006}$	$\boxed{.434 \pm .009}$	$\mathbf{.349 \pm .007}$
svmguide4	$.427 \pm .009$	$.418 \pm .009$	$\boxed{.430 \pm .010}$	$\mathbf{.417 \pm .009}$	$.425 \pm .009$	$\mathbf{.393 \pm .011}$

Table 3.7: Average computational time (in seconds) per problem instance for multi-class classification problem

Dataset	SAA	KL-DRO	1-WDRO	1-SDRO	2-WDRO	2-SDRO
MNIST	2.423	5.245	5.826	4.810	4.920	2.606
IRIS	0.986	1.612	1.803	1.243	1.226	1.033
wine	1.669	2.331	2.104	1.813	1.905	1.826
vowel	19.507	23.438	26.826	25.343	29.447	26.123
vehicle	3.620	7.177	17.789	14.337	18.837	18.146
svmguide4	27.189	29.988	39.635	31.576	37.132	31.361

3.6 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 customized batch gradient descent with bisection search algorithm, we have shown that the resulting DRO problem is tractable under mild assumptions, greatly spans the tractability of Wasserstein DRO. Analysis on the worst-case distribution indicates that Sinkhorn DRO hedges a more reasonable set of adverse scenarios and thus less conservative compared with Wasserstein DRO, which is then demonstrated via extensive numerical experiments. Based on theoretical and numerical findings, we conclude that the Sinkhorn distance is a promising candidate for modeling distributional ambiguities in decision-making under uncertainty from the perspective of computational tractability, modeling rationality and out-of-sample performance.

In the meantime, several topics worth in-depth investigating are left for future works. A meaningful research question is the choice of the optimal hyper-parameters in Sinkhorn DRO, such as the radius of the ambiguity set $\overline{\rho}$, the entropic regularization parameters ϵ, and reference measures ν. This paper focuses on regularizing Wasserstein distance with the entropic regularization – the Sinkhorn distance, but extensions to other types of

regularization are possible. Exploring and discovering the benefits of Sinkhorn DRO in other types of applications may lead to future research directions.

Appendices

APPENDIX A
TWO-SAMPLE TEST WITH KERNEL PROJECTED WASSERSTEIN DISTANCE

A.1 Preliminary Technical Results

Theorem 9 (Pinsker's Inequality [44]). *Consider two discrete probability distributions* $p = \{p_i\}_{i=1}^n$ *and* $q = \{q_i\}_{i=1}^n$*, then it holds that*

$$\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \geq \frac{1}{2}\|p - q\|_1^2.$$

Proposition 2 (Lipschitz Properties of Retraction Operator [26]). *There exists constants* L_1, L_2 *such that the following inequalities hold:*

$$\|\mathit{Retr}_s(\zeta) - s\| \leq L_1\|\zeta\|$$
$$\|\mathit{Retr}_s(\zeta) - (s + \zeta)\| \leq L_2\|\zeta\|^2.$$

Inspired from Appendix A.3 in [87], we are able to compute the constants in Proposition 2 explicitly: $L_1 = 1$ and $L_2 = \frac{1}{2}$. The proof is provided below.

Proof. By definition, we have that

$$\begin{aligned}
\|\mathrm{Retr}_s(\zeta) - s\|_2^2 &= \left\| \frac{s+\zeta}{\|s+\zeta\|} - s \right\|_2^2 \\
&= 2\left(1 - \frac{1}{\|s+\zeta\|_2}\right) \\
&= 2\left(1 - (1 + \sum_i \zeta_i^2)^{-1/2}\right) \\
&\leq \sum_i \zeta_i^2 = \|\zeta\|_2^2.
\end{aligned}$$

where the second and the third equality is by using the relation $s^{\mathrm{T}}\zeta = 0$, and the inequality is based on the relation $2(1 - (1 + z)^{-1/2}) \leq z$ with $z = \sum_i \zeta_i^2$. Then it holds that

$\|\text{Retr}_s(\zeta) - (s+\zeta)\|_2 \leq \|\zeta\|$.

Secondly, we can see that

$$\begin{aligned}\|\text{Retr}_s(\zeta) - (s+\zeta)\|_2^2 &= \left\| \frac{s+\zeta}{\|s+\zeta\|} - (s+\zeta) \right\|_2^2 \\ &= (1 - \|s+\zeta\|_2)^2 \\ &= \left(1 - \sqrt{1 + \sum_i \zeta_i^2}\right)^2 \\ &\leq \frac{1}{4}\|\zeta\|_2^4,\end{aligned}$$

where the inequality is based on the relation that $(1-(1+z)^{1/2})^2 \leq z^2/4$ with $z = \sum_i \zeta_i^2$. Consequently it holds that $\|\text{Retr}_s(\zeta) - (s+\zeta)\|_2 \leq \frac{1}{2}\|\zeta\|^2$. □

Theorem 10 (McDiarmid's Inequality [107])**.** *Let $X_1, \ldots, X_n$ be independent random variables, where X_i has the support $\mathcal{X}_i$. Let $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ be any function with the $(c_1, \ldots, c_n)$ bounded difference property, i.e., for $i \in \{1, \ldots, n\}$ and for any $(x_1, \ldots, x_n), (x_1', \ldots, x_n')$ that differs only in the i-th corodinate, we have*

$$|f(x_1, \ldots, x_n) - f(x_1', \ldots, x_n')| \leq c_i.$$

Then for any $t > 0$, we have

$$Pr\left\{ |f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \geq t \right\} \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Lemma 8 (Equivalent Definition for Sub-Gaussian variables (Lemma 2.3.2 in [70]))**.** *Assume that $\mathbb{E}[\zeta] = 0$ and*

$$\mathbb{P}\{|\zeta| \geq t\} \leq 2C\exp\left(-\frac{t^2}{2\sigma^2}\right), \quad t > 0,$$

for some $C \geq 1$ and $\sigma > 0$. Then the random variable ζ is sub-Gaussian with constant $\tilde{\sigma}^2 = 12(2C+1)\sigma^2$.

Theorem 11 (Poincare's Inequality)**.** *Denote by μ^n the product of μ on $\otimes_{i=1}^n \mathbb{R}^d$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfies the Poincare's inequality, i.e., there exists $M > 0$ for $X \sim \mu$ so that*

Var$[f(X)] \leq M\mathbb{E}[\|\nabla f(X)\|^2]$ *for any* f *satisfying* $\mathbb{E}[f(X)^2] < \infty$ *and* $\mathbb{E}[\|\nabla f(X)\|_2^2] < \infty$. *Consider a function* f *on* $\otimes_{i=1}^n \mathbb{R}^d$ *satisfying* $\mathbb{E}|f(X)| < \infty$ *and* $\sum_{i=1}^n \|\nabla_i f(X)\|^2 \leq \alpha^2$, *and* $\max_{1\leq i\leq n} \|\nabla_i f(X)\| \leq \beta$ *almost surely. Then the following inequality holds for* $X \sim \mu^n$*:*

$$\textit{Pr}\left\{ f(X) - \mathbb{E}[f(X)] > t \right\} \leq \exp\left(-\frac{1}{K} \min(t/\beta, t^2/\alpha^2) \right).$$

A.2 Introduction to Manifold Optimization

A brief introduction to manifold optimization can be found in [80]. In this section we list some related operators for solving manifold optimization problems. Traditional manifold optimization concerns with solving the following problem:

$$\min_{x \in \mathcal{M}} \ f(x), \tag{A.1}$$

where $\mathcal{M}$ is a Riemannian manifold and f is a real-valued function on $\mathcal{M}$. A tangent vector ζ_x to $\mathcal{M}$ at a point x is defined as a mapping so that there exists a curve γ on $\mathcal{M}$ satisfying

$$\gamma(0) = x, \quad \zeta_x[u] = \frac{\mathrm{d}(u(\gamma(t)))}{\mathrm{d}t} \mid_{t=0}, \ \forall u \in \mathfrak{E}(\mathcal{M}),$$

where $\mathfrak{E}(\mathcal{M})$ stands for the collection of real-valued functions defined in a neighborhood of x. Denote by $T_x\mathcal{M}$ as the collection of all tangent vectors to $\mathcal{M}$ at a point x, which is called the tangent space to $\mathcal{M}$ at x. Define $\mathcal{P}_x(z)$ as the projection of z into the tangent space at x. Based on definitions listed above, we can define necessary operators for manifold optimization. The Riemannian gradient of f at x is denoted as $\mathrm{Grad} f(x)$, which can be obtained by projecting the gradient of f at x in the Euclidean space into the tangent space to $\mathcal{M}$ at x:

$$\mathrm{Grad} f(x) = \mathcal{P}_x(\nabla f(x)).$$

Typical Riemannian manifolds include the Sphere and Stiefel manifold defined as follows:

$$\mathrm{Sphere}(n-1) := \{x \in \mathbb{R}^n : \ \|x\|_2 = 1\},$$
$$\mathrm{St}(n,p) := \{X \in \mathbb{R}^{n \times p} : \ X^{\mathrm{T}} X = I_p\}.$$

We can express the tangent space together with the projection operator for these two types of manifolds in analytical form:

$$T_x\text{Sphere}(n-1) = \{z:\ z^{\mathrm{T}}x = 0\}, \quad \mathcal{P}_x(z) = (I - xx^{\mathrm{T}})z$$
$$T_x\text{St}(n,p) = \{Z:\ Z^{\mathrm{T}}X + X^{\mathrm{T}}Z = 0\}, \quad \mathcal{P}_X(Z) = Z - X\frac{X^{\mathrm{T}}Z + Z^{\mathrm{T}}X}{2}.$$

When using first-order methods to solve a manifold optimization problem, one also needs to define the retraction operator associated with $\mathcal{M}$, which is denoted as Retr. It is a smooth mapping from the tangent budle $\cup_{x\in\mathcal{M}} T_x\mathcal{M}$ to $\mathcal{M}$ satisfying that for any $x \in \mathcal{M}$,

- $\text{Retr}_x(0_x) = x$, where 0_x denotes the zero element in $T_x\mathcal{M}$;
- $\lim_{\zeta\in T_x\mathcal{M},\zeta\to 0} \frac{\|\text{Retr}_x(\zeta)-(x+\zeta)\|}{\|\zeta\|} = 0.$

When $\mathcal{M}$ is a sphere, we choose the following retraction operator which can be implemented efficiently:

$$\text{Retr}_x\big(\zeta\big) = \frac{x+\zeta}{\|x+\zeta\|}, \quad x \in \text{Sphere}(n-1).$$

See [54] and [162] for discussions of retraction operators on the Stiefel manifold. The general iteration update of first-order methods for manifold optimization problem can be expressed as

$$x^{t+1} = \text{Retr}_{x^t}(-\tau^t\zeta^t),$$

where τ^t is a well-defined step size and ζ^t is the Riemannian gradient at x^t. The computation of the projected Wasserstein distance relates to the optimization on a Stiefel manifold, while the computation of the KPW distance relates to the optimization on a sphere. A recent paper [26] investigated the Riemannian gradient methods that are guaranteed to converge into stationary points globally, the key proof technique of which relies on Proposition 2. We follow the similar proof idea to establish the convergence analysis for computing the KPW distance.

A.3 Technical Proofs in Section 2.2

Proof of Remark 1. When taking the kernel function $K(x, y) = \langle x, y \rangle$, the space

$$\mathcal{F} = \{a : \ a^{\mathrm{T}} a \leq 1\}.$$

Note that the cost function $c(x, y) = \|x - y\|_2^2$ satisfies $c(mx, my) = m^2 c(x, y)$ for any $m \in \mathbb{R}$. Hence we can argue that the maximizer of the KPW distance is obtained when $a^{\mathrm{T}} a = 1$, i.e.,

$$\mathcal{KPW}(\mu, \nu) = \max_{\substack{f:\, \mathbb{R}^D \to \mathbb{R}, \\ f(z) = a^{\mathrm{T}} z, a^{\mathrm{T}} a = 1}} W\left(f \# \mu, f \# \nu\right).$$

This indicates that the KPW distance reduces into the PW distance. □

Proof of Proposition 1. It is easy to see that $\mu = \nu$ implies $\mathcal{KPW}(\mu, \nu) = 0$. Now we show the converse. For fixed $x \in \mathcal{X}$, $y \in \mathbb{R}^d$ and a distribution μ, define the operator K_μ with the action y as a mapping $K_\mu y : \ \mathcal{X} \to \mathbb{R}^d$ so that

$$K_\mu y(x') = \int (K_x y)(x') \, \mathrm{d}\mu(x) = \int K(x', x) y \, \mathrm{d}\mu(x).$$

When $\mathcal{KPW}(\mu, \nu) = 0$, we can see that

$$f \# \mu = f \# \nu, \quad \forall f \in \mathcal{F},$$

which implies

$$
\begin{aligned}
0 = & \sup_{f:\ \|f\|_{\mathcal{H}}^2 \le 1} \left\| \mathbb{E}_{f\#\mu}[x] - \mathbb{E}_{f\#\nu}[y] \right\|_2 \\
= & \sup_{f:\ \|f\|_{\mathcal{H}}^2 \le 1} \sup_{a:\ \|a\|_2 \le 1} \left(\mathbb{E}_{\mu}[\langle f(x), a \rangle] - \mathbb{E}_{\nu}[\langle f(y), a \rangle] \right) \\
= & \sup_{f:\ \|f\|_{\mathcal{H}}^2 \le 1} \sup_{a:\ \|a\|_2 \le 1} \left(\mathbb{E}_{\mu}[\langle f, K_x a \rangle_{\mathcal{H}}] - \mathbb{E}_{\nu}[\langle f, K_y a \rangle_{\mathcal{H}}] \right) \\
= & \sup_{f:\ \|f\|_{\mathcal{H}}^2 \le 1} \sup_{a:\ \|a\|_2 \le 1} \langle f, (K_\mu - K_\nu) a \rangle \\
= & \sup_{a:\ \|a\|_2 \le 1} \|(K_\mu - K_\nu) a\|_{\mathcal{H}}.
\end{aligned}
$$

Equivalently, $\|(K_\mu - K_\nu)a\|_{\mathcal{H}} = 0$ for any a so that $\|a\|_2 \le 1$. Since $\mathcal{H}$ is a Hilbert space, we imply that $(K_\mu - K_\nu)a$ is a zero function for any a satisfying $\|a\|_2 \le 1$. For any function $f \in \mathcal{C}(X)$, we make the expansion

$$
\begin{aligned}
& \|\mathbb{E}_\mu[f(x)] - \mathbb{E}_\nu[f(y)]\|_2 \\
\le & \|\mathbb{E}_\mu[f(x)] - \mathbb{E}_\mu[g(x)]\|_2 + \|\mathbb{E}_\mu[g(x)] - \mathbb{E}_\nu[g(y)]\|_2 + \|\mathbb{E}_\nu[g(y)] - \mathbb{E}_\nu[f(y)]\|_2 .
\end{aligned}
$$

The first term satisfies that

$$
\|\mathbb{E}_\mu[f(x)] - \mathbb{E}_\mu[g(x)]\|_2 \le \mathbb{E}_\mu[\|f(x) - g(x)\|_2] < \varepsilon,
$$

and the third term can be upper bounded likewise. For the second term, we have that

$$
\begin{aligned}
& \|\mathbb{E}_\mu[g(x)] - \mathbb{E}_\nu[g(y)]\|_2 \\
= & \sup_{a:\ \|a\|_2 \le 1} \left(\mathbb{E}_{\mu}[\langle g(x), a \rangle] - \mathbb{E}_{\nu}[\langle g(y), a \rangle] \right) \\
= & \sup_{a:\ \|a\|_2 \le 1} \left(\mathbb{E}_{\mu}[\langle g, K_x a \rangle] - \mathbb{E}_{\nu}[\langle g, K_y a \rangle] \right) \\
= & \sup_{a:\ \|a\|_2 \le 1} \langle g, (K_\mu - K_\nu) a \rangle = 0,
\end{aligned}
$$

where the last equality is because that $(K_\mu - K_\nu)a$ is a zero function for any a satisfying

$\|a\|_2 \leq 1$. Hence, $\|\mathbb{E}_\mu[f(x)] - \mathbb{E}_\nu[f(y)]\|_2 < 2\varepsilon$ for any $\varepsilon > 0$ and $f \in \mathcal{C}_b(\mathcal{X})$. Then we conclude that the distribution $\mu = \nu$. $\square$

A.4 Technical Proofs in Section 2.3

A.4.1 Deviation of Duality Reformulation

We first present the proof of the dual reformulation of the inner minimization problem in (2.4). By definition, the primal formulation can be expressed as:

$$\min_{\pi \geq 0} \left\{ \sum_{i,j} \pi_{i,j} c_{i,j} - \eta \sum_{i,j} \pi_{i,j} (\log \pi_{i,j} - 1) : \quad \sum_{j} \pi_{i,j} = \frac{1}{n}, \sum_{i} \pi_{i,j} = \frac{1}{m} \right\}. \tag{A.2}$$

The Lagrangian function becomes

$$L(\pi, u, v) = \sum_{i,j} \pi_{i,j} c_{i,j} - \eta \sum_{i,j} \pi_{i,j} (\log \pi_{i,j} - 1) + \sum_{i} u_i \left(\sum_{j} \pi_{i,j} - \frac{1}{n} \right) + \sum_{j} v_j \left(\sum_{i} \pi_{i,j} - \frac{1}{m} \right)$$

Then the dual problem becomes

$$\begin{aligned}
&\max_{u,v} \left\{ \min_{\pi \geq 0} L(\pi, u, v) \right\} \\
=&\max_{u,v} -\frac{1}{n} \sum_{i} u_i - \frac{1}{m} \sum_{j} v_j + \min_{\pi \geq 0} \sum_{i,j} \pi_{i,j} \big[c_{i,j} + u_i + v_j \big] - \eta \pi_{i,j} (\log \pi_{i,j} - 1) \\
=&\max_{u,v} -\frac{1}{n} \sum_{i} u_i - \frac{1}{m} \sum_{j} v_j - \sum_{i,j} \max_{\pi_{i,j} \geq 0} \big\{ -\pi_{i,j} \big[c_{i,j} + u_i + v_j \big] + \eta \pi_{i,j} (\log \pi_{i,j} - 1) \big\} \\
=&\max_{u,v} -\frac{1}{n} \sum_{i} u_i - \frac{1}{m} \sum_{j} v_j - \sum_{i,j} (\eta \phi)^* (u_i + v_j + c_{i,j}) \\
=&\max_{u,v} -\frac{1}{n} \sum_{i} u_i - \frac{1}{m} \sum_{j} v_j - \eta \sum_{i,j} \exp \left(-\frac{u_i + v_j + c_{i,j}}{\eta} \right)
\end{aligned}$$

where $\phi(w) = w \log w - w$ and ϕ^* denotes its conjugate [132]. Moreover, the dual optimal value equals the primal optimal value because the Slater's condition [27] for finite-dimensional optimization is satisfied. Take $u_i' = -u_i/\eta$ and $v_j' = -v_j/\eta$, the dual problem becomes

$$\max_{u',v'} \frac{\eta}{n} \sum_{i} u_i' + \frac{\eta}{m} \sum_{j} v_j' - \eta \sum_{i,j} \exp \left(-\frac{c_{i,j}}{\eta} + u_i' + v_j' \right).$$

Therefore, the whole problem (2.4) becomes

$$\max_{u,v,s} \frac{\eta}{n}\sum_i u_i + \frac{\eta}{m}\sum_j v_j - \eta\sum_{i,j}\exp\left(-\frac{c_{i,j}}{\eta} + u_i + v_j\right).$$

Or equivalently, we write it as the minimization problem:

$$-\eta \times \left\{\min_{u,v,s} \; -\frac{1}{n}\sum_i u_i - \frac{1}{m}\sum_j v_j + \eta\sum_{i,j}\exp\left(-\frac{c_{i,j}}{\eta} + u_i + v_j\right)\right\}.$$

Remark 13. *By adding the entropic regularization term $\eta H(\pi)$, we are able to derive an unconstrained optimization formulation on the sphere, thus reducing the computational cost for computing KPW distance. Besides, the induced optimal transport mapping between projected samples is usually stochastic instead of deterministic, which is robust to potential data outliers.*

A.4.2 Proof of Theorem 1

Assume that $\hat{f}$ is an optimal solution to the problem (2.2). Let S be the subspace

$$S = \left\{\sum_{i=1}^{n}\sum_{j=1}^{m}(K_{x_i} - K_{y_j})a_{i,j} : \; a_{i,j} \in \mathbb{R}^d\right\}.$$

Denote by $S_\perp$ the orthogonal complement of S. Given a set $\mathcal{X}$, denote by $f_{\mathcal{X}}$ a function that lies in the set $\mathcal{X}$. Then by the projection theorem, there exists $\hat{f}_S$ and $\hat{f}_{S_\perp}$ such that $\hat{f} = \hat{f}_S + \hat{f}_{S_\perp}$ and $\|\hat{f}\|_{\mathcal{H}}^2 = \|\hat{f}_S\|_{\mathcal{H}}^2 + \|\hat{f}_{S_\perp}\|_{\mathcal{H}}^2$. It remains to show that $\hat{f}_S$ shares the same

objective value with $\hat{f}$. For fixed i, j, we have that

$$\begin{aligned}
\|\hat{f}(x_i) - \hat{f}(y_j)\|_2 &= \max_{a_{i,j}:\ \|a_{i,j}\|_2 \le 1} \langle \hat{f}(x_i) - \hat{f}(y_j), a_{i,j} \rangle \\
&= \max_{a_{i,j}:\ \|a_{i,j}\|_2 \le 1} \langle \hat{f}(x_i), a_{i,j} \rangle - \langle \hat{f}(y_j), a_{i,j} \rangle \\
&= \max_{a_{i,j}:\ \|a_{i,j}\|_2 \le 1} \langle \hat{f}, K_{x_i} a_{i,j} \rangle - \langle \hat{f}, K_{y_j} a_{i,j} \rangle \\
&= \max_{a_{i,j}:\ \|a_{i,j}\|_2 \le 1} \langle \hat{f}, (K_{x_i} - K_{y_j}) a_{i,j} \rangle \\
&= \max_{a_{i,j}:\ \|a_{i,j}\|_2 \le 1} \langle \hat{f}_S, (K_{x_i} - K_{y_j}) a_{i,j} \rangle = \|\hat{f}_S(x_i) - \hat{f}_S(y_j)\|_2,
\end{aligned}$$

where the second last equality is because $\hat{f}_{S_\perp}$ is orthogonal to the subspace S. It follows that $\|\hat{f}(x_i) - \hat{f}(y_j)\|_2^2 = \|\hat{f}_S(x_i) - \hat{f}_S(y_j)\|_2^2$. Therefore, there always exists an optimal solution that lies in the subspace S, which means that there exists an optimal solution to (2.2) that admits the following expression:

$$\hat{f} = \sum_{i=1}^{n} \sum_{j=1}^{m} (K_{x_i} - K_{y_j}) a_{i,j}.$$

Defining $a_{x,i} = \sum_{j=1}^{m} a_{i,j}$ and $a_{y,j} = \sum_{i=1}^{n} a_{i,j}$ completes the proof.

Remark 14. *From the proof we can also see that the representer theorem holds if replacing the square of the ℓ_2 norm in* (2.2) *with any p-th power of the ℓ_2 norm for $p \geq 2$. However, we find the development of optimization algorithms for the square of the ℓ_2 norm case is the simplest.*

A.4.3 Proof of Theorem 2

In the following we give a iteration complexity analysis about Algorithm 2, the proof of which largely follows the idea in [85]. In particular, we first establish the descent lemma for the update of each block of variables and then argue that the objective function is lower bounded. Based on these two facts, we finally build the iteration complexity result for Algorithm 2.

Lemma 9 (Lipschitzness of $\nabla_s F(u,v,s)$)**.** *Let $\{u^t, v^t, s^t\}_t$ be the sequence generated from Algorithm 2. The following inequality holds for any $s \in \mathbb{S}^{d(n+m)-1}$ and $\lambda \in [0,1]$:*

$$\|\nabla_s F(u^{t+1}, v^{t+1}, \lambda s + (1-\lambda)s^t) - \nabla_s F(u^{t+1}, v^{t+1}, s^t)\| \le \varrho\lambda\|s^t - s\|,$$

where $\varrho = \frac{2\|AU\|_\infty^2}{\eta} + \frac{4\|AU\|_\infty^4}{\eta^2}$ and $\|AU\|_\infty = \max_{i,j} \|A_{i,j}U\|_2$.

Proof of Lemma 9. An intermediate result is that

$$\begin{aligned}
\sum_i \pi_{i,j}(u^{t+1}, v^{t+1}, s^t) &= \sum_i \exp\left(-\frac{1}{\eta}c_{i,j}[s^t] + u_i^{t+1}\right)\exp\left(v_j^{t+1}\right) \\
&= \sum_i \exp\left(-\frac{1}{\eta}c_{i,j}[s^t] + u_i^{t+1}\right)\exp\left(v_j^{t}\right)\frac{1/m}{\sum_i \pi_{i,j}(u^{t+1}, v^t, s^t)} \\
&= \frac{1}{m}\frac{\sum_i \pi_{i,j}(u^{t+1}, v^t, s^t)}{\sum_i \pi_{i,j}(u^{t+1}, v^t, s^t)} = 1/m.
\end{aligned}$$

Then we can assert that $\sum_{i,j} \pi_{i,j}(u^{t+1}, v^t, s^t) = 1$. For fixed s^t, define $s^\lambda = \lambda s + (1-\lambda)s^t$. Then we have that

$$\begin{aligned}
&\|\nabla_s F(u^{t+1}, v^{t+1}, s^t) - \nabla_s F(u^{t+1}, v^{t+1}, s^\lambda)\| \\
=&\frac{2}{\eta}\left\|\sum_{i,j} \pi_{i,j}(u^{t+1}, v^{t+1}, s^t)U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}Us^t - \sum_{i,j} \pi_{i,j}(u^{t+1}, v^{t+1}, s^\lambda)U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}Us^\lambda\right\| \\
\le&\frac{2}{\eta}\left\|\sum_{i,j} \pi_{i,j}(u^{t+1}, v^{t+1}, s^t)U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U(s^t - s^\lambda)\right\| \\
&+ \frac{2}{\eta}\left\|\sum_{i,j} U^{\mathrm{T}}\left[\pi_{i,j}(u^{t+1}, v^{t+1}, s^t) - \pi_{i,j}(u^{t+1}, v^{t+1}, s^\lambda)\right]A_{i,j}^{\mathrm{T}}A_{i,j}U\right\| \\
\le&\frac{2}{\eta}\left\|\sum_{i,j} \pi_{i,j}(u^{t+1}, v^{t+1}, s^t)U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U\right\|\|s^\lambda - s^t\| \\
&+ \frac{2}{\eta}\left\|\sum_{i,j} \left[\pi_{i,j}(u^{t+1}, v^{t+1}, s^t) - \pi_{i,j}(u^{t+1}, v^{t+1}, s^\lambda)\right]U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U\right\|
\end{aligned}$$

where the first inequality is based on the constraint that $\|s^\lambda\| \le \lambda\|s\| + (1-\lambda)\|s^t\| = 1$.

To upper bound the first term, we find

$$\begin{aligned}&\left\|\sum_{i,j}\pi_{i,j}(u^{t+1},v^{t+1},s^t)U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U\right\|\\ \leq&\sum_{i,j}\pi_{i,j}(u^{t+1},v^{t+1},s^t)\|U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U\|_2\leq\max_{i,j}\|A_{i,j}U\|_2^2.\end{aligned}$$

To bound the second term, we find that

$$\begin{aligned}&\left\|\sum_{i,j}\left[\pi_{i,j}(u^{t+1},v^{t+1},s^t)-\pi_{i,j}(u^{t+1},v^{t+1},s^\lambda)\right]U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U\right\|\\ \leq&\max_{i,j}\|A_{i,j}U\|_2^2\|\pi(u^{t+1},v^{t+1},s^\lambda)-\pi(u^{t+1},v^{t+1},s^t)\|_1,\end{aligned}$$

where

$$\|\pi(u^{t+1},v^{t+1},s^\lambda)-\pi(u^{t+1},v^{t+1},s^t)\|_1:=\sum_{i,j}\left|\pi_{i,j}(u^{t+1},v^{t+1},s^\lambda)-\pi_{i,j}(u^{t+1},v^{t+1},s^t)\right|.$$

Denote by $H(\pi,s;\eta)$ the objective function for (2.3). Based on the strong convexity property, we have that

$$\begin{aligned}&\langle\nabla_\pi H(\pi(u^{t+1},v^{t+1},s^\lambda),s^\lambda;\eta)-\nabla_\pi H(\pi(u^{t+1},v^{t+1},s^t),s^\lambda;\eta),\pi(u^{t+1},v^{t+1},s^\lambda)-\pi(u^{t+1},v^{t+1},s^t)\rangle\\ \geq&\eta\|\pi(u^{t+1},v^{t+1},s^\lambda)-\pi(u^{t+1},v^{t+1},s^t)\|_1^2\end{aligned}$$

Moreover, by simple calculation we find

$$\begin{aligned}\nabla_\pi H(\pi(u,v,s),s)&=\left[c_{i,j}+\eta\log(\pi_{i,j}(u,v,s))\right]_{i,j}\\&=\left[\eta(u_i+v_j)\right]_{i,j},\end{aligned}$$

where the second equality is by substituting the formulation of $\pi_{i,j}(u,v,s)$. Hence, we find

that the gradient $\nabla_\pi H(\pi(u,v,s),s)$ only depends on u and v, which implies

$$\begin{aligned}&\langle \nabla_\pi H(\pi(u^{t+1},v^{t+1},s^t),s^t;\eta) - \nabla_\pi H(\pi(u^{t+1},v^{t+1},s^t),s^\lambda;\eta), \pi(u^{t+1},v^{t+1},s^\lambda) - \pi(u^{t+1},v^{t+1},s^t)\rangle \\ \geq & \eta \|\pi(u^{t+1},v^{t+1},s^\lambda) - \pi(u^{t+1},v^{t+1},s^t)\|_1^2.\end{aligned}$$

It follows that

$$\begin{aligned}&\eta\|\pi(u^{t+1},v^{t+1},s^\lambda) - \pi(u^{t+1},v^{t+1},s^t)\|_1 \\ \leq & \|\nabla_\pi H(\pi(u^{t+1},v^{t+1},s^t),s^t;\eta) - \nabla_\pi H(\pi(u^{t+1},v^{t+1},s^t),s^\lambda;\eta)\|_\infty \\ = & \max_{i,j} \left| \|A_{i,j}Us^\lambda\|_2^2 - \|A_{i,j}Us^t\|_2^2 \right| \\ \leq & 2\max_{i,j} \|A_{i,j}U\|_2^2 \|s^\lambda - s^t\|.\end{aligned}$$

where the inequality is by applying the following relation:

$$\begin{aligned}\|Ax_1\|_2^2 - \|Ax_2\|_2^2 &= (x_1-x_2)^{\mathrm{T}}(A^{\mathrm{T}}Ax_1) + x_2^{\mathrm{T}}A^{\mathrm{T}}A(x_1-x_2) \\ &\leq \|x_1-x_2\|\|A^{\mathrm{T}}Ax_1\| + \|x_2^{\mathrm{T}}A^{\mathrm{T}}A\|\|x_1-x_2\| \\ &\leq 2\|A\|^2\|x_1-x_2\|.\end{aligned}$$

In summary, the second term can be upper bounded as

$$\begin{aligned}&\left\| \sum_{i,j} \left[\pi_{i,j}(u^{t+1},v^{t+1},s^t) - \pi_{i,j}(u^{t+1},v^{t+1},s^\lambda)\right] U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U \right\| \\ \leq & \frac{2\left(\max_{i,j}\|A_{i,j}U\|_2^2\right)^2}{\eta}\|s^\lambda - s^t\|.\end{aligned}$$

Then applying the condition that $\|s^\lambda - s^t\| = \lambda\|s - s^t\|$ completes the proof. □

Lemma 10 (Decrease of F in s)**.** *Let $\{u^t, v^t, s^t\}_t$ be the sequence generated from Algorithm 2. The following inequality holds for any $k \geq 1$:*

$$F(u^{t+1},v^{t+1},s^{t+1}) - F(u^{t+1},v^{t+1},s^t) \leq -\frac{1}{8\|AU\|_\infty^2 L_2/\eta + 2\varrho L_1^2}\|\xi^{t+1}\|^2.$$

Proof of Lemma 10. Note that

$$
\begin{aligned}
&\left|F(u^{t+1},v^{t+1},s^{t+1})-F(u^{t+1},v^{t+1},s^{t})-\langle\nabla_t F(u^{t+1},v^{t+1},s^{t}),s^{t+1}-s^{t}\rangle\right|\\
=&\left|\int_0^1\langle\nabla_s F(u^{t+1},v^{t+1},\lambda s^{t+1}+(1-\lambda)s^{t})-\nabla_s F(u^{t+1},v^{t+1},s^{t}),s^{t+1}-s^{t}\rangle\,\mathrm{d}\lambda\right|\\
\le&\int_0^1\|\nabla_s F(u^{t+1},v^{t+1},\lambda s^{t+1}+(1-\lambda)s^{t})-\nabla_s F(u^{t+1},v^{t+1},s^{t})\|\|s^{t+1}-s^{t}\|\,\mathrm{d}\lambda\\
\le&\int_0^1\varrho\lambda\|s^{t+1}-s^{t}\|^2\,\mathrm{d}\lambda\\
=&\frac{\varrho}{2}\|s^{t+1}-s^{t}\|^2=\frac{\varrho}{2}\left\|\mathrm{Retr}_{s^t}\big(-\tau\xi^{t+1}\big)-s^{t}\right\|^2\\
\le&\frac{\varrho\tau^2L_1^2}{2}\|\xi^{t+1}\|^2.
\end{aligned}
$$

where the second inequality is by applying Lemma 9, and the last inequality is by applying Proposition 2. Moreover, we have that

$$
\begin{aligned}
&\langle\nabla_s F(u^{t+1},v^{t+1},s^{t}),s^{t+1}-s^{t}\rangle\\
=&\langle\nabla_s F(u^{t+1},v^{t+1},s^{t}),-\tau\xi^{t+1}\rangle+\langle\nabla_s F(u^{t+1},v^{t+1},s^{t}),\mathrm{Retr}_{s^t}\big(-\tau\xi^{t+1}\big)-(s^{t}-\tau\xi^{t+1})\rangle\\
\le&-\tau\|\xi^{t+1}\|^2+\|\nabla_s F(u^{t+1},v^{t+1},s^{t})\|_2\|\mathrm{Retr}_{s^t}\big(-\tau\xi^{t+1}\big)-(s^{t}-\tau\xi^{t+1})\|\\
\le&-\tau\|\xi^{t+1}\|^2+\|\zeta^{t+1}\|_2\cdot L_2\tau^2\|\xi^{t+1}\|^2\\
\le&-\tau\|\xi^{t+1}\|^2+\frac{2\|AU\|_\infty^2L_2\tau^2}{\eta}\|\xi^{t+1}\|^2.
\end{aligned}
$$

Combining those inequalities above implies that

$$
F(u^{k+1},v^{k+1},t^{k+1})-F(u^{k+1},v^{k+1},t^{k})\le-\tau\left(1-\left[\frac{2\|AU\|_\infty^2L_2}{\eta}+\frac{\varrho}{2}L_1^2\right]\tau\right)\|\xi^{t+1}\|^2.
$$

Taking $\tau=\frac{1}{4\|AU\|_\infty^2L_2/\eta+\varrho L_1^2}$ gives the desired result. □

Lemma 11 (Decrease of F in v)**.** *Let $\{u^t,v^t,s^t\}_t$ be the sequence generated from Algorithm 2. The following inequality holds for any $k\ge1$:*

$$
F(u^{t+1},v^{t+1},s^{t})-F(u^{t+1},v^{t},s^{t})\le-\frac{1}{2}\|1/m-\pi(u^{t+1},v^{t},s^{t})^{\mathrm{T}}1\|_1^2.
$$

where

$$\|1/m - \pi(u^{t+1}, v^t, s^t)\|_1 = \sum_j \left| \frac{1}{m} - \sum_i \pi_{i,j}(u^{t+1}, v^t, s^t) \right|.$$

Proof of Lemma 11. According to the expression of F, we have that

$$\begin{aligned}
&F(u^{t+1}, v^{t+1}, s^t) - F(u^{t+1}, v^t, s^t) \\
=&\sum_{i,j} \pi_{i,j}(u^{t+1}, v^{t+1}, s^t) - \sum_{i,j} \pi_{i,j}(u^{t+1}, v^t, s^t) + \frac{1}{m}\sum_{j=1}^m (v_j^t - v_j^{t+1}) \\
=&\frac{1}{m}\sum_{j=1}^m (v_j^t - v_j^{t+1}) = -\frac{1}{m}\sum_{j=1}^m \log \frac{1/m}{\sum_i \pi_{i,j}(u^{t+1}, v^t, s^t)},
\end{aligned}$$

where the second equality is because that

$$\begin{aligned}
\sum_i \pi_{i,j}(u^{t+1}, v^{t+1}, s^t) &= \frac{1}{m}, \\
\sum_j \pi_{i,j}(u^{t+1}, v^t, s^t) &= \frac{1}{n}.
\end{aligned}$$

Therefore, applying the Pinsker's inequality in Theorem 9 implies that

$$F(u^{t+1}, v^{t+1}, s^t) - F(u^{t+1}, v^t, s^t) \leq -\frac{1}{2}\left(\sum_j \left| \frac{1}{m} - \sum_i \pi_{i,j}(u^{t+1}, v^t, s^t) \right| \right)^2.$$

□

Lemma 12 (Decrease of F in u)**.** *Let $\{u^t, v^t, s^t\}_t$ be the sequence generated from Algorithm 2. The following inequality holds for any $t \geq 1$:*

$$F(u^{t+1}, v^t, s^t) - F(u^t, v^t, s^t) \leq -\frac{1}{2}\|1/n - \pi(u^t, v^t, s^t)1\|_2^2.$$

where

$$\|1/n - \pi(u^t, v^t, s^t)1\|_2^2 = \sum_i \left| \frac{1}{n} - \sum_j \pi_{i,j}(u^k, v^k, t^k) \right|^2.$$

Proof of Lemma 12. For fixed $i \in [n]$, define

$$h_i = \sum_j \pi_{i,j}(u^{t+1}, v^t, s^t) - \sum_j \pi_{i,j}(u^t, v^t, s^t) - \frac{1}{n} \log \frac{1/n}{\sum_j \pi_{i,j}(u^t, v^t, s^t)}$$

According to the expression of F,

$$F(u^{t+1}, v^t, s^t) - F(u^t, v^t, s^t) = \sum_i h_i,$$

and it suffices to provide an upper bound for $h_i, i \in [n]$. By substituting the expression of u^{t+1} into h_i, we have that

$$\begin{aligned} h_i &= \sum_j \pi_{i,j}(u^t, v^t, s^t) \left[\frac{1/n}{\sum_j \pi_{i,j}(u^t, v^t, s^t)} - 1 \right] - \frac{1}{n} \log \frac{1/n}{\sum_j \pi_{i,j}(u^t, v^t, s^t)} \\ &= \frac{1}{n} - \big(\pi(u^t, v^t, s^t)1\big)_i - \frac{1}{n} \log \frac{1/n}{\big(\pi(u^t, v^t, s^t)1\big)_i} \end{aligned}$$

Define the function

$$\ell(x) = \frac{1}{n} - x - \frac{1}{n} \log \frac{1/n}{x} + (x - 1/n)^2.$$

We can see that this function attains its maximum at $x = 1/n$, with $\ell(1/n) = 0$. It follows that

$$h_i \leq - \left(\big(\pi(u^t, v^t, s^t)1\big)_i - \frac{1}{n} \right)^2.$$

The proof is completed. □

Lemma 13. *Let $\{u^t, v^t, s^t\}_t$ be the sequence generated from Algorithm 2, which is terminated when the following conditions hold:*

$$\|\xi^{t+1}\| \leq \epsilon_1, \quad \|1/n - \pi(u^t, v^t, s^t)1\|_2 \leq \frac{\epsilon_2}{4\|AU\|_\infty^2}, \quad \|1/m - \pi(u^{t+1}, v^t, s^t)^{\mathrm{T}}1\|_1 \leq \frac{\epsilon_2}{4\|AU\|_\infty^2}.$$

Then $\{u^T, v^T, s^T\}$ is an (ϵ_1, ϵ_2) stationary point of (2.5)*.*

Proof of Lemma 13. The condition $\|\xi^{t+1}\| \leq \epsilon_1$ directly implies that

$$\|\text{Grad}_s F(u^T, v^T, s^T)\| \leq \epsilon_1.$$

Suppose that

$$\pi(u^T, v^T, s^T)1 = r, \quad \pi(u^T, v^T, s^T)^{\mathrm{T}}1 = c,$$

where $\|1/n - r\|_2 \leq \epsilon_2/(4\|AU\|_\infty^2)$ and $\|1/m - c\|_1 \leq \epsilon_2/(4\|AU\|_\infty^2)$. Then we find that

$$F(u^T, v^T, s^T) = \min_{\pi} \left\{ \sum_{i,j} \pi_{i,j} M_{i,j} - \eta H(\pi) : \sum_j \pi_{i,j} = r_i, \sum_i \pi_{i,j} = c_j \right\},$$

and

$$\min_{u,v} F(u, v, s^T) = \min_{\pi} \left\{ \sum_{i,j} \pi_{i,j} M_{i,j} - \eta H(\pi) : \sum_j \pi_{i,j} = \frac{1}{n}, \sum_i \pi_{i,j} = \frac{1}{m} \right\},$$

where $M_{i,j} = \|A_{i,j} U s^T\|_2^2$. It follows that

$$\begin{aligned} &F(u^T, v^T, s^T) - \min_{u,v} F(u, v, s^T) \\ \leq &\eta \log(mn) + 2\|1/m - c_1\|_1 \times \|AU\|_\infty^2 \leq \epsilon_2, \end{aligned}$$

where the last inequality is by taking $\eta = \epsilon_2/(2\log(mn))$.

□

Lemma 14 (Lower Boundedness of F)**.** *Denote by* (u^*, v^*, s^*) *the global optimum of* (2.5)*. Then we have that*

$$F(u^*, v^*, s^*) \geq 1 - \frac{1}{\eta}\|AU\|_\infty^2.$$

Proof of Lemma 14. It is easy to show that

$$\sum_{i,j} \pi_{i,j}(u^*, v^*, s^*) = 1.$$

Moreover, for any (i,j), we have that $c_{i,j} \leq \|AU\|_\infty^2$. It follows that

$$\exp\left(-\frac{1}{\eta}\|AU\|_\infty^2 + u_i^* + v_j^*\right) \leq \pi_{i,j} \leq 1,$$

and therefore $u_i^* + v_j^* \leq \frac{1}{\eta}\|AU\|_\infty^2$ for any (i,j). Hence we conclude that

$$\sum_{i,j} \pi_{i,j}(u^*, v^*, s^*) - \frac{1}{n}\sum_{i=1}^n u_i - \frac{1}{m}\sum_{j=1}^m v_j \geq 1 - \frac{1}{\eta}\|AU\|_\infty^2.$$

□

In the following we give a re-statement of Theorem 2 and the formal proof.

Theorem (Re-statement of Theorem 2)**.** *Choose parameters*

$$\tau = \frac{1}{4\|AU\|_\infty^2 L_2/\eta + \varrho L_1^2}, \quad \eta = \frac{\epsilon_2}{2\log(mn)}, \quad \varrho = \frac{2\|AU\|_\infty^2}{\eta} + \frac{4\|AU\|_\infty^4}{\eta^2},$$

and Algorithm 2 terminates when

$$\|\xi^{t+1}\| \leq \epsilon_1, \quad \|1/n - \pi(u^t, v^t, s^t)1\|_2 \leq \frac{\epsilon_2}{4\|AU\|_\infty^2}, \quad \|1/m - \pi(u^{t+1}, v^t, s^t)^{\mathrm{T}}1\|_1 \leq \frac{\epsilon_2}{4\|AU\|_\infty^2}.$$

We say that $(\hat{u}, \hat{v}, \hat{s})$ *is a* (ϵ_1, ϵ_2)*-stationary point of* (2.5) *if*

$$\|\mathit{Grad}_s F(\hat{u}, \hat{v}, \hat{s})\| \leq \epsilon_1,$$
$$F(\hat{u}, \hat{v}, \hat{s}) - \min_{u,v} F(u, v, \hat{s}) \leq \epsilon_2,$$

where $\mathit{Grad}_s F(u, v, s)$ *denotes the partial derivative of* F *with respect to the variable* s *on the sphere* $\mathbb{S}^{d(n+m)-1}$*. Then Algorithm 2 returns an* (ϵ_1, ϵ_2)*-stationary point in iterations*

$$T = \mathcal{O}\left(\log(mn) \cdot \left[\frac{1}{\epsilon_2^3} + \frac{1}{\epsilon_1^2 \epsilon_2}\right]\right).$$

Proof of Theorem 2. We can build the one-iteration descent result based on Lemma 10,

Lemma 11, and Lemma 12:

$$
\begin{aligned}
&F(u^{t+1}, v^{t+1}, s^{t+1}) - F(u^t, v^t, s^t) \\
\leq& -\left(\frac{1}{2}\|1/n - \pi(u^t, v^t, s^t)1\|_2^2 + \frac{1}{2}\|1/m - \pi(u^{t+1}, v^t, s^t)^{\mathrm{T}}1\|_1^2 + \frac{1}{8\|AU\|_\infty^2 L_2/\eta + 2\varrho L_1^2}\|\xi^{t+1}\|_2^2\right) \\
=& -\frac{1}{2}\left(\|1/n - \pi(u^t, v^t, s^t)1\|_2^2 + \|1/m - \pi(u^{t+1}, v^t, s^t)^{\mathrm{T}}1\|_1^2 + \frac{\eta^2\|\zeta^{t+1}\|^2}{2\|AU\|_\infty^2\eta(2L_2 + L_1^2) + 4\|AU\|_\infty^4 L_1^2}\right)
\end{aligned}
$$

Then we have that

$$
\begin{aligned}
&F(u^T, v^T, s^T) - F(u^0, v^0, s^0) \\
\leq& -\frac{1}{2}\sum_{t=0}^{T-1}\left(\|1/n - \pi(u^t, v^t, s^t)1\|_2^2 + \|1/m - \pi(u^{t+1}, v^t, s^t)^{\mathrm{T}}1\|_1^2 + \frac{\eta^2\|\zeta^{t+1}\|^2}{2\|AU\|_\infty^2\eta(2L_2 + L_1^2) + 4\|AU\|_\infty^4 L_1^2}\right) \\
\leq& -\frac{1}{2}\cdot\min\left\{1, \frac{1}{2\|AU\|_\infty^2\eta(2L_2 + L_1^2) + 4\|AU\|_\infty^4 L_1^2}\right\} \\
&\qquad\times\sum_{t=0}^{T-1}\left(\|1/n - \pi(u^t, v^t, s^t)1\|_2^2 + \|1/m - \pi(u^{t+1}, v^t, s^t)^{\mathrm{T}}1\|_1^2 + \eta^2\|\xi^{t+1}\|_2^2\right) \\
\leq& -\frac{1}{2}T\cdot\min\left\{1, \frac{1}{2\|AU\|_\infty^2\eta(2L_2 + L_1^2) + 4\|AU\|_\infty^4 L_1^2}\right\}\cdot\min\left\{\epsilon_1^2, \frac{\epsilon_2^2}{16\|AU\|_\infty^4}, \frac{\epsilon_2^2}{16\|AU\|_\infty^4}\right\}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
T \leq& [F(u^0, v^0, t^0) - F(u^T, v^T, s^T)]\max\left\{2, 4\|AU\|_\infty^2\eta(2L_2 + L_1^2) + 8\|AU\|_\infty^4 L_1^2\right\} \\
&\qquad\max\left\{\frac{1}{\epsilon_1^2}, \frac{16\|AU\|_\infty^4}{\epsilon_2^2}, \frac{16\|AU\|_\infty^4}{\epsilon_2^2}\right\} \\
\leq& \left(F(u^0, v^0, t^0) - 1 + \frac{\|AU\|_\infty^2}{\eta}\right)\max\left\{2, 4\|AU\|_\infty^2\eta(2L_2 + L_1^2) + 8\|AU\|_\infty^4 L_1^2\right\} \\
&\qquad\max\left\{\frac{1}{\epsilon_1^2}, \frac{16\|AU\|_\infty^4}{\epsilon_2^2}, \frac{16\|AU\|_\infty^4}{\epsilon_2^2}\right\} \\
=& \mathcal{O}\left(\log(mn)\cdot\left[\frac{1}{\epsilon_2^3} + \frac{1}{\epsilon_1^2\epsilon_2}\right]\right).
\end{aligned}
$$

□

A.5 Technical Proofs in Section 2.4

A.5.1 Proof of Theorem 3

Proof of Lemma 1. Denote $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$. By the bias-variation decomposition, we have that

$$\mathbb{E}[(\mathcal{KP}W(\hat{\mu}_n, \mu))^{1/p}] \leq \sup_{f\in\mathcal{F}} \mathbb{E}[(W(f\#\hat{\mu}_n, f\#\mu))^{1/p}] + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left((W(f\#\hat{\mu}_n, f\#\mu))^{1/p} - \mathbb{E}[(W(f\#\hat{\mu}_n, f\#\mu))^{1/p}]\right)\right].$$

For fixed $f \in \mathcal{F}$, we can see that

$$\mathbb{E}[(W(f\#\hat{\mu}_n, f\#\mu))^{1/p}] \leq c_p n^{-\frac{1}{(2p)\vee d}} (\log n)^{\zeta_{p,d}/p}$$

where c_p is a constant depending only on p and

$$\zeta_{p,d} = \begin{cases} 1, & \text{if } d = 2p, \\ 0, & \text{otherwise.} \end{cases}$$

Now we start to upper bound the variation term. Define the empirical process

$$X_f = (W(f\#\hat{\mu}_n, f\#\mu))^{1/p} - \mathbb{E}[(W(f\#\hat{\mu}_n, f\#\mu))^{1/p}].$$

It is easy to see that $\mathbb{E}[X_f] = 0$. Moreover, we can show that for fixed f, the random variable X_f is sub-exponential. Denote by $Z = \{z_i\}_{i=1}^n$ and $Z' = \{z_i'\}_{i=1}^n$ i.i.d. samples from $f\#\mu$. Take $g(Z) = (W(f\#\hat{\mu}_n, f\#\mu))^{1/p}$. Then we have that

$$|g(Z) - g(Z'_{(i)})| \leq (W(f\#\hat{\mu}_n, f\#\hat{\mu}'_n))^{1/p} \leq n^{-1/(2\vee p)} \|Z - Z'\|_2.$$

It follows that

$$\sum_{i=1}^n \|\nabla_i g(Z)\|^2 \leq n^{-2/(2\vee p)}, \quad \max_{1\leq i\leq n} \|\nabla_i g(Z)\| \leq n^{-1/p}.$$

Then the Poincare's inequality in Theorem 11 implies that

$$\Pr\{X_f \geq t\} \leq \exp\left(-K^{-1}\min\{tn^{1/p}, t^2 n^{2/(2\vee p)}\}\right).$$

Hence we conclude that X_f is sub-exponential with parameters $(\sqrt{K/2}n^{-1/(2\vee p)}, (K/2)n^{-1/p})$.

For the function space $\mathcal{F}$, define the corresponding metric

$$\mathsf{d}(f, f') = \|f - f'\|_{\mathcal{H}}.$$

Let $X \sim \mu$. Then for any $f, f' \in \mathcal{F}$, we have that

$$\begin{aligned}
&|X_f - X_{f'}| \\
\leq &\mathbb{E}\big[\,(W(f\#\hat{\mu}_n, f'\#\hat{\mu}_n))^{1/p} + (W(f\#\mu, f'\#\mu))^{1/p}\,\big] + \mathbb{E}\big[\,(W(f\#\hat{\mu}_n, f'\#\hat{\mu}_n))^{1/p} + (W(f\#\mu, f'\#\mu))^{1/} \\
\leq &2\left(\mathbb{E}\|f(X) - f'(X)\|_2^p\right)^{1/p} + \left(\frac{1}{n}\sum_{i=1}^n \|f(X_i) - f'(X_i)\|_2^p\right)^{1/p} + \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \|f(X_i) - f'(X_i)\|_2^p\right)^{1/p}\right]
\end{aligned}$$

Note that the following upper bound holds for any $f, f' \in \mathcal{F}$ and $x \in \mathbb{R}^D$:

$$\begin{aligned}
\|f(x) - f'(x)\|_2 &= \max_{a:\ \|a\|_2 \leq 1} \langle f(x) - f'(x), a\rangle \\
&= \max_{a:\ \|a\|_2 \leq 1} \langle f(x), a\rangle - \langle f'(x), a\rangle \\
&= \max_{a:\ \|a\|_2 \leq 1} \langle f, K_x a\rangle_{\mathcal{H}_K} - \langle f', K_x a\rangle_{\mathcal{H}_K} \\
&= \max_{a:\ \|a\|_2 \leq 1} \langle f - f', K_x a\rangle_{\mathcal{H}_K} \\
&\leq \|f - f'\|_{\mathcal{H}_K} \times \max_{a:\ \|a\|_2 \leq 1} \|K_x a\|_{\mathcal{H}_K} \\
&= \|f - f'\|_{\mathcal{H}_K} \times \max_{a:\ \|a\|_2 \leq 1} \sqrt{a^{\mathrm{T}} K(x, x) a} \\
&= \sqrt{B}\|f - f'\|_{\mathcal{H}_K}.
\end{aligned}$$

As a consequence, substituting this upper bound into the relation above implies that

$$|X_f - X_{f'}| \leq 4\sqrt{B}\mathsf{d}(f, f').$$

Applying the ϵ-net argument similar to the Dudley's entropy integral bound [155, Theorem 5.22] gives

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} X_f\right] \leq \inf_{\epsilon>0}\left\{4\sqrt{B}\epsilon + \sqrt{2K}n^{-1/(2\vee p)}\sqrt{\log\mathcal{N}(\mathcal{F},\mathsf{d},\epsilon)} + (K/2)n^{-1/p}\log\mathcal{N}(\mathcal{F},\mathsf{d},\epsilon)\right\}$$

Taking $\mathcal{N}(\mathcal{F},\mathsf{d},\epsilon) = \left\lceil\frac{1}{\epsilon}\right\rceil$ and $\epsilon = n^{-1/p}$ implies that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} X_f\right] \lesssim n^{-1/(2\vee p)}\sqrt{\log(n)} + n^{-1/p}\log(n).$$

□

Proof of Lemma 2. We start to upper bound the variance term

$$(\mathcal{KPW}(\hat{\mu}_n,\mu))^{1/p} - \mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n,\mu))^{1/p}].$$

Denote by $X = \{x_i\}_{i=1}^n$ and $X' = \{x_i'\}_{i=1}^n$ i.i.d. samples from μ, and let $g(X) = (\mathcal{KPW}(\hat{\mu}_n,\mu))^{1/p}$. Based on the triangular inequality, we find that

$$\begin{aligned}
|g(X) - g(X')| &\leq n^{-1/p}\left(\sum_{i=1}^n \max_{f\in\mathcal{F}} \|f(x_i) - f(x_i')\|_2\right)^{1/p} \\
&\leq n^{-1/p}\left(\sum_{i=1}^n L\|x_i - x_i'\|\right)^{1/p} \\
&\leq n^{-1/(2\vee p)}L^{1/p}\|X - X'\|.
\end{aligned}$$

It follows that

$$\sum_{i=1}^n \|\nabla_i g(Z)\|^2 \leq n^{-2/(2\vee p)}L^{2/p}, \quad \max_{1\leq i\leq n}\|\nabla_i g(Z)\| \leq n^{-1/p}L^{1/p}.$$

Then the Poincare's inequality in Theorem 11 implies that

$$\Pr\{\left|(\mathcal{KPW}(\hat{\mu}_n,\mu))^{1/p} - \mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n,\mu))^{1/p}]\right| \geq t\} \leq \exp\left(-K^{-1}\min\{tn^{1/p}L^{-1/p}, t^2n^{2/(2\vee p)}L^{-2/p}\}\right).$$

Substituting the right-hand-side with α completes the proof. □

Proof of Theorem 3. Based on the triangular inequality, we can see that

$$\left| \left(\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m)\right)^{1/p} - \left(\mathcal{KPW}(\mu, \nu)\right)^{1/p} \right| \leq \left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p} + \left(\mathcal{KPW}(\hat{\nu}_m, \nu)\right)^{1/p}.$$

It suffices to upper bound $\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p}$ and $\left(\mathcal{KPW}(\hat{\nu}_m, \nu)\right)^{1/p}$ separately. By the bias-variance decomposition,

$$\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p} \leq \mathbb{E}[\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p}] + \left(\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p} - \mathbb{E}[\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p}] \right),$$

where the first term quantifies the bias for empirical estimation, and the second term quantifies the variance of estimation. The bias term can be upper bounded by applying Lemma 1, and the variance term can be upper bounded by applying Lemma 2. In summary, with probability at least $1 - \alpha$, it holds that

$$\begin{aligned}\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p} \lesssim & \max\left\{ n^{-1/p} K \log(1/\alpha), n^{-1/(2\vee p)} \sqrt{K \log(1/\alpha)} \right\} L^{1/p} \\ & + n^{-\frac{1}{(2p)\vee d}} (\log n)^{\zeta_{p,d}/p} + n^{-1/(2\vee p)} \sqrt{\log(n)} + n^{-1/p} \log(n).\end{aligned}$$

The upper bound for $\left(\mathcal{KPW}(\hat{\nu}_m, \nu)\right)^{1/p}$ can be proceeded similarly.

□

A.5.2 Testing Performance

Based on the finite-sample guarantee in Theorem 3, we are able to characterize the performance of the KPW test. To make the type-I error below than α, we reject the null hypothesis as long as the empirical statistic $\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m) \geq \gamma_{m,n}$, where

$$\begin{aligned}\gamma_{m,n}^{1/p} \sim & \max\left\{ N^{-1/p} K \log(1/\alpha), N^{-1/(2\vee p)} \sqrt{K \log(1/\alpha)} \right\} L^{1/p} \\ & + N^{-\frac{1}{(2p)\vee d}} (\log N)^{\zeta_{p,d}/p} + N^{-1/(2\vee p)} \sqrt{\log(n)} + N^{-1/p} \log(n).\end{aligned}$$

For the alternative hypothesis, assume that target distributions μ and ν satisfy $\mathcal{KPW}(\mu,\nu) > \gamma_{m,n}$. Then the type-II error can be upper bounded as

$$\begin{aligned}
&\Pr{}_{\mathcal{H}_1}\left(\mathcal{KPW}(\hat{\mu}_n,\hat{\nu}_m) < \gamma_{m,n}\right)\\
=&\Pr{}_{\mathcal{H}_1}\left(\mathcal{KPW}(\hat{\mu}_n,\hat{\nu}_m) - \mathcal{KPW}(\mu,\nu) < \gamma_{m,n} - \mathcal{KPW}(\mu,\nu)\right)\\
=&\Pr{}_{\mathcal{H}_1}\left(\mathcal{KPW}(\mu,\nu) - \mathcal{KPW}(\hat{\mu}_n,\hat{\nu}_m) > \mathcal{KPW}(\mu,\nu) - \gamma_{m,n}\right)\\
\leq&\Pr{}_{\mathcal{H}_1}\left(|\mathcal{KPW}(\mu,\nu) - \mathcal{KPW}(\hat{\mu}_n,\hat{\nu}_m)| > \mathcal{KPW}(\mu,\nu) - \gamma_{m,n}\right)\\
\leq&\frac{\mathbb{E}\left(\mathcal{KPW}(\mu,\nu) - \mathcal{KPW}(\hat{\mu}_n,\hat{\nu}_m)\right)^2}{\left(\mathcal{KPW}(\mu,\nu) - \gamma_{m,n}\right)^2}.
\end{aligned}$$

A.5.3 Finite-sample Guarantee for $p \in [1,2)$

In this subsection, we discuss the finite-sample guarantee for KPW distance with p-Wasserstein distance for $p \in [1,2)$. Note that it is not necessary to rely on the Poincare inequality or projection poincare inequality to obtain the result. We first present several technical lemmas before showing the final result.

Lemma 15. *Based on Assumption 1, for $f \in \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$, we have*

$$\|f(x)\|_2 \leq \sqrt{B}, \quad \forall x \in \mathbb{R}^D.$$

Proof of Lemma 15. For fixed $x \in \mathcal{X}$, the norm of $f(x)$ can be upper bounded as the following:

$$\|f(x)\|_2^2 = \langle f(x), f(x)\rangle = \langle f, K_x f(x)\rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}}\|K_x f(x)\|_{\mathcal{H}} \leq \|K_x f(x)\|_{\mathcal{H}}.$$

In particular,

$$\begin{aligned}\|K_x f(x)\|_{\mathcal{H}}^2 &= \langle K_x f(x), K_x f(x)\rangle_{\mathcal{H}} \\ &= \langle \big(K_x f(x)\big) f(x), f(x)\rangle \\ &= \langle K(x, f(x)) f(x), f(x)\rangle \\ &= f(x)^{\mathrm{T}} K(x, f(x)) f(x) \\ &\leq B\|f(x)\|_2^2\end{aligned}$$

Combining those two relations above implies the desired result. □

Lemma 16. *For $p \in [1, 2)$, the bias term of empirical KPW distance can be upper bounded as*

$$\mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p}] \lesssim n^{-\frac{1}{(2p)\vee d}} (\log n)^{\zeta_{p,d}/p} + n^{1/2-1/p}\sqrt{\log(n)} + n^{-1/p}.$$

where $\zeta_{p,d} = 1$ if $d = 2p$ and $\zeta_{p,d} = 0$ otherwise.

Proof of Lemma 16. Following the similar argument as in Lemma 1, we can see that

$$\begin{aligned}\mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p}] \leq \sup_{f\in\mathcal{F}} \mathbb{E}[(W(f\#\hat{\mu}_n, f\#\mu))^{1/p}] \\ + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \Big((W(f\#\hat{\mu}_n, f\#\mu))^{1/p} - \mathbb{E}[(W(f\#\hat{\mu}_n, f\#\mu))^{1/p}]\Big)\right],\end{aligned}$$

and the first term can also be bounded similarly. To upper bound the second term, define the empirical process $\{X_f\}$ as in Lemma 1. For fixed f, the random variable X_f can be shown to be sub-Gaussian. Denote by $Z = \{z_i\}_{i=1}^n$ and $Z'_{(i)}$ the sample set so that the i-th element is different. Take $g(Z) = (W(f\#\hat{\mu}_n, f\#\mu))^{1/p}$. Then we have that

$$\begin{aligned}|g(Z) - g(Z'_{(i)})| &\leq (W(f\#\hat{\mu}_n, f\#\hat{\mu}'_n))^{1/p} \leq \left(\frac{1}{n}\|f(z_i) - f(z'_i)\|_2^p\right)^{1/p} \\ &\leq n^{-1/p} 2\sqrt{B}.\end{aligned}$$

Therefore, applying the McDiarmid's inequality in Theorem 10 implies

$$\Pr\{|X_f| \geq u\} \leq 2\exp\left(-\frac{u^2}{2Bn^{1-2/p}}\right).$$

Applying Lemma 8 implies that for fixed ℓ, the random variable X_f is sub-Gaussian with the parameter $\sigma^2 = 36Bn^{1-2/p}$. Then applying the ϵ-net argument similar to the Dudley's entropy integral bound [155, Theorem 5.22] gives

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} X_f\right] \leq \inf_{\epsilon>0}\left\{4\sqrt{B}\epsilon + \sqrt{36Bn^{1-2/p}}\sqrt{2\log\mathcal{N}(\mathcal{F},\mathsf{d},\epsilon)}\right\}.$$

Taking $\mathcal{N}(\mathcal{F},\mathsf{d},\epsilon) = \left\lceil\frac{1}{\epsilon}\right\rceil$ and $\epsilon = n^{-1/p}$ implies that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} X_f\right] \lesssim n^{1/2-1/p}\sqrt{\log(n)} + n^{-1/p}.$$

□

Lemma 17. *For $p \in [1,2)$, with with probability at least $1-\alpha$, it holds that*

$$\left|(\mathcal{KPW}(\hat{\mu}_n,\mu))^{1/p} - \mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n,\mu))^{1/p}]\right| \leq n^{1/2-1/p}\sqrt{2B\log\frac{2}{\alpha}}.$$

Proof of Lemma 17. Denote by $Z = \{z_i\}_{i=1}^n$ and $Z'_{(i)}$ the sample set so that the i-th element is different. Take $g(Z) = (\mathcal{KPW}(\hat{\mu}_n,\mu))^{1/p}$. Then we can see that

$$|g(Z) - g(Z'_{(i)})| \leq (\mathcal{KPW}(\hat{\mu}_n,\hat{\mu}'_n))^{1/p} \leq n^{-1/p}2\sqrt{B}.$$

Then applying the McDiarmid's inequality in Theorem 10 implies

$$\Pr\left\{\left|(\mathcal{KPW}(\hat{\mu}_n,\mu))^{1/p} - \mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n,\mu))^{1/p}]\right| \geq u\right\} \leq 2\exp\left(-\frac{u^2}{2Bn^{1-2/p}}\right).$$

□

Based on Lemma 16 and Lemma 17, we obtain the uncertainty quantification result in Theorem 4.

A.6 Implementation Details for Computing KPW Distance

The variable s is initialized to be a uniform random vector over sphere. The dual variable v is initialized to be a Gaussian random vector with unit covariance. When updating the block of variables u^{t+1} and v^{t+1}, we make the change of variables $(u')^{t+1} = \exp(u^{t+1})$ and $(v')^{t+1} = \exp(v^{t+1})$. We update $(u')^{t+1}$ and $(v')^{t+1}$ instead to accelerate the computation:

$$(u')^{t+1} = \left\{ \frac{1/n}{\sum_j \exp\left(-\frac{1}{\eta}c_{i,j} + (v'_j)^t\right)} \right\}_i$$
$$(v')^{t+1} = \left\{ \frac{1/m}{\sum_i \exp\left(-\frac{1}{\eta}c_{i,j} + (u'_i)^{t+1}\right)} \right\}_j,$$

and we further store the matrix A with $A_{i,j} = \exp\left(-\frac{1}{\eta}c_{i,j}\right)$ in advance to reduce the computational cost. The transport mapping $\pi^{t+1} \triangleq (\pi_{i,j}(u^{t+1}, v^{t+1}, s^t))_{i,j}$ can be formulated without going through a for loop but only with multiplication operators:

$$\pi^{t+1} = (u')^{t+1} \mathbin{.*} A \mathbin{.*} [(v')^{t+1}]^{\mathrm{T}},$$

where the operator $.*$ means we multiply two objects componentiwisely in terms of array broadcasting. When updating ζ^{t+1}, we first formulate the matrix V^{t+1} with

$$V_{i,j}^{t+1} = \sum_{i,j} \pi_{i,j}^{t+1} A_{i,j}^{\mathrm{T}} A_{i,j}$$

and then continue the matrix multiplication procedure in (2.6i). Denote by G_i the i-th row block of the gram matrix G, then

$$\begin{aligned} V^{t+1} &= \left\{ \sum_{i,j} \pi_{i,j}^{t+1} (G_i + G_{n+j})^{\mathrm{T}} (G_i + G_{n+j}) \right\}_{i,j} \\ &= \left\{ \sum_{i,j} \pi_{i,j}^{t+1} (G_i^{\mathrm{T}} G_i + G_{n+j}^{\mathrm{T}} G_{n+j} + G_{n+j}^{\mathrm{T}} G_i + G_i^{\mathrm{T}} G_{n+j}) \right\}_{i,j}. \end{aligned}$$

Consequently, we can compute each of the four components in the formula above without executing double for loops and then sum them up to obtain the matrix V^{t+1}. During the numerical implementation, we also find that the computation is sensitive to the choice of η. This phenomenon has also been observed when using Sinkhorn's algorithm to compute Wasserstein distance or projected Wasserstein distance. When η is too small, the iteration update may have numerical instability issues. When η is too large, the obtained solution is far away from the optimal solution to the original KPW distance. We have tried the best to tune this parameter to make the algorithm maintain the best performance. How to tune this hyper-parameter systematically is left for future works.

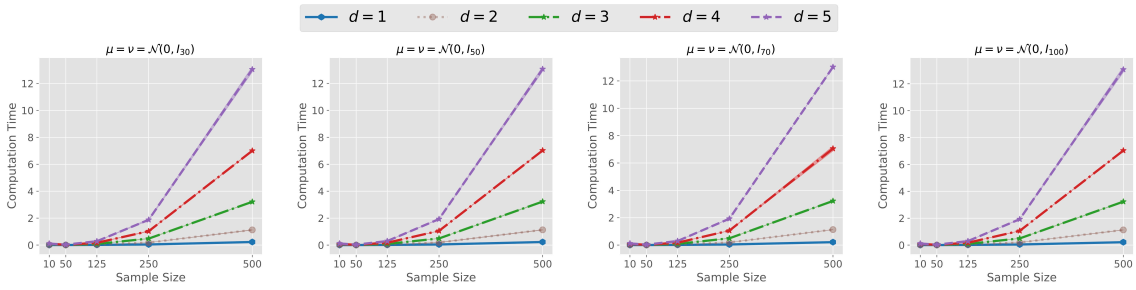

Figure A.1: Mean computation time for computing $\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_n)$ for varying n. Results are averaged over 10 independent trials.

A.7 Details about Experiment

A.7.1 Sample Complexity

In this experiment, we fix hyper-parameters $\sigma^2 = 1, \rho = 0.5$ for computing KPW distances. The values of empirical KPW distances across different choices of sample size are reported in Figure 2.1, and the corresponding computation time is reported in Figure A.1. From the plot we can see that it is efficient to compute KPW distances with reasonably small sample size n and projected dimension d.

A.7.2 Configurations

All methods are implemented using python 3.7 (Pytorch 1.1) on a MacBook Pro labtop with 32GB of memory. When running the code, there is no swapping of memory and the average CPU frequency is 3.2 GHZ. We compute the projected Wasserstein distance based on the official code in https://github.com/fanchenyou/PRW. We run the MMD-O test based on the code in https://github.com/fengliu90/DK-for-TST. We run the MMD-NTK test based on the code in https://github.com/xycheng/NTK-MMD. From extensive experiments we realize that MMD-NTK is the most computationally efficient test, but its power does not scale the best. On the other hand, this method can be useful when performing a test for the large-sampled case, while our method may be intractable to compute in short time. We run the ME test based on the code in https://github.com/wittawatj/interpretable-test.

A.7.3 Implementation of Cross-Validation

The candidate choices of hyper-parameters ρ and σ^2 are within the set

$$\{(\rho, \sigma^2) : \ \sigma^2 = a \cdot \hat{\sigma}^2 : \ a \in \{0.5, 1, 2\}, \rho \in \{0.25, 0.5, 0.75\}\},$$

where $\hat{\sigma}^2$ denotes the empirical median of pairwise distances between observations. To choose ρ and σ^2, we further split the training set into the training and validation dataset, which contain 70% and 30% data, respectively. For each choice of hyper-parameters we use the training dataset to obtain a nonlinear projector and examine its hold-out performance on the validation dataset, which is quantified as the negative of the p-value for two-sample tests between two collection of samples in the validation dataset. We choose hyper-parameters ρ and σ^2 with the best hold-out performance.

A.7.4 Tests for Synthetic Datasets

When studying tests on Gaussian distributions, we take both the training and testing sample sizes N to be 50. When reproducing the experiments corresponding to the left two figures in Fig. 2.3, we take the dimension $D \in \{20, 40, 60, 80, 100, 120, 140, 160\}$. When reproducing the experiments corresponding to the right two figures, we take the sample size $n = m \in \{80, 100, 140, 180, 250\}$.

A.7.5 Tests for MNIST handwritten digits

Table A.1 present the type-I error for various tests in MNIST dataset, from which we can see that all tests have the type-I error close to $\alpha = 0.05$.

A.7.6 Human activity detection

The pre-processing of data is as follows. We first remove frames in which the person is standing still or with little movements. Then we delete the first few frames to make the

Table A.1: Average type-I error and standard error for two-sample tests in *MNIST* dataset across different choices of sample size.

N	MMD-NTK	MMD-O	ME	PW	KPW
200	0.057±0.0010	0.056±0.0006	0.044±0.0003	0.056±0.0004	0.061±0.0005
250	0.051±0.0003	0.060±0.0001	0.065±0.0002	0.046±0.0003	0.048±0.0002
300	0.068±0.0006	0.055±0.0003	0.059±0.0007	0.056±0.0002	0.053±0.0001
400	0.049±0.0007	0.058±0.0002	0.041±0.0002	0.061±0.0006	0.056±0.0006
500	0.061±0.0006	0.054±0.0004	0.060±0.0002	0.049±0.0003	0.047±0.0004
Avg.	0.057	0.056	0.053	0.054	0.053

action of bending consist of 500 frames. Next we delete the last few frames to make the action of throwing consist of 355 frames. We take the window size $W = 100$. To perform online change point detection, we pre-train a nonlinear projector using the data before time index 300 and compute the null statistics for many times to obtain the true threshold. Then we compute the detection statistic by comparing the distribution between the block of data before time 300 and the data from the sliding window. We reject the null hypothesis and claim a change is happened if the statistic is above the threshold. The plot of the detection statistic over time after the time index 400 is presented in Fig. A.2, and the delay detection time corresponding to all users are reported in Table 2.2.

A.8 Impact of Hyper-parameters

A.8.1 Impact of Projected Dimension d

We prefer to choose the projected dimension d with relatively small values since the testing statistic will have poor sample complexity rate and is expansive to compute for large d. In this section, we examine the testing performance for different choices of d. In particular, we perform the KPW test on Gaussian distributions (with diagonal covariance matrices, $D = 128$ and $n = m = 50$) and Gaussian mixture distributions (with $D = 100$ and $n = m = 100$) following the setup in Section 2.5.1, the results of which are reported in Fig. A.3. From the plot we can see that the testing power is generally better for $d > 1$, which suggests that using vector-valued RKHS is better than using classical scalar-valued RKHS.

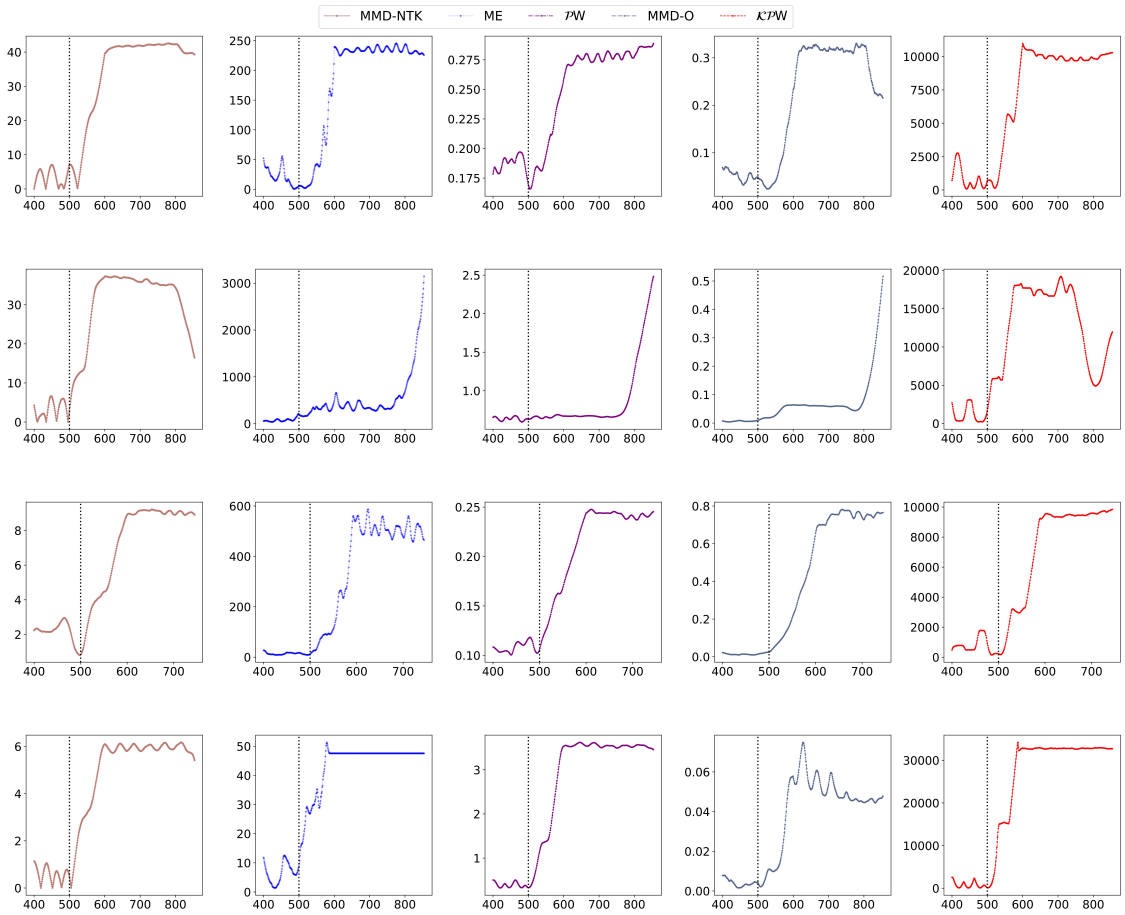

Figure A.2: Comparison of detection statistics from bending to throwing for various testing procedures. Black dash line indicates the true change-point. Each row corresponds to detection results for each user.

Moreover, we observe the performance is insensitive to the choice of d as long as we take $d > 1$.

A.8.2 Impact of Entropic Regularization Parameter η

As pointed out in [65], the entropic regularization in (2.4) could alerady improve the sample complexity result of Wasserstein distance. We perform experiments in this subsection to validate the impact of the entropic regularization parameter η for the performance of KPW test. The generated data follows Gaussian distributions (with $n = m = 100$) or Gaussian mixture distributions (with $n = m = 200$) with different choices of dimension D and fixed sample size. Benchmark methods include 1) *KPW test with* $\eta = 0$ (here Wasserstein distance

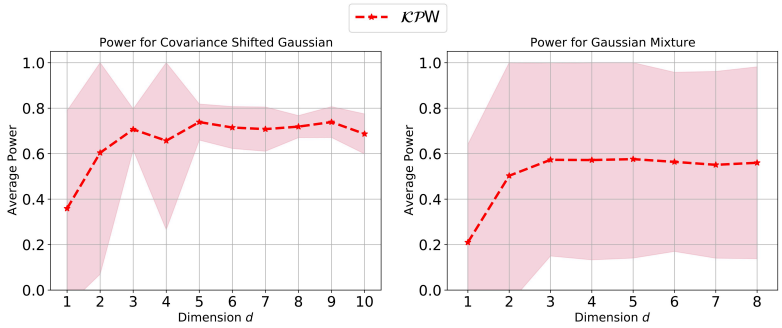

Figure A.3: Average power for KPW test across different choices of projected dimension d. Left: Gaussian distribution; Right: Gaussian mixture distribution. Results are averaged over 10 independent trials.

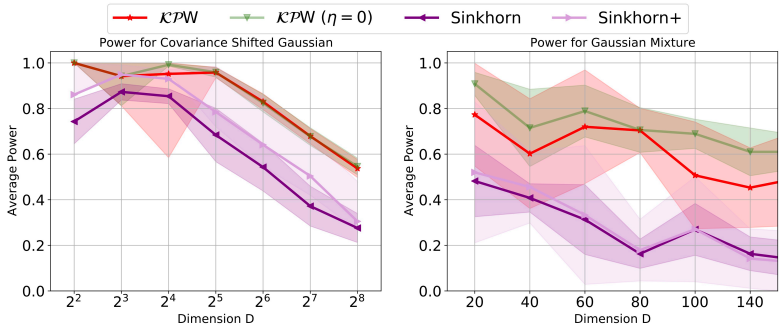

Figure A.4: Average power for KPW tests and Sinkhorn tests across different choices of data dimension D. Left: Gaussian distribution; Right: Gaussian mixture distribution. Results are averaged over 10 independent trials.

is computed exactly and we apply alternating optimization procedure as a heuristic); 2) *Sinkhorn test* with the same η as in the KPW test (in which we take the Sinkhorn divergence as the statistic and all training and testing samples are used); 3) *Sinkhorn+* (using all data and post-selecting η with the best performance). Experiment results are reported in Fig. A.4, from which we can see that even Sinkhorn+ test has the curse of dimension issue. Moreover, the KPW test with $\eta = 0$ has similar performance as the KPW test. Hence, we can assert that the KPW test is capable of alleviating the curse of dimension mainly due to the kernel projection operator instead of the entropic regularization.

A.9 Societal Impact

Two-sample testing is not only a fundamental problem in statistics but also growing increasing attention in machine learning. On the one hand, it plays a key role in modern applications such as anomaly detection and health care. On the other hand, it can help to design better algorithms for artificial intelligence such as GANs. Our work shows a competitive performance for dealing with high-dimensional data by nonlinear dimensionality reduction using kernel trick. It identifies the difference between two collections of samples by extracting the most representative nonlinear features. We hope this work can be applied to design more powerful algorithms in those areas.

APPENDIX B
SINKHORN DISTRIBUTIONALLY ROBUST OPTIMIZATION

B.1 Sufficient condition for Condition 1

Proposition 3. *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 $\widehat{P}$ has a finite mean, denoted as $\overline{x}$. Moreover, $\nu\{z : 0 \leq c(\overline{x}, z) < \infty\} = 1$ and*

$$\Pr_{x\sim\widehat{P}}\{c(x, \overline{x}) < \infty\} = 1.$$

(III) *Assumption 2*(III) *holds, and there exists $\lambda > 0$ such that*

$$\int e^{f(z)/(\lambda\epsilon)} e^{-2^{1-p} c(\overline{x},z)/\epsilon} \, \mathrm{d}\nu(z) < \infty.$$

We make some remarks for the sufficient conditions listed above. The first condition can be satisfied by taking the cost function as the p-th power of the metric defined on $\mathcal{Z}$ for any $p \geq 1$. The second condition requires the nominal distribution $\widehat{P}$ is finite almost surely, e.g., it can be a subguassian distribution with respect to the cost function c. Combining three conditions together and levering concentration arguments completes the proof of Proposition 3.

B.2 Detailed Experiment Setup

All the experiments are preformed on a MacBook Pro laptop with 32GB of memory running python 3.7. Unless otherwise specified, the SAA, Wasserstein DRO, and KL-divergence DRO baseline models are solved exactly based on the interior point method-based solver Mosek [5]. Optimization hyper-parameters (including step size, maximum iterations, number of levels, etc.) are tuned such that the training error after 100 iterations is decreased the fastest. To solve the Sinkhorn DRO model, we use V-SGD gradient estimator for newsvendor and portfolio optimization problems, and use V-MLMC estimator for multi-class classification problem. When solving the subproblem (3.7), the iteration is terminated when $\frac{\|\texttt{obj}_{\ell+1}-\texttt{obj}_\ell\|}{1+\|\texttt{obj}_\ell\|} \leq \texttt{1e-2}$, where $\texttt{obj}_\ell$ denotes the objective function obtained at the ℓ-th iteration. We also use 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 or optimization algorithms for solving baseline models.

B.2.1 Setup for Newsvendor Problem

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:

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

B.2.2 Setup for Mean-risk Portfolio Optimization

From [110, Eq. (27)] we can see that the 1-Wasserstein DRO formulation with radius ρ for the portfolio optimization problem becomes

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

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

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

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

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

Table B.1: Basic statistics of classification datasets

Dataset	# of Features	# of Classes	Training Size	Testing Size
MNIST	50	10	345	69655
IRIS	10	3	37	133
wine	13	3	44	134
vowel	30	11	594	396
vehicle	20	4	507	339
svmguide4	50	6	367	245

B.2.3 Setup for Linear Classification Incorportating Structural Information

In this example, we solve the SAA, KL-divergence DRO problem using stochastic gradient descent. Also, Wasserstein DRO models can be reformulated as

$$\min_{B,\lambda\geq 0} \max_{\|z_i\|_\infty \leq 1, i\in[n]} \lambda\rho + \frac{1}{N}\sum_{i=1}^{n} \big[h_B(z_i) - \lambda c(\hat{x}_i, z_i)\big]. \tag{B.1}$$

To approximately solve such a convex-non-concave problem, we implement a gradient descent ascent heuristic outlined in Algorithm 5. Finally, we report basic statistics of classification datasets in this example in Table B.1.

Algorithm 5 Heuristic Gradient Descent Ascent algorithm for solving (B.1). We use default values $\alpha = \texttt{5e-6}, m = 64, n_{\text{critic}} = 5$.

Require: Learning rate α, batch size m, number of inner iterations n_{critic}, initial guess (θ, λ).

while (θ, λ) not converged **do**
  for $t = 0, 1, \ldots, n_{\text{critic}}$ **do**
    Sample a subset of indices $\{n_i\}_{i=1}^m$ from $[n]$.
    Compute $g_{n_i} \leftarrow \frac{1}{N}\nabla_{z_{n_i}}\big[h_B(z_{n_i}) - \lambda c(\hat{z}_{n_i}, z_{n_i})\big]$ for $i \in [m]$.
    Update $z_{n_i} \leftarrow \text{Proj}_{\mathcal{Z}}\big[z_{n_i} + \alpha\text{RMSProp}(z_{n_i}, g_{n_i})\big]$.
  end for
  Sample a subset of indices $\{n_i\}_{i=1}^m$ from $[n]$.
  Compute $g_\lambda \leftarrow \nabla_\lambda\left\{\lambda\rho + \frac{1}{m}\sum_{i=1}^m \big[h_B(z_{n_i}) - \lambda c(\hat{z}_{n_i}, z_{n_i})\big]\right\}$.
  Compute $g_B \leftarrow \nabla_B\big\{\frac{1}{m}\sum_{i=1}^m h_B(z_{n_i})\big\}$.
  Update $\lambda \leftarrow (\lambda - \alpha\text{RMSProp}(\lambda, g_\lambda))_+$.
  Update $B \leftarrow B - \alpha\text{RMSProp}(B, g_B)$.
end while
Output (λ, B)**.**

B.3 Additional Validation Experiments

B.3.1 Comparison of Optimization Algorithms

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 $\widehat{P}$ as the empirical one based on samples $\{(a_i, b_i)\}_{i=1}^n$. As a consequence, the inner objective function in (3.7) 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 $\Theta = \{\theta : \ \|\theta\|_2^2 \leq 0.999 \cdot \frac{\lambda}{2}\}$. For the data generation procedure, we first generate a ground truth predictor $\theta^* \sim \mathcal{N}(0, 100 \cdot I_d)$. We then generate inputs $a_i \sim \mathcal{N}(0, I_d)$ and $b_i = \langle a_i, \theta^* \rangle + \zeta_i$, where the noise ζ_i follows the Gaussian distribution such that the response b_i has the signal-to-noise ratio 0.2. In this experiment, we take the training sample size $n = 500$ and data dimension $d = 50$.

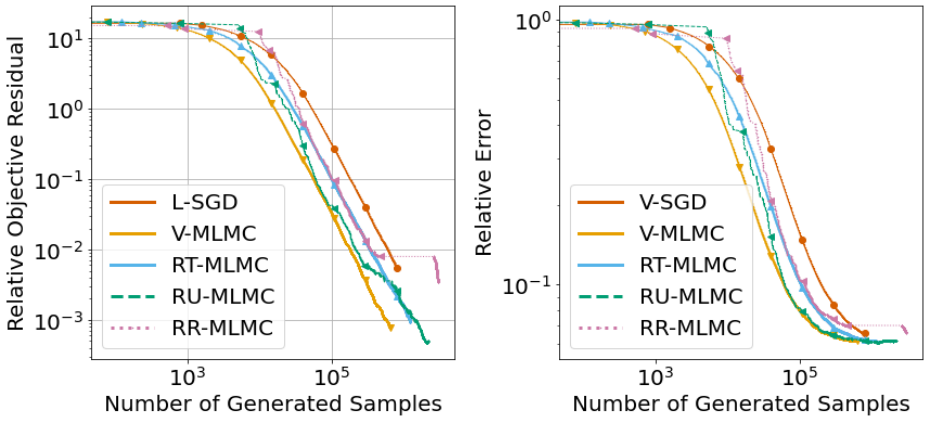

Figure B.1: Comparison results of V-SGD, V-MLMC, RT-MLMC, RU-MLMC, and RR-MLMC on robust linear regression in terms of relative objective residual (left plot) and relative prediction error (right plot).

The quality of proposed gradient estimators is examined in a single BSMD step with specified hyper-parameters $(\lambda, \epsilon) = (5 \cdot 10^3, 10^{-2})$. For baseline comparison, we also study the performance of two unbiased gradient estimators in literature [82]. However, the

variance of them are unbounded, so that there is no convergence analysis for those two methods.

RU-MLMC Estimator: at point θ, first sample a *random level* ι following distribution $Q_{\text{RU}} = \{q_\ell\}_{\ell=0}^{\infty}$ with $\mathbb{P}(\iota = \ell) = q_\ell$, then construct

$$v^{\text{RU-MLMC}}(\theta) := \frac{1}{q_\iota} G^\iota(\theta, \zeta^\iota). \tag{B.2}$$

RR-MLMC Estimator: at point θ, first sample a *random level* L following distribution $Q_{\text{RR}} = \{q_\ell\}_{\ell=0}^{\infty}$ with $\mathbb{P}(L = \ell) = q_\ell$, then construct

$$v^{\text{RR-MLMC}}(\theta) := \sum_{\ell=0}^{L} p_\ell G^\ell(x, \zeta^\ell), \tag{B.3}$$

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

To quantify the performance of a given solution θ, we denote the relative objective residual as $\frac{F(\theta)-F(\theta^*)}{1+|F(\theta^*)|}$, where θ^* is the optimal solution of F. We compute this optimal solution by optimizing the closed-form expression of F directly via OSMM software [145]. We also denote the relative error of θ as $\frac{\|\theta-\theta^*\|}{1+\|\theta^*\|}$, where θ^* is the ground truth optimal solution when the data distribution is known. Fig. B.1 reports the relative objective residual and relative error in terms of the number of generated samples. From the plot, we can see the V-SGD scheme does not have competitive performance, which is consistent with our theoretical analysis that V-SGD has the worst complexity order. In contrast, using other MLMC methods, we can obtain optimal solutions with small sample complexity. Although the RU-MLMC and RR-MLMC schemes have competitive performance, we can see there exist some oscillations during the optimization procedure. One possible explanation is that the variance values of those gradient estimators are unbounded, making those two approaches unstable.

B.3.2 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 3.5. Hence, the smaller the metric is, the better the given decision has.

Fig. B.2 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. B.3 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.

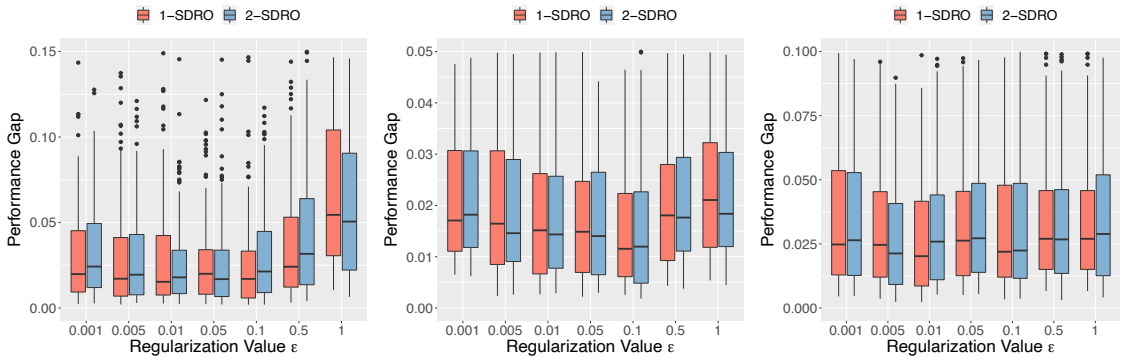

Figure B.2: Performance of Sinkhorn DRO models for newsvendor problem versus different choices of regularization values ϵ. For figures from left to right, we specify the data distribution as exponential distribution, gamma distribution, and equiprobable mixture of two truncated normal distributions, respectively.

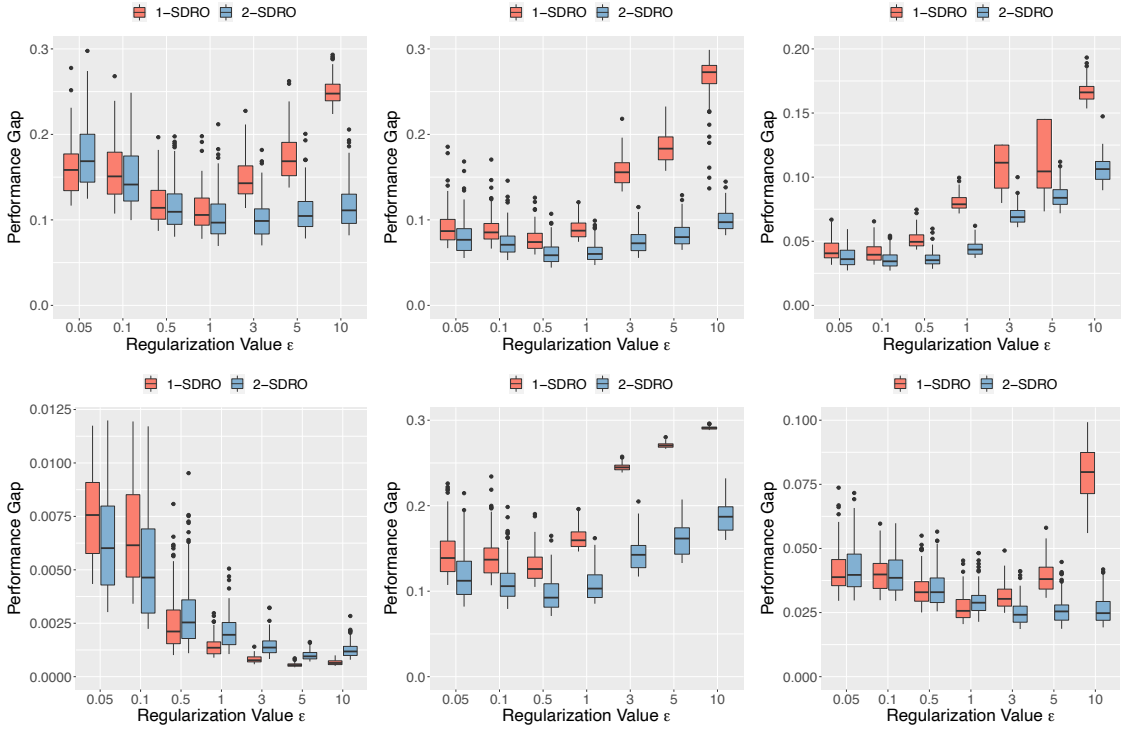

Figure B.3: Performance of Sinkhorn DRO models for portfolio problem versus different choices of regularization values ϵ. For those fix figures from left to right, from top to bottom, we specify the problem parameters (sample size n and data dimension d) as $(30, 30), (100, 30), (400, 30), (100, 5), (100, 20), (100, 100)$, respectively.

B.4 Proofs of Technical Results in Section 3.3.2

Proof. Proof of Remark 7 We can reformulate the dual objective function as

$$v(\lambda;\epsilon) = \lambda\rho + \lambda\epsilon \int \log\left(\int \exp\left(\frac{f(z)-\lambda c(x,z)}{\lambda\epsilon}\right) \mathrm{d}\nu(z)\right) \mathrm{d}\widehat{P}(x).$$

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

$$\begin{aligned}
&\lim_{\epsilon\to 0} \lambda\epsilon \int \log\left(\int \exp\left(\frac{f(z)-\lambda c(x,z)}{\lambda\epsilon}\right) \mathrm{d}\nu(z)\right) \mathrm{d}\widehat{P}(x) \\
&= \int \lim_{\beta\to\infty} \frac{\lambda}{\beta} \log\left(\int \exp\left(\frac{\big[f(z)-\lambda c(x,z)\big]\beta}{\lambda}\right) \mathrm{d}\nu(z)\right) \mathrm{d}\widehat{P}(x) \\
&= \int \lim_{\beta\to\infty} \lambda \nabla \log\left(\int \exp\left(\frac{\big[f(z)-\lambda c(x,z)\big]\beta}{\lambda}\right) \mathrm{d}\nu(z)\right) \mathrm{d}\widehat{P}(x) \\
&= \int \left[\lim_{\beta\to\infty} \frac{\int \exp\left(\frac{\left[f(z)-\lambda c(x,z)\right]\beta}{\lambda}\right) \big[f(z)-\lambda c(x,z)\big] \mathrm{d}\nu(y)}{\int \exp\left(\frac{\left[f(z)-\lambda c(x,z)\right]\beta}{\lambda}\right) \mathrm{d}\nu(y)}\right] \mathrm{d}\widehat{P}(x) \\
&= \int \operatorname*{ess\,sup}_{\nu} \big[f(\cdot)-\lambda c(x,\cdot)\big] \, \mathrm{d}\widehat{P}(x).
\end{aligned}$$

Particularly, when $\mathrm{supp}(\nu) = \mathcal{Z}$, it holds that

$$\operatorname*{ess\,sup}_{\nu} \big[f(\cdot)-\lambda c(x,\cdot)\big] = \sup_{z} \ \big[f(z)-\lambda c(x,z)\big],$$

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

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

$$V_{\mathrm{D}} = \inf_{\lambda\geq 0} \ \left\{\lambda\overline{\rho} + \mathbb{E}_{(a,b)\sim\widehat{P}}\left[\lambda\epsilon \log\left(\mathbb{E}_{a'\sim\mathcal{N}(a,\eta I_d)} \ \exp\left(\frac{(\theta^{\mathrm{T}}a'-b)^2}{\lambda\epsilon}\right)\right)\right]\right\}. \tag{B.4}$$

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

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

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

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

Substituting this expression into (B.4) gives the desired result.

□ □

Proof. Proof of Corollary 2 We now introduce the epi-graphical variables $s_i, i = 1, \ldots, n$

to reformulate V_{D} as

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

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

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

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. □ □

B.5 Proofs of Technical Results in Section 3.3.3

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

Lemma 18. *For fixed τ and a reference probability distribution $\mathbb{Q} \in \mathcal{P}(\mathcal{Z})$, consider the optimization problem*

$$v(\tau) = \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{\mathbb{P}} \left[f(z) - \tau \log \left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(z) \right) \right] \right\}. \tag{B.5}$$

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

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

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

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

it holds that

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

and $\lim_{\tau \downarrow 0} v(\tau) = v(0)$. The optimal solution in (B.5) *has the expression*

$$\mathrm{d}\mathbb{P}(z) = \frac{e^{f(z)/\tau}}{\int e^{f(u)/\tau} \, \mathrm{d}\mathbb{Q}(u)} \, \mathrm{d}\mathbb{Q}(z).$$

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

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

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

Proof. Proof of Lemma 18 We reformulate $v(\tau)$ based on the importance sampling trick:

$$v(\tau) = \sup_{L: \, L \geq 0} \left\{ \int \left[f(z) L(z) - \tau L(z) \log L(z) \right] \mathrm{d}\mathbb{Q}(z) : \ \int L(z) \, \mathrm{d}\mathbb{Q}(z) = 1 \right\}.$$

Then the remaining part follows the discussion in [84, Section 2.1]. □ □

Proof. Proof of Lemma 3 Based on Definition 6 of Sinkhorn distance, we reformulate V as

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

By the disintegration theorem [31] we represent the joint distribution γ such that $\mathrm{d}\gamma(x,z) = \mathrm{d}\widehat{P}(x)\,\mathrm{d}\gamma_x(z)$ holds for any (x,z), where γ_x is the conditional distribution of γ given the first marginal of γ equals x. Thereby the constraint is equivalent to

$$\int \mathbb{E}_{\gamma_x}\left[c(x,z) + \epsilon \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)} \right) \right] \mathrm{d}\widehat{P}(x) \leq \rho, \quad \gamma_x \in \mathcal{P}(\mathcal{Z}), x \in \mathrm{supp}(\widehat{P}).$$

We remark that any feasible solution γ satisfies that $\gamma \ll \widehat{P} \otimes \nu$ and hence $\gamma_x \ll \nu$. Consequently the term $\log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)}\right)$ is well-defined. 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

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

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

$$\epsilon \int \mathbb{E}_{\gamma_x}\left[\log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)} \right) \right] \mathrm{d}\widehat{P}(x) \leq \overline{\rho}, \quad \gamma_x \in \mathcal{P}(\mathcal{Z}), x \in \mathrm{supp}(\widehat{P}).$$

Similarly, the objective function of (`Sinkhorn Primal`) can be written as $\int \mathbb{E}_{\gamma_x}[f(z)]\,\mathrm{d}\widehat{P}(x)$. Consequently, the primal problem (`Sinkhorn Primal`) can be reformulated as a generalized KL-divergence DRO problem

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

□ □

Proof. Proof of Lemma 4 Recall from Remark 8 that the primal problem V can be reformulated as

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

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

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

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

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

Since the optimization over $\gamma_x, \forall x$ is separable for each x, by defining

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

and swap the supremum and the integration, we obtain

$$V \leq \inf_{\lambda \geq 0} \ \left\{ \lambda \overline{\rho} + \int v_x(\lambda) \, \mathrm{d}\widehat{P}(x) \right\}. \tag{B.6}$$

When there exists $\lambda > 0$ such that Condition 1 holds, by leveraging a well-known reformulation on entropy regularized linear optimization in Lemma 18, we can see that almost surely,

$$v_x(\lambda) = \lambda \epsilon \log \left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}} \left[e^{f(z)/(\lambda \epsilon)} \right] \right) < \infty.$$

Substituting this expression into (B.6) implies that $V \leq V_{\mathrm{D}} < \infty$. Suppose on the contrary that for any $\lambda > 0$,

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

then intermediately we obtain $V \leq V_{\mathrm{D}} = \infty$, and the weak duality still holds.

□ □

Proof. Proof of Lemma 5 We first show that $\lambda^* < \infty$. Denote by $v(\lambda)$ the objective function for the dual problem, then

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

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

$$\begin{aligned}
&\lim_{\lambda\to\infty} \lambda\epsilon \int \log\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\right) \mathrm{d}\widehat{P}(x) \\
&= \int \lim_{\beta\to 0} \frac{\epsilon}{\beta} \log\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}\left[e^{\beta f(z)/\epsilon}\right]\right) \mathrm{d}\widehat{P}(x) \\
&= \int \lim_{\beta\to 0} \epsilon\nabla_\beta \log\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}\left[e^{\beta f(z)/\epsilon}\right]\right) \mathrm{d}\widehat{P}(x) \\
&= \int \lim_{\beta\to 0} \epsilon \frac{1}{\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}\left[e^{\beta f(z)/\epsilon}\right]} \left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}\left[\frac{f(z)}{\epsilon} e^{(\beta f(z))/\epsilon}\right]\right) \mathrm{d}\widehat{P}(x) \\
&= \int \mathbb{E}_{\mathbb{Q}_{x,\epsilon}}[f(z)] \, \mathrm{d}\widehat{P}(x),
\end{aligned}$$

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 $v(\lambda) < \infty$, which guarantees the existence of the dual minimizer. Hence $\lambda^* < \infty$, which implies that either $\lambda^* = 0$ or λ^* satisfies Condition 1. □ □

Proof. Proof of Lemma 6 Suppose the dual minimizer $\lambda^* = 0$, then taking the limit of the dual objective function gives

$$\lim_{\lambda\to 0} v(\lambda) = \int H^u(x) \, \mathrm{d}\widehat{P}(x) < \infty,$$

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 $x \in \text{supp}(\widehat{P})$: for any t so that $\mathbb{Q}_{x,\epsilon}\{f(z) > t\} = 0$, we have that

$$\int 1\{f(z) > t\} e^{-c(x,z)/\epsilon} \, \mathrm{d}\nu(z) = 0,$$

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

$$\int 1\{f(z) > t\} \, \mathrm{d}\nu(z) = 0.$$

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

$$0 \leq \int 1\{f(z) > t\} e^{-c(x,z)/\epsilon} \, \mathrm{d}\nu(z) \leq \int 1\{f(z) > t\} \, \mathrm{d}\nu(z) = 0,$$

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

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

Then we show that almost surely for all x,

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

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

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

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

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

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

$$\begin{aligned} v(\lambda) &= \lambda\overline{\rho} + \int v_x(\lambda)\,\mathrm{d}\widehat{P}(x) \\ &\leq H^u + \lambda\overline{\rho} + \lambda\epsilon \int_D \log\left(\mathfrak{h}_x + e^{-H^{\mathrm{gap}}(x)/(\lambda\epsilon)}\mathfrak{h}_x^c\right)\,\mathrm{d}\widehat{P}(x). \end{aligned}$$

We can see that

$$\lim_{\lambda\to 0}\ \lambda\overline{\rho} + \lambda\epsilon \int_D \log\left(\mathfrak{h}_x + e^{-H^{\mathrm{gap}}(x)/(\lambda\epsilon)}\mathfrak{h}_x^c\right)\,\mathrm{d}\widehat{P}(x) = 0,$$

and

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

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

$$v(\overline{\lambda}) \leq H^u + \overline{\lambda}\overline{\rho} + \overline{\lambda}\epsilon \int_D \log\left(\mathfrak{h}_x + e^{-H^{\mathrm{gap}}(x)/(\overline{\lambda}\epsilon)}\mathfrak{h}_x^c\right)\,\mathrm{d}\widehat{P}(x) < v(0),$$

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

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

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

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

The gradient of $v(\lambda)$ becomes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

minimizer $\lambda^* = 0$.

□ □

Proof. Proof of Lemma 7. Since $\lambda^* > 0$, based on the optimality condition of the dual problem, we have that

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

Or equivalently, we have that

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

Re-arranging the term completes the proof.

□ □

Proof. Proof of Theorem 5. The feasibility result in Theorem 5(I) can be easily shown by considering the reformulation of V in Lemma 3 and the non-negativity of KL-divergence. When $\overline{\rho} = 0$, one can see that

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

Therefore, the strong duality result holds in this case. The proof for $\overline{\rho} > 0$ can be found in the main context. It remains to show the second part of Theorem 5(III). We consider a sequence of real numbers $\{R_j\}_j$ such that $R_j \to \infty$ and take the objective function $f_j(z) = f(z)1\{z \leq R_j\}$. Hence, there exists $\lambda > 0$ satisfying $\Pr_{x\sim\widehat{P}}\left\{x:\ \mathbb{E}_{\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right] = \infty\right\} = 0$. According to the necessary condition in Lemma 6, 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 5(III) to show that for sufficiently large j, it holds that

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

Taking $j \to \infty$ both sides implies that $V = \infty$, which completes the proof.

□ □

Proof. Proof of Corollary 1 According to the definition of Sinkhorn distance, we first reformulate V_λ as

$$\begin{aligned} V_\lambda &= \sup_{\mathbb{P},\gamma\in\Gamma(\widehat{P},\mathbb{P})} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[f(z)] - \lambda \mathbb{E}_\gamma \left[c(x,z) + \epsilon \log\left(\frac{\mathrm{d}\gamma(x,z)}{\mathrm{d}\widehat{P}(x)\,\mathrm{d}\nu(z)} \right) \right] \right\} \\ &= \sup_{\gamma_x\in\mathcal{P}(\Omega), x\in\mathcal{Z}} \left\{ \int \mathbb{E}_{\gamma_x} \left[f(z) - \lambda\epsilon \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)} \right) \right] \mathrm{d}\widehat{P}(x) \right\} \\ &\qquad + \lambda\epsilon \int \log\left(\int e^{-c(x,u)/\epsilon}\,\mathrm{d}\nu(u) \right) \mathrm{d}\widehat{P}(x), \end{aligned}$$

where the second relation is by decomposing γ with $\gamma(x,z) = \widehat{P}(x)\otimes\gamma_x(z)$. By the principal of interchangability [143], it holds that

$$\begin{aligned} V_\lambda &= \int \sup_{\gamma_x\in\mathcal{P}(\Omega)} \mathbb{E}_{\gamma_x} \left[f(z) - \lambda\epsilon \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)} \right) \right] \mathrm{d}\widehat{P}(x) + C \\ &= \lambda\epsilon \int \log\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}[e^{f(z)/(\lambda\epsilon)}] \right) \mathrm{d}\widehat{P}(x) + C, \end{aligned}$$

where the last relation holds by applying Lemma 18. □ □

B.6 Preliminaries on Stochastic Mirror Descent (SMD)

In this section, we provide some preliminaries on SMD that can be useful for proving Theorem 6. Consider the minimization of the objective function $F(\theta) = \mathbb{E}[f_\theta(z)]$ 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 4 (Stochastic Oracles of Gradient Estimate)**.** (I) *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 $\mathbb{E}[G(\theta, \xi)] \in \partial F(\theta)$, where $\partial F(\theta)$ is the subdifferential set of $F(\cdot)$ at θ. Also, suppose there exists a constant $M_* > 0$ so that*

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

(II) *Assume the objective function $F(\theta)$ is S-smooth, and we have the stochastic oracle such that for given θ we can generate a stochastic vector $G(\theta, \xi)$ such that $\mathbb{E}[G(\theta, \xi)] = \nabla F(\theta)$. Also, suppose there exists $\sigma > 0$ so that*

$$\mathbb{V}\mathrm{ar}\big[G(\theta, \xi)\big] \leq \sigma^2, \quad \forall \theta \in \Theta.$$

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

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

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 [69, 114, 129, 143]. For the sake of completeness, we provide the proof for the case where the loss function $F(\theta)$ is smooth in θ.

Lemma 19 (SMD for Nonsmooth Convex Optimization)**.** *Under Assumption 4*(I)*, 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

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

Proof. Proof of Lemma 19 The proof follows from [143, Section 8.2.3]. □ □

Lemma 20 (SMD for Smooth Convex Optimization)**.** *Assume the loss function $F(\theta)$ is convex in θ and Assumption 4*(II) *holds. Let the estimation of optimal solution at the iteration j be $\tilde{\theta}_{1:j} = \frac{1}{j}\sum_{t=1}^{j}\theta_t$. Also, suppose the norm function $\|\cdot\|$ satisfies that the dual norm $\|\cdot\|_* \leq c\|\cdot\|_2$. When taking constant step size $\gamma \in (0, \kappa/(2Sc^2)]$, it holds that*

$$\mathbb{E}\big[F(\tilde{\theta}_{1:T}) - F(\theta^*)\big] \leq \frac{\gamma Sc^2\sigma^2}{\kappa} + \frac{2V(\theta_1, \theta^*)}{\gamma T}.$$

Proof. Proof of Lemma 20 For each $u \in \Theta$ and $\theta \in \Theta^*$, and $y \in \mathbb{R}^d$, based on [143, Lemma 8.3], one has that

$$V\big(P_\theta(y), u\big) \leq V(\theta, u) + \langle y, u - \theta\rangle + \frac{1}{2\kappa}\|y\|_*^2. \tag{B.7}$$

Based on this identity with $\theta := \theta_t$, $y := \gamma G(\theta_t, \xi^t)$, and $u := \theta^*$, we obtain

$$V\big(\theta_{t+1}, \theta^*\big) \leq V(\theta_t, \theta^*) + \gamma\langle G(\theta_t, \xi^t), \theta^* - \theta_t\rangle + \frac{\gamma^2}{2\kappa}\|G(\theta_t, \xi^t)\|_*^2. \tag{B.8}$$

As a consequence, we have the relation

$$\mathbb{E}\langle G(\theta_t, \xi^t), \theta_t - \theta^*\rangle \leq \frac{\mathbb{E}V\big(\theta_t, \theta^*\big) - \mathbb{E}V(\theta_{t+1}, \theta^*)}{\gamma} + \frac{\gamma}{2\kappa}\mathbb{E}\|G(\theta_t, \xi^t)\|_*^2.$$

On the other hand, conditioned on x_t, we have that

$$-\mathbb{E}\langle G(\theta_t, \xi^t), \theta_t - \theta^* \rangle = -\langle \nabla F(\theta_t), \theta_t - \theta^* \rangle \leq F(\theta^*) - F(\theta_t).$$

Based on those two relations above, we obtain that conditioned on x_t,

$$\begin{aligned} F(\theta_t) - F(\theta^*) &\leq \frac{\mathbb{E}V\big(\theta_t, \theta^*\big) - \mathbb{E}V(\theta_{t+1}, \theta^*)}{\gamma} + \frac{\gamma}{2\kappa}\mathbb{E}\|G(\theta_t, \xi^t)\|_*^2 \\ &\leq \frac{\mathbb{E}V\big(\theta_t, \theta^*\big) - \mathbb{E}V(\theta_{t+1}, \theta^*)}{\gamma} + \frac{\gamma c^2}{2\kappa}\mathbb{V}\mathrm{ar}\big(G(\theta_t, \xi^t)\big) + \frac{\gamma c^2}{2\kappa}\|\nabla F(\theta_t) - \nabla F(\theta^*)\|_2^2 \\ &\leq \frac{\mathbb{E}V\big(\theta_t, \theta^*\big) - \mathbb{E}V(\theta_{t+1}, \theta^*)}{\gamma} + \frac{\gamma c^2}{2\kappa}\mathbb{V}\mathrm{ar}\big(G(\theta_t, \xi^t)\big) + \frac{\gamma S c^2}{\kappa}\big[F(\theta_t) - F(\theta^*)\big]. \end{aligned}$$

This implies that

$$\mathbb{E}\big[F(\theta_t) - F(\theta^*)\big] \leq \frac{1}{1 - \gamma S c^2/\kappa}\left[\frac{\mathbb{E}V\big(\theta_t, \theta^*\big) - \mathbb{E}V(\theta_{t+1}, \theta^*)}{\gamma} + \frac{\gamma S c^2}{2\kappa}\mathbb{V}\mathrm{ar}\big(G(\theta_t, \xi^t)\big)\right].$$

When taking the step size $\gamma \in (0, \kappa/(2Sc^2)]$, it holds that for any t,

$$\mathbb{E}\big[F(\theta_t) - F(\theta^*)\big] \leq \frac{2\big\{\mathbb{E}V\big(\theta_t, \theta^*\big) - \mathbb{E}V(\theta_{t+1}, \theta^*)\big\}}{\gamma} + \frac{\gamma S c^2 \sigma^2}{\kappa}. \tag{B.9}$$

Finally, the estimate of optimal solution $\tilde{\theta}_{1:T}$ satisfies that

$$\begin{aligned} \mathbb{E}\big[F(\tilde{\theta}_{1:T}) - F(\theta^*)\big] &= \mathbb{E}\left[F\left(\frac{1}{T}\sum_{t=1}^{T}\theta_t\right) - F(\theta^*)\right] \\ &\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}F\left(\theta_t\right) - F(\theta^*)\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[F(\theta_t) - F(\theta^*)\big] \\ &\leq \frac{\gamma S c^2 \sigma^2}{\kappa} + \frac{2V(\theta_1, \theta^*)}{\gamma T}, \end{aligned}$$

where the first inequality is based on the Jensen's inequality and the convexity of $F(\theta)$, and the second inequality is by the relation (B.9). □ □

B.7 Proofs of Technical Results in Section 3.4.2

Remark 15 (Computation Costs of Gradient Estimators)**.** *The cost for generating V-SGD gradient estimator $v^{\text{V-SGD}}(\theta)$ is of $O(n_L^{\circ}2^L)$. For V-MLMC scheme, we set the number of inner approximation level $n_\ell = \lceil 2^{-\ell}N \rceil$. The cost for generating V-MLMC gradient estimator $v^{\text{V-MLMC}}(\theta)$ is of $O(NL + 2^L)$. For RT-MLMC scheme, we take the probability $q_\ell \propto 2^{-\ell}$. The cost for generating RT-MLMC gradient estimator $v^{\text{RT-MLMC}}(\theta)$ is of $O(n_L^{\circ}L)$.*

Proof. Proof of Remark 15 Since V-SGD 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 cost of V-SGD estimator is of $O(n_L^{\circ}2^L)$.

Since for fixed ℓ, the cost of generating a single $G^\ell(\theta, \zeta_i^\ell)$ requires generating the random sampling parameters $\{z_j^\ell\}_{j\in[2^L]}$ of size 2^ℓ, the cost of V-MLMC estimator can be bounded as

$$O\left(\sum_{\ell=0}^{L} n_\ell^{\circ}2^\ell\right) = O\left(\sum_{\ell=0}^{L}\lceil 2^{-\ell}N\rceil 2^\ell\right) = O\left(\sum_{\ell=0}^{L}(2^{-\ell}N+1)2^\ell\right) = O\left(NL+2^L\right).$$

The cost of RT-MLMC estimator can be bounded as

$$O\left(n_L^{\circ}\sum_{\ell=0}^{L} q_\ell 2^\ell\right) = O\left(n_L^{\circ}\sum_{\ell=0}^{L}\frac{2^{-\ell}}{C}2^\ell\right) = O(n_L^{\circ}L/C) = O(n_L^{\circ}L),$$

where the constant $C = \sum_{\ell=0}^{L} 2^{-\ell} = O(1)$.

□ □

B.7.1 Proof of Theorem 6(I)

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 4 (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*

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

Proof. Proof of Proposition 4 Note that we can establish the following error bound:

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

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 19, if we take the step size $\gamma = \sqrt{\frac{2\kappa V(\theta_1, \overline{\theta}^*)}{TM_*^2}}$, then it holds that

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

□ □

Now we show complexity result for the V-SGD scheme. Without loss of generality, we take the batch size $n_L^{\circ} = 1$. According to Proposition 4, parameters for V-SGD scheme satisfy

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

To obtain δ-optimal solution, we set

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

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

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

B.7.2 Proof of Theorem 6(II)

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

$$\mathbb{E}[v(\theta)] = \nabla \overline{F}(\theta), \quad \mathbb{V}\text{ar}\left[v(\theta)\right] \leq \sigma^2.$$

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 5 (BSMD for Smooth Convex Optimization)**.** *Assume that the approximation function $\overline{F}$ is S-smooth. When taking constant step size $\gamma = \sqrt{\frac{2\kappa V(\theta_1, \overline{\theta}^*)}{Sc^2\sigma^2 T}}$, it holds that*

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

Specially, we can show the approximation function $F^\ell(\theta)$ defined in (3.8) is indeed smooth, and therefore Proposition 5 can be used to prove Theorem 6(II).

Lemma 21. *Under Assumption 3*(II)*, 3*(III)*, and 3*(IV)*, the functions F and F^ℓ are S_F-*

smooth with

$$S_F := (S_f + L_f^2/(\lambda\epsilon))\exp(B/(\lambda\epsilon)) + L_f^2/(\lambda\epsilon)\exp(2B/(\lambda\epsilon)).$$

The proof of Lemma 21 follows the similar argument as in [82, Proposition 4.1]. We provide a full proof here for the sake of completeness.

Proof. Proof of Lemma 21 Observe that

$$\begin{aligned}
&\|\nabla F(\theta_1) - \nabla F(\theta_2)\|_2 \\
\leq&(\lambda\epsilon)^{-1}\mathbb{E}_{\widehat{P}}\Big\|\phi'\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_1}(z)/(\lambda\epsilon)}\right)\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_1}(z)/(\lambda\epsilon)}\nabla f_{\theta_1}(z) \\
&\qquad\qquad - \phi'\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_2}(z)/(\lambda\epsilon)}\right)\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_2}(z)/(\lambda\epsilon)}\nabla f_{\theta_2}(z)\Big\|_2 \\
\leq&(\lambda\epsilon)^{-1}\mathbb{E}_{\widehat{P}}\left\|\phi'\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_1}(z)/(\lambda\epsilon)}\right)\left[\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_1}(z)/(\lambda\epsilon)}\nabla f_{\theta_1}(z) - \mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_2}(z)/(\lambda\epsilon)}\nabla f_{\theta_2}(z)\right]\right\|_2 \\
&+ (\lambda\epsilon)^{-1}\mathbb{E}_{\widehat{P}}\left\|\left[\phi'\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_1}(z)/(\lambda\epsilon)}\right) - \phi'\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_2}(z)/(\lambda\epsilon)}\right)\right]\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_2}(z)/(\lambda\epsilon)}\nabla f_{\theta_2}(z)\right\|_2
\end{aligned}$$

The first term on the RHS can be bounded as

$$\begin{aligned}
&\mathbb{E}_{\widehat{P}}\left\|\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_1}(z)/(\lambda\epsilon)}\nabla f_{\theta_1}(z) - \mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_2}(z)/(\lambda\epsilon)}\nabla f_{\theta_2}(z)\right\|_2 \\
\leq&\mathbb{E}_{\widehat{P}}\left\|\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_1}(z)/(\lambda\epsilon)}\left[\nabla f_{\theta_1}(z) - \nabla f_{\theta_2}(z)\right]\right\|_2 \\
&\qquad + \mathbb{E}_{\widehat{P}}\left\|\left[\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_1}(z)/(\lambda\epsilon)} - \mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_2}(z)/(\lambda\epsilon)}\right]\nabla f_{\theta_2}(z)\right\|_2 \\
\leq&e^{B/(\lambda\epsilon)}S_f\|\theta_1 - \theta_2\| + L_f^2/(\lambda\epsilon)e^{B/(\lambda\epsilon)}\|\theta_1 - \theta_2\|_2,
\end{aligned}$$

where the last inequality is because $f_\theta(z)$ is bounded by B, L_f-Lipschitz, and S_f-smooth. The second term on the RHS can be bounded as

$$\begin{aligned}
&L_f/(\lambda\epsilon)e^{B/(\lambda\epsilon)}\mathbb{E}_{\widehat{P}}\left\|\phi'\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_1}(z)/(\lambda\epsilon)}\right) - \phi'\left(\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_2}(z)/(\lambda\epsilon)}\right)\right\|_2 \\
\leq&L_f e^{B/(\lambda\epsilon)}\mathbb{E}_{\widehat{P}}\left\|\mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_1}(z)/(\lambda\epsilon)} - \mathbb{E}_{\mathbb{Q}_{x,\epsilon}}e^{f_{\theta_2}(z)/(\lambda\epsilon)}\right\|_2 \\
\leq&L_f^2/(\lambda\epsilon)e^{2B/(\lambda\epsilon)}\|\theta_1 - \theta_2\|_2.
\end{aligned}$$

Hence, we conclude that the function F is S_F-smooth, where

$$S_F = L_f^2/(\lambda\epsilon)e^{2B/(\lambda\epsilon)} + e^{B/(\lambda\epsilon)}S_f + L_f^2/(\lambda\epsilon)e^{B/(\lambda\epsilon)}.$$

The smoothness of the function F^ℓ can be finished in a similar manner. □ □

Now we are ready to show complexity results for V-SGD, V-MLMC, and RT-MLMC schemes.

V-SGD Without loss of generality, we take the batch size $n_L^{\mathrm{o}} = 1$. According to Proposition 5, parameters for V-SGD scheme satisfy

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

To obtain δ-optimal solution, we set

$$2\Delta_F \leq \frac{\delta}{2}, \quad \sqrt{\frac{2S_F c^2 \sigma^2 V(\theta_1, \overline{\theta}^*)}{\kappa T}} \leq \frac{\delta}{2}.$$

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

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

V-MLMC When taking the inner approximation sample size $n_\ell = \lceil 2^{-\ell} N \rceil$ for some $N > 0$, it holds that

$$\Delta_F := \lambda\epsilon \exp\left(2B/(\lambda\epsilon)\right) \cdot 2^{-(L+1)}, \quad \sigma^2 := L_f^2 \exp\left(4B/(\lambda\epsilon)\right)(L+1)N^{-1}.$$

To obtain δ-optimal solution, we set

$$2\Delta_F \leq \frac{\delta}{2}, \quad \frac{2S_F c^2 V(\theta_1, \overline{\theta}^*)}{\kappa T_{\mathrm{in}}} \leq \frac{\delta}{2}, \quad \sigma^2 \leq \frac{\delta}{2}.$$

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

ments:

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

RT-MLMC Without loss of generality, we take the batch size $n_L^{\text{o}} = 1$. When taking the probability $q_\ell \propto 2^{-\ell}$, it holds that

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

To obtain δ-optimal solution, we set

$$2\Delta_F \leq \frac{\delta}{2}, \quad \frac{2S_F c^2 \sigma^2 V(\theta_1, \overline{\theta}^*)}{\kappa T_{\text{in}}} \leq \frac{\delta^2}{4}.$$

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

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

B.7.3 Sampling Algorithm in Remark 10

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:

$$X_{k+1} = X_k - \gamma \nabla V_\epsilon(X_k) + \sqrt{2\gamma} Z_{k+1}, \quad \text{where } Z_{k+1} \sim \mathcal{N}(0, I_d). \tag{B.10}$$

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

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

Hence, the iteration (B.10) returns a distribution that is τ-close to $\mathbb{Q}_\epsilon$ in terms of KL-divergence distance within $O(1/\tau)$ iterations.

B.7.4 Proof of Remark 11

If employing the BSAA technique, the estimation of optimal solution of (3.7) is given by the optimal value of the following problem, where the objective function is a biased estimate of the objective in (3.7):

$$\min_{\theta \in \Theta} \left\{ \hat{F}_{n,m}(\theta) := \frac{\lambda\epsilon}{n} \sum_{i=1}^{n} \log \left(\frac{1}{m} \sum_{j=1}^{m} e^{f_\theta(z_{i,j})/(\lambda\epsilon)} \right) \right\}. \tag{B.11}$$

Here $\{x_i\}_{i=1}^n$ are samples i.i.d. generated from $\widehat{P}$, 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 [81, Corollary 4.2], we present the following sample complexity analysis of BSAA problem.

Proposition 6 (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 $|f_{\theta_1}(z) - f_{\theta_2}(z)| \leq L_f \|\theta_1 - \theta_2\|_2$.*

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

Suppose we specify parameters for (B.11) *as*

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

then with probability at least $1-\alpha$*, the solution to the SAA problem* (B.11) *is an* δ*-optimal solution of* (3.7)*.*

The sample complexity of BSAA problem is of $\tilde{O}(\delta^{-3})$, which is much worse than $\tilde{O}(\delta^{-2})$, 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. Proof of Proposition 6 We first verify the technical assumption imposed in [81, 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

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

(c) The variance

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

(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 [81, Corollary 4.2], we know that to obtain δ-optimal solution with probability at least $1-\alpha$, sample sizes m, n need to satisfy

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

and

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

□ □

B.8 Proofs of Technical Results in Section 3.4.2

Here we provide bounds for two basic statistics: $\mathbb{V}\mathrm{ar}\left(a^\ell(\theta,\zeta^\ell)\right)$ and $\mathbb{V}\mathrm{ar}\left(A^\ell(\theta,\zeta^\ell)\right)$. First,

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

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

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

Lemma 22 (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 $\delta > 0$, there exists $\epsilon_1 > 0$ such that for any $\epsilon \in (0,\epsilon_1)$, it holds that*

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

We first present the complexity of estimating objective value of a feasible solution θ based on MLMC estimators.

Proposition 7 (Complexity of MLMC Objective Estimators)**.** *Assume that Assumption 3*(II) *holds, then with properly chosen hyper-parameters of objective estimators in* (3.13)*, the following results hold:*

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

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

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

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

Proof. Proof of Proposition 7 First, we pick L such that $|F^L(\theta)-F(\theta)|\leq\frac{\delta}{4}$, 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 $\mathbb{V}\mathrm{ar}(V(\theta)) < \infty$.

$$\left\{|F(\theta)-V(\theta)| > \frac{\delta}{2}\right\} \subseteq \left\{|F^L(\theta)-V(\theta)| > \frac{\delta}{4}\right\}. \tag{B.12}$$

Then by the relation (B.12) and Lemma 22, it holds that

$$\Pr\left\{|F(\theta)-V(\theta)| > \frac{\delta}{2}\right\} \leq \Pr\left\{|F^L(\theta)-V(\theta)| > \frac{\delta}{4}\right\} \leq 2\exp\left(-\frac{\delta^2}{16(\delta'+2)\mathbb{V}\mathrm{ar}(V(\theta))}\right). \tag{B.13}$$

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

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

The concentration for V-SGD scheme becomes

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

To make the desired coverage probability, we take

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

The concentration for V-MLMC scheme becomes

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

To make the desired coverage probability, we take

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

Similar to the analysis of V-MLMC scheme, we specify the following parameter to make RT-MLMC scheme satisfies the desired coverage probability:

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

□

Finally, we provide the proof for Theorem 7.

Proof. Proof of Theorem 7 When running the BSMD step, we obtain $\widehat{\theta}$ such that

$$\mathbb{E}\big[F(\widehat{\theta}) - F(\theta^*)\big] \leq \delta.$$

Based on Markov's inequality, it holds that

$$\Pr\Big\{F(\widehat{\theta}) - F(\theta^*) \leq 2\delta\Big\} \geq \frac{1}{2}.$$

When running the BSMD step for $m := \lceil \log_2 \frac{2}{\eta} \rceil$ times, it holds that

$$\Pr\Big\{\min_{i\in[m]} F(\widehat{\theta}_i) - F(\theta^*) \leq 2\delta\Big\} \geq 1 - \frac{1}{2^m} \geq 1 - \frac{\eta}{2}.$$

We specify the error probability $\alpha = \frac{\eta}{2m}$ in Proposition 7 when running the sampling step. Then with probability at least $1 - m\alpha = 1 - \frac{\eta}{2}$, it holds that

$$\max_{i\in[m]} \Big|F(\widehat{\theta}_i) - V(\widehat{\theta}_i)\Big| \leq \delta.$$

Combining those two relations, it holds that

$$\Pr\Big\{\Big|\min_{i\in[m]} V(\widehat{\theta}_i) - F(\theta^*)\Big| \leq 3\delta\Big\} \geq 1 - \frac{\eta}{2} - m\alpha = 1 - \eta.$$

The overall computation cost in Algorithm 4 is

$$m*\Big\{\text{Cost(estimating optimal solution of (3.7) once)}+\text{Cost(estimating objective value of (3.7) once)}\Big\}.$$

The memory cost is

$$\max\Big\{\text{Cost(estimating optimal solution of (3.7) once)}, \text{Cost(estimating objective value of (3.7) once)}\Big\}.$$

(I) When running the BSMD step with V-SGD scheme and estimating the objective values

using V-MLMC or RT-MLMC scheme, the computation cost becomes

$$m * O(\delta^{-3}) + m * \tilde{O}\left(\delta^{-2} \cdot \log \frac{m}{\eta}\right) = O\left(\delta^{-3} \cdot \text{polylog}\frac{1}{\eta}\right).$$

The memory cost becomes

$$\max\left\{O(\delta^{-1}), \tilde{O}\left(\delta^{-2} \cdot \log \frac{m}{\eta}\right)\right\} = \tilde{O}\left(\delta^{-2} \cdot \text{polylog}\frac{1}{\eta}\right).$$

(II) When additionally the smoothness assumption holds, if running the BSMD step and estimating the objective values using V-MLMC or RT-MLMC scheme, the complexity becomes

$$m * \tilde{O}(\delta^{-2}) + m * \tilde{O}\left(\delta^{-2} \cdot \log \frac{m}{\eta}\right) = \tilde{O}\left(\delta^{-2} \cdot \text{polylog}\frac{1}{\eta}\right).$$

The memory cost becomes

$$\max\left\{\tilde{O}(\delta^{-1}), \tilde{O}\left(\delta^{-2} \cdot \log \frac{m}{\eta}\right)\right\} = \tilde{O}\left(\delta^{-2} \cdot \text{polylog}\frac{1}{\eta}\right).$$

□

B.9 Proofs of Technical Results in Section 3.4.2

A key technique to show Theorem 8 is the following complexity result on bisection search with inexact oracles.

Lemma 23 (Complexity for Noisy Bisection [40, Lemma 33])**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a B-Lipschitz and convex function defined on the interval $[\ell, u]$, and $\mathcal{G} : \mathbb{R} \to \mathbb{R}$ be an oracle so that $|\mathcal{G}(y) - f(y)| \leq \tilde{\delta}$ for all y. With at most*

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

calls to $\mathcal{G}$, the algorithm `OneDimMinimier` *[40, Algorithm 8] outputs y' so that*

$$f(y') - \min_y \ f(y) \leq 4\tilde{\delta}.$$

Proof. Proof of Theorem 8 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

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

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

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

and

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

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

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

In summary, $F^*(\lambda)$ is a L_λ-Lipschitz and convex function defined on $[\lambda_\ell, \lambda_u]$. Applying Lemma 23 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

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

times. □

B.10 Proof of the Technical Result in Appendix B.1

We first present an useful technical lemma before showing Proposition 3.

Lemma 24. *Under the first condition of Proposition 3, for any $x \in \mathcal{Z}$, it holds that*

$$\int e^{-c(x,z)/\epsilon} \, \mathrm{d}\nu(z) \geq e^{-2^{p-1}c(x,\overline{x})/\epsilon} \int e^{-2^{p-1}c(\overline{x},z)/\epsilon} \, \mathrm{d}\nu(z).$$

Proof. Proof of Lemma 24 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

$$\int e^{-c(x,z)/\epsilon} \, \mathrm{d}\nu(z) \geq \exp\left(-2^{p-1}c(x,\overline{x})/\epsilon\right) \int e^{-2^{p-1}c(\overline{x},z)/\epsilon} \, \mathrm{d}\nu(z).$$

The proof is completed.

□ □

Proof. Proof of Proposition 3 One can see that for any $x \in \mathrm{supp}(\widehat{P})$, it holds that

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

where the first inequality is based on the lower bound in Lemma 24, 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 $x \in \text{supp}(\widehat{P})$, $c(x, \overline{x}) < \infty$. Moreover,

$$0 < \int e^{-2^{p-1}c(\overline{x},z)/\epsilon} \, \mathrm{d}\nu(z) \leq \int e^{-c(\overline{x},z)/\epsilon} \, \mathrm{d}\nu(z) < \infty,$$

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

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

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

□ □

BIBLIOGRAPHY

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

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

[3] Ahmed, M., Mahmood, A. N., and Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60:19–31.

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

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

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

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

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

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

[10] Baldassarre, L., Rosasco, L., Barla, A., and Verri, A. (2010). Vector field learning via spectral filtering. In *Machine Learning and Knowledge Discovery in Databases*, pages 56–71, Berlin, Heidelberg. Springer Berlin Heidelberg.

[11] Bayraksan, G. and Love, D. K. (2015). Data-driven stochastic programming using phi-divergences. In *The Operations Research Revolution*, pages 1–19. INFORMS.

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

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

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

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

[16] Binkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying MMD GANs. In *International Conference on Learning Representations*.

[17] Blair, C. (1985). Problem complexity and method efficiency in optimization (a. s. nemirovsky and d. b. yudin). *SIAM Review*, 27(2):264–265.

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

[19] Blanchet, J., Glynn, P. W., Yan, J., and Zhou, Z. (2019a). Multivariate distributionally robust convex regression under absolute error loss. In *Advances in Neural Information Processing Systems*, volume 32, pages 11817–11826.

[20] Blanchet, J. and 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.

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

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

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

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

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

[26] Boumal, N., Absil, P.-A., and Cartis, C. (2018). Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33.

[27] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[28] Brouard, C., d'Alché Buc, F., and Szafranski, M. (2011). Semi-supervised penalized output kernel regression for link prediction. In *28th International Conference on Machine Learning (ICML 2011)*, pages 593–600.

[29] Caponnetto, A., Micchelli, C. A., Pontil, M., and Ying, Y. (2008). Universal multi-task kernels. *Journal of Machine Learning Research*, 9(52):1615–1646.

[30] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3).

[31] Chang, J. T. and Pollard, D. (2001). Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317.

[32] Chen, R. and Paschalidis, I. C. (2019). Selecting optimal decisions via distributionally robust nearest-neighbor regression. In *Advances in Neural Information Processing Systems*.

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

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

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

[36] Cheng, X. and Xie, Y. (2021a). Kernel mmd two-sample tests for manifold data. *arXiv preprint arXiv:2105.03425*.

[37] Cheng, X. and Xie, Y. (2021b). Neural tangent kernel maximum mean discrepancy. In *Advances in Neural Information Processing Systems*, volume 34.

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

[39] Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 2606–2615.

[40] Cohen, M. B., Lee, Y. T., Miller, G., Pachocki, J., and Sidford, A. (2016). Geometric median in nearly linear time. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 9–21.

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

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

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

[44] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience.

[45] Cudeck, R. (2000). Exploratory factor analysis. In *Handbook of applied multivariate statistics and mathematical modeling*, pages 265–296. Elsevier.

[46] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transportation distances. In *Advances in neural information processing systems*.

[47] del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., and Rodriguez-Rodriguez, J. M. (1999). Tests of goodness of fit based on the l_2-wasserstein distance. *Annals of Statistics*, 27(4):1230–1239.

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

[49] Deming, W. E. and Stephan, F. F. (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.

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

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

[52] Duchi, J. C., Glynn, P. W., and Namkoong, H. (2021). Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 0(0).

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

[54] Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353.

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

[56] Fothergill, S., Mentis, H., Kohli, P., and Nowozin, S. (2012). Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 1737–1746. Association for Computing Machinery.

[57] Fournier, N. and Guillin, A. (2015). On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738.

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

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

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

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

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

[63] Gao, R. and Kleywegt, A. J. (2017b). Distributionally robust stochastic optimization with dependence structure. *arXiv preprint arXiv:1701.04200*.

[64] Genevay, A. (2019). *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE).

[65] Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). Sample complexity of sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 1574–1583.

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

[67] Genevay, A., Peyré, G., and Cuturi, M. (2018a). Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1608–1617.

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

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

[70] Gin, E. and Nickl, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, USA.

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

[72] Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

[73] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.

[74] Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009). A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 22, pages 673–681.

[75] Györfi, L. and Van Der Meulen, E. C. (1991). *A Consistent Goodness of Fit Test Based on the Total Variation Distance*, pages 631–645. Springer Netherlands, Dordrecht.

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

[77] Hildreth, C. (1957). A quadratic programming procedure. *Naval Research Logistics Quarterly*, 4(1):79–85.

[78] Hotelling, H. (1931). The generalization of student's ratio. *Annals of Mathematical Statistics*, 2(3):360–378.

[79] HQuang, M., Bazzani, L., and Murino, V. (2013). A unifying framework for vector-valued manifold regularization and multi-view learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 100–108.

[80] Hu, J., Liu, X., Wen, Z., and Yuan, Y. (2019). A brief introduction to manifold optimization. *arXiv preprint arXiv:1906.05450*.

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

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

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

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

[85] Huang, M., Ma, S., and Lai, L. (2021a). A riemannian block coordinate descent method for computing the projection robust wasserstein distance. *arXiv preprint arXiv:2012.05199*.

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

[87] Jiang, B., Ma, S., So, A. M.-C., and Zhang, S. (2017). Vector transport-free svrg with general retraction for riemannian optimization: Complexity analysis and practical implementation. *arXiv preprint arXiv:1705.09059*.

[88] Jitkrittum, W., Szabó, Z., Chwialkowski, K., and Gretton, A. (2016). Interpretable distribution features with maximum testing power. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 181–189.

[89] Jolliffe, I. (1986). *Principal Component Analysis*. Springer Verlag.

[90] Jr., F. J. M. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

[91] Kadri, H., Rabaoui, A., Preux, P., Duflos, E., and Rakotomamonjy, A. (2013). Functional regularized least squares classi cation with operator-valued kernels. *arXiv preprint arXiv:1301.2655*.

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

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

[94] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[95] Ledoux, M. (1999). Concentration of measure and logarithmic sobolev inequalities. *Séminaire de probabilités de Strasbourg*, 33:120–216.

[96] Lei, J. (2020). Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1).

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

[98] Li, J., Huang, S., and So, A. M.-C. (2019). A first-order algorithmic framework for wasserstein distributionally robust logistic regression. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3937–3947.

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

[100] Lin, T., Zheng, Z., Chen, E., Cuturi, M., and Jordan, M. (2021). On projection robust optimal transport: Sample complexity and model misspecification. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 262–270.

[101] Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. (2020). Learning deep kernels for non-parametric two-sample tests. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6316–6326.

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

[103] Lloyd, J. R. and Ghahramani, Z. (2015). Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, pages 829–837.

[104] Lopez-Paz, D. and Oquab, M. (2018). Revisiting classifier two-sample tests. In *International Conference on Learning Representations*.

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

[106] Luo, F. and 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.

[107] McDiarmid, C. (1989). *On the method of bounded differences*, pages 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press.

[108] Micchelli, C. A. and Pontil, M. A. (2005). On learning vector-valued functions. *Neural Computation*, 17(1):177–204.

[109] Minh, H. Q. and Sindhwani, V. (2011). Vector-valued manifold regularization. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 57–64.

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

[111] Mueller, J. and Jaakkola, T. (2015). Principal differences analysis: Interpretable characterization of differences between distributions. In *Advances in Neural Information Processing Systems*, volume 28.

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

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

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

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

[116] Nguyen, V. A., Si, N., and Blanchet, J. (2020). Robust bayesian classification using an optimistic score ratio. In *International Conference on Machine Learning*, pages 7327–7337.

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

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

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

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

[121] Pfanzagl, J. and Sheynin, O. (1996). Studies in the history of probability and statistics xliv a forerunner of the t-distribution. *Biometrika*, 83(4):891–898.

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

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

[124] Poor, H. and Hadjiliadis, O. (2008). *Quickest detection*. Cambridge University Press.

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

[126] Pratt, J. W. and Gibbons, J. D. (1981). *Kolmogorov-Smirnov Two-Sample Tests*, pages 318–344. Springer New York, New York, NY.

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

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

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

[130] Ramdas, A., Garcia, N., and Cuturi, M. (2017). On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2).

[131] Reddi, S. J., Ramdas, A., PAczos, B., Singh, A., and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, page 3571–3577.

[132] Rockafellar, R. T. (1970). *Convex analysis*. Princeton Mathematical Series. Princeton University Press.

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

[134] Savage, D., Zhang, X., Yu, X., Chou, P., and Wang, Q. (2014). Anomaly detection in online social networks. *Social networks*, 39:62–70.

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

[136] Schober, P. and Vetter, T. (2019). Two-sample unpaired t tests in medical research. *Anesthesia and analgesia*, 129:911.

[137] Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In Helmbold, D. and Williamson, B., editors, *Computational Learning Theory*, pages 416–426.

[138] Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319.

[139] Selvi, A., Belbasi, M. R., Haugh, M. B., and Wiesemann, W. (2022). Wasserstein logistic regression with mixed features. In *Advances in Neural Information Processing Systems*.

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

[141] Shafieezadeh Abadeh, S., Mohajerin Esfahani, P. M., and Kuhn, D. (2015). Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, volume 28.

[142] Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). *Lectures on stochastic programming: modeling and theory*. SIAM.

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

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

[145] Shen, X., Ali, A., and Boyd, S. (2022). Minimizing oracle-structured composite functions. *Optimization and Engineering*.

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

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

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

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

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

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

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

[153] Van Parys, B. P., Goulart, P. J., and Kuhn, D. (2015). Generalized gauss inequalities via semidefinite programming. *Mathematical Programming*, 156(1-2):271–302.

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

[155] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

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

[157] Wang, J., Gao, R., and Xie, Y. (2021a). Two-sample test using projected wasserstein distance. In *Proceedings of IEEE International Symposium on Information Theory*.

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

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

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

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

[162] Wen, Z. and Yin, W. (2012). A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434.

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

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

[165] Xie, L. and Xie, Y. (2021). Sequential change detection by optimal weighted ℓ_2 divergence. *IEEE Journal on Selected Areas in Information Theory*, pages 1–1.

[166] Xie, L., Zou, S., Xie, Y., and Veeravalli, V. V. (2021). Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory*, 2(2).

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

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

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

[170] Young, G. A., Severini, T. A., Young, G. A., Smith, R., Smith, R. L., et al. (2005). *Essentials of statistical inference*, volume 16. Cambridge University Press.

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

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

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

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

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