RELIABLE DECISION-MAKING UNDER UNCERTAINTY THROUGH THE LENS OF STATISTICS AND OPTIMIZATION

A Dissertation
Presented to
The Academic Faculty

By

Jie Wang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering
College of Engineering

Georgia Institute of Technology

May 2025

RELIABLE DECISION-MAKING UNDER UNCERTAINTY THROUGH THE LENS OF STATISTICS AND OPTIMIZATION

Thesis committee:

Dr. Yao Xie
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. George Lan
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Alexander Shapiro
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Xin Chen
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Rui Gao
McCombs School of Business
University of Texas at Austin

Date approved: April 7, 2025

To my family

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Prof. Yao Xie, for introducing me to the field of statistical learning and its applications in machine learning and healthcare, as well as for her unwavering support throughout my Ph.D. journey. Working with her on exciting research topics has been a privilege, and this thesis would not have been possible without her invaluable guidance. Her passion for research, keen insights, and remarkable ability to bridge theoretical advancements with practical applications have profoundly enriched my academic experience. I am especially appreciative of her constant encouragement, which has been instrumental in helping me navigate challenges in both research and life.

I would also like to express my sincere gratitude to my committee members – Prof. Rui Gao, Prof. Xin Chen, Prof. George Lan, and Prof. Alexander Shapiro—for their time, effort, and invaluable guidance in both my research and personal growth. I am especially thankful to Prof. Rui Gao for his support and mentorship in my research on operations research problems, as well as for his insightful suggestions during my job search. As an expert in operations research, he has taught me how to formulate meaningful research questions and approach them from a rigorous theoretical perspective. I also greatly admire his ability to integrate distributionally robust optimization with a broad range of practical applications. I would like to thank Prof. Xin Chen for his guidance in business analytics and optimization-related reserarch. He is always eager to explain the research problems and insights with great patience whenever I have a question. I am also grateful for his support throughout my job search process. I would like to thank Prof. George Lan for his guidance and insights into my research in continuous optimization. After my thesis proposal, he engaged in extensive discussions with me, offering extensive suggestions and potential improvements that have provided valuable future research directions. These discussions have inspired me to explore exciting new problems in my future career. I would like to thank Prof. Alexander Shapiro for his guidance and insights for my research on stochastic programming. His advice and

encouragements have been a great source of inspiration whenever I had the opportunity to speak with him. Additionally, I have greatly benefited from reading his textbook on stochastic programming during the early stage of my Ph.D. study, which played a pivotal role in shaping my research focus in this field.

I would like to extend my heartfelt gratitude to my colleagues and collaborators who made this work possible. In particular, I sincerely appreciate the invaluable contributions of my collaborators: Prof. Yao Xie, Prof. Rui Gao, Prof. Shenghao Yang, Dr. Yanyan Dong, Dr. Hoover H.F. Yin, Prof. Hongyuan Zha, Prof. Daniel Kuhn, Prof. Andreas Krause, Prof. Rishikesan Kamaleswaran, Dr. Ronald Moore, Prof. Minshuo Chen, Prof. Tuo Zhao, Prof. Wenjing Liao, Prof. Santanu S. Dey, Prof. Congduan Li, Prof. Yifan Hu, Prof. Urbashi Mitra, Dr. Talha Bozkus, Mr. Zhiyuan Jia, Dr. Yueyao Yu, Mr. Yiheng Zhang, Dr. Jun Ma, Prof. March Boedihardjo, Prof. Niao He, Prof. Xin Chen, Prof. Sherman S.M. Chow, Prof. Wenye Li, and Prof. Yin Zhang. I am also deeply grateful to all the members of Prof. Yao Xie's lab and my peers at the H. Milton Stewart School of Industrial and Systems Engineering (ISyE).

Finally, I wish to express my profound gratitude to my parents, Yujuan Cao and Chuanming Wang. I also thank my girlfriend, Prof. Yongchun Li, for her unconditional love and support. To them, I dedicate this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

In this thesis, we develop computationally efficient algorithms with statistical guarantees for problems of decision-making under uncertainty, particularly in the presence of large-scale, noisy, and high-dimensional data. In Chapter 2, we propose a kernelized projected Wasserstein distance for high-dimensional hypothesis testing, which finds the nonlinear mapping that maximizes the discrepancy between projected distributions. In Chapter 3, we provide an in-depth analysis of the computational and statistical guarantees of the kernelized projected Wasserstein distance. In Chapter 4, we study the variable selection problem in two-sample testing, aiming to select the most informative variables to determine whether two datasets follow the same distribution. In Chapter 5, we present a novel framework for distributionally robust stochastic optimization (DRO), which seeks an optimal decision that minimizes expected loss under the worst-case distribution within a specified set. This worst-case distribution is modeled using a variant of the Wasserstein distance based on entropic regularization. In Chapter 6, we incorporate Phi-divergence regularization into the infinity-type Wasserstein DRO, which is a formulation particularly useful for adversarial machine learning tasks. Chapter 7 concludes with an overview of promising future research directions.

CHAPTER 1
INTRODUCTION

Problems of decision making under uncertainty frequently occur in real-life applications, such as industrial engineering, computer science, and management. The existing literature proposed to formulate, analyze and solve these problems through the lens of optimization, in which they assume that the input optimization parameters can be obtained or estimated accurately from the 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, we propose solutions to address this challenge in various decision-making problems with large-scale, noisy, and 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 for the proposed modeling using 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 computing the KPW distance. By developing a representer theorem, we reformulate the problem as a finite-dimensional

optimization and employ a block coordinate descent algorithm that 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 study a special class of the KPW distance in which the data is projected into one dimension — termed the kernel max-sliced (KMS) Wasserstein distance. We establish sharp statistical and computational guarantees for this distance, which have practical implications for applications such as hypothesis testing, generative modeling, and distributionally robust optimization. Specifically:

(I) We provide a non-asymptotic estimate on the KMS p-Wasserstein distance between two empirical distributions based on n samples, referred to as the *finite-sample guarantees*. Our result shows that when the samples are drawn from identical populations, the rate of convergence is $n^{-1/(2p)}$, which is dimension-free and optimal in the worst case scenario.

(II) We analyze the computation of KMS 2-Wasserstein distance between two empirical distributions based on n samples. First, we show that computing this distance exactly is NP-hard. Consequently, we are prompted to propose a semidefinite relaxation (SDR) as an approximate heuristic with various guarantees.

- We develop an efficient first-order method with biased gradient oracles to solve the SDR, the complexity of which for finding a δ-optimal solution is $\widetilde{\mathcal{O}}\left(n^3\delta^{-3}\right)$. In comparison, the complexity of the interior point method for solving SDR is $\widetilde{\mathcal{O}}(n^{6.5})$ [25].
- We derive theoretical guarantees for the optimal solutions from the SDR. We show that there exists an optimal solution from SDR that is at most rank-k, where

$k \triangleq 1 + \lfloor \sqrt{2n + 9/4} - 3/2 \rfloor$, whereas computing the KMS distance exactly requires a rank-1 solution. We also provide a corresponding rank reduction algorithm designed to identify such low-rank solutions from the pool of optimal solutions of SDR.

In Chapter 4, we consider the variable selection problem for two-sample tests, aiming to select the most informative variables to determine whether two collections of samples follow the same distribution. Our approach involves maximizing a variance-regularized kernel maximum mean discrepancy (MMD) statistic, which in turn approximately maximizes testing power. We focus on three kernel types: linear, quadratic, and Gaussian. Specifically:

(I) From the computational perspective, we leverage mixed-integer programming techniques to solve the MMD optimization problem for variable selection. For linear kernel, we reformulate the optimization as an *inhomogeneous* quadratic maximization with ℓ_2 and ℓ_0 norm constraints (see Section 4.4.1), called **S**parse **T**rust **R**egion **S**ubproblem (STRS). Despite its NP-hardness, we provide an exact mixed-integer semi-definite programming formulation together with exact and approximation algorithms for solving this problem. To the best of our knowledge, this study is new in the literature. For quadratic and Gaussian kernels, the MMD optimization becomes a sparse maximization of a non-concave function (see Section 4.4.1), which is intractable in general. We propose a heuristic algorithm that iteratively optimizes a quadratic approximation of the objective function, which is also a special case of STRS.

(II) From the statistical perspective, we derive the rate of testing power of our framework under appropriate conditions. We demonstrate that when the training sample size n_{Tr} is sufficiently large, the type-II error decays in the order of $n_{\mathrm{Te}}^{-1/2}$, where n_{Te} denotes the testing sample size. For the three focused types of kernels, the training sample size requirement is almost independent of the data dimension D but dependent on the number of selected variables d: For linear, quadratic, and Gaussian kernels, to achieve satisfac-

tory performance, the training sample sizes are at least $\Omega(d^2 \log \frac{D}{d}), \Omega(d^4 \log \frac{D}{d})$, and $\Omega(d \log \frac{D}{d})$, respectively.

(III) By combining both viewpoints, it becomes evident that there exists a balance between computational tractability and statistical guarantees. While the Gaussian kernel requires a smaller sample size to be statistically powerful, the corresponding MMD optimization is challenging. Conversely, the linear or quadratic kernel may require more samples to be statistically powerful, but the optimization is easier to solve.

In Chapter 5, 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 measures 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 6, we incorporate ϕ-divergence regularization, which includes entropic regularization as a special case, into the ∞-type Wasserstein DRO formulation. The ∞-type Wasserstein DRO has important connections with the adversarial robsut machine learning approach. We demonstrate several notable properties of this new formulation, as summarized below:

(I) The dual formulation of the regularized ∞-type Wasserstein DRO is a variant of optimized certainty equivalent (OCE) risk measure studied in [26]. It can also be interpreted as a smooth approximation of the pointwise maximization of the loss function over a compact set. The worst-case distribution that corresponds to the regularized ∞-type Wasserstein DRO is also derived. In contrast to the un-regularized formulation that *deterministically* transports each data to its extreme perturbation, the worst-case distribution of our formulation transports each data towards the entire domain set through specific absolutely continuous distributions. This observation indicates that our formulation is well-suited for adversarial defense where the data distribution after adversarial attack manifests as absolutely continuous, such as through the addition of white noise to the data.

(II) We adopt the idea of stochastic approximation to solve our reformulation by iteratively obtaining a stochastic gradient estimator and next performing projected gradient update. To tackle the difficulty that one cannot obtain the unbiased gradient estimator, we

introduce and analyze stochastic gradient methods with biased oracles inspired from Hu et al. [155]. Our proposed algorithm achieves $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity for finding ϵ-optimal solution for convex loss and general choices of ϕ-divergence, and $\widetilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity for finding ϵ-stationary point for nonconvex loss and KL-divergence. These sample complexity results are near-optimal up to a near-constant factor.

(III) We derive several statistical properties of our proposed formulation. First, we analyze the regularization effects of our formulation and show that it is asymptotically equivalent to regularized ERM formulations under three different scalings of the regularization parameter and the robustness level. Second, we investigate the generalization properties of our adversarial training framework. In particular, the optimal value of the optimization with empirical data acts as a confidence upper bound for its population counterpart, with only a negligible residual error. We further derive specific generalization error bounds for both linear and neural network function classes.

CHAPTER 2
KERNEL PROJECTED WASSERSTEIN DISTANCE WITH APPLICATIONS TO HYPOTHESIS TESTING

In this chapter, 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. This work is mainly summarized in [307].

2.1 Background

As a fundamental problem in statistical inference [331], 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, 66, 264] to identify abnormal observations that follow a distinct distribution compared with typical observations. Similarly, in change-point detection [252, 322, 323], two-sample testing is essential to detect abrupt changes in streaming data. Other notable examples include model criticism [36, 79, 204], causal inference [206], and health care [266].

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 [149], Student's t-test [248], 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 [216, 254], total variation distance [141], and Wasserstein distance [93, 256]. 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) [76, 138, 139]. Although the power of MMD test with the median choice of kernel bandwidth decays quickly when the dimension of distributions increases [257], this test with properly chosen bandwidth does not have the curse of dimensionality issue for low-dimensional manifold data as pointed out in [76]. 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 [306, 322] 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) [171], kernel PCA [268], factor analysis [91], etc. Several works leverage this idea to design tests for high-dimensional data. [228] and [322] first design tests by finding the worst-case linear projector that maximizes the distance between projected sample points in one dimension. Later [199] and [306] 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 [163, 201]. 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 [130], 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 [60, 150, 172, 223] 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 [116]. 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

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

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 [201, 306]. 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

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

In particular, the Hilbert space $\mathcal{H}_K$ satisfies the reproducing property, i.e., $\langle f, K_x y\rangle_{\mathcal{H}_K} = \langle f(x), y\rangle$ for $\forall f \in \mathcal{H}_K$.

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 [18, 62, 219, 223] 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 [136] 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.

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 $\frac{1}{N_p}\sum_{t=1}^{N_p} 1\{T_t \geq T\}$.

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 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 [267, 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 [127] 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 [148] 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 [163]. 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 [190, 201], and is satisfied for various exponential measures such as the Gaussian distribution. See [189] for more examples.*

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

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

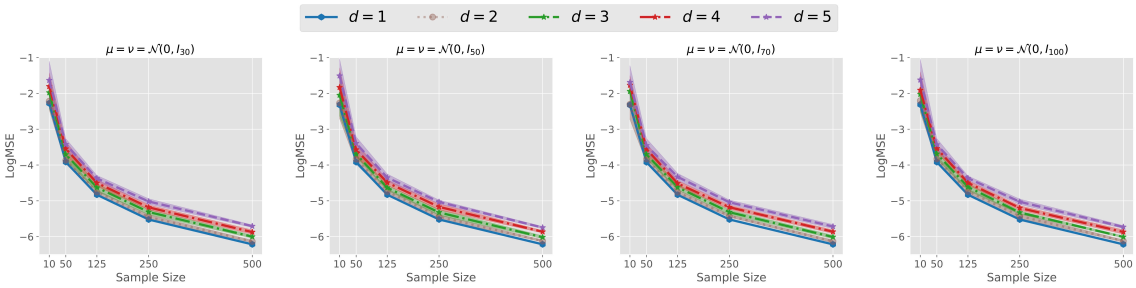

Figure 2.1: Average values of KPW distances between empirical distributions $\hat{\mu}_n$ and $\hat{\nu}_n$ as the sample size n varies. Results are averaged for 10 independent trials and the shaded areas show the corresponding error bars.

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 [201], 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

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

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*

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

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

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.

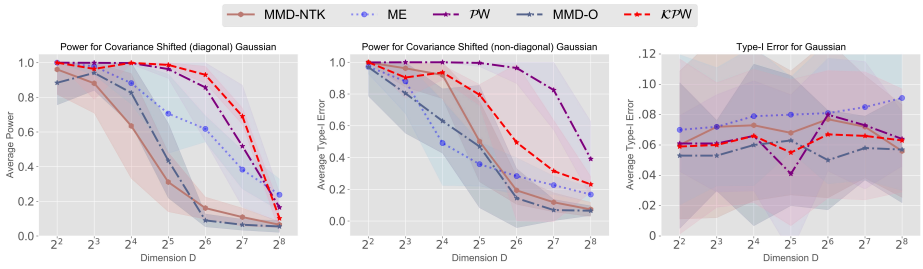

Figure 2.2: Testing results on Gaussian distributions across different choices of dimension D. Left: power for Gaussian distributions, where the shifted covariance matrix is still diagonal; Middle: power for Gaussian distributions, where the shifted covariance matrix is non-diagonal; Right: Type-I error.

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**

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 [306]; (ii) MMD-O: the MMD test with a Gaussian kernel whose bandwidth is optimized [202]; (iii) MMD-NTK: the test that combines both neural networks and MMD [77]; and (iv) ME: the mean embedding test with optimized hyper-parameters [170]. 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 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

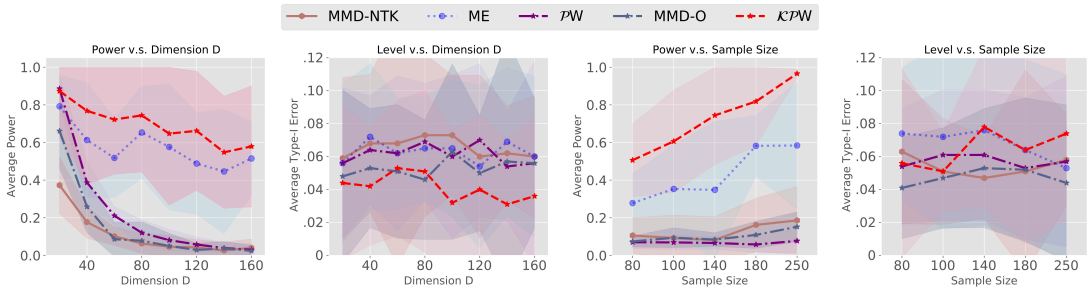

Figure 2.3: Testing results on Gaussian-mixture distributions. Left two: type-I and type-II errors across different choices of dimension D with fixed sample size $n = m = 200$; Right two: type-I and type-II errors across different choices of sample size $n = m$ with fixed dimension $D = 140$.

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

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

in Appendix A.7.4.

2.5.2 Tests for MNIST handwritten digits

We now perform two-sample tests on the MNIST dataset [187]. 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 [202, 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 $\alpha = 0.05$.

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 [115]. 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 of research interest to determine the optimal hyper-parameters for the KPW test, including the projected subspace dimension d and the matrix-valued kernel function K. Second, 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

STATISTICAL AND COMPUTATIONAL GUARANTEES OF KERNEL MAX-SLICED WASSERSTEIN DISTANCES

In this chapter, we study the kernel max-sliced (KMS) Wasserstein distance that finds an optimal nonlinear mapping that reduces data into 1 dimension before computing the Wasserstein distance. Its theoretical properties have not yet been fully developed. To fill the gap, we provide sharp finite-sample guarantees under milder technical assumptions compared with state-of-the-art for the KMS p-Wasserstein distance between two empirical distributions with n samples for general $p \in [1, \infty)$. Algorithm-wise, we show that computing the KMS 2-Wasserstein distance is NP-hard, and then we further propose a semidefinite relaxation (SDR) formulation (which can be solved efficiently in polynomial time) and provide a relaxation gap for the obtained solution. We provide numerical examples to demonstrate the good performance of our scheme for high-dimensional two-sample testing. This work is mainly summarized in [302].

3.1 Introduction

Optimal transport (OT) has achieved much success in various areas, such as generative modeling [130, 208, 244, 246], distributional robust optimization [124, 126, 305], nonparametric testing [256, 303, 309, 314, 322], domain adaptation [17, 85, 86, 88, 313], etc. See [247] for comprehensive reviews on these topics. The sample complexity of the Wasserstein distance has been an essential building block for OT in statistical inference. It studies the relationship between a population distribution μ and its empirical distribution $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ with $x_i \sim \mu$ in terms of the "Wasserstein distance". Unfortunately, the sample size n needs to be exponentially large in data dimension to achieve an accurate enough estimation [116], called the *curse of dimensionality* issue.

To tackle the challenge of high dimensionality, it is meaningful to combine OT with projection operators in low-dimensional spaces. Researchers first attempted to study the sliced Wasserstein distance [55, 64, 98, 177, 178, 230, 235], which computes the average of the Wasserstein distance between two projected distributions using random one-dimensional projections. Since a single random projection contains little information to distinguish two high-dimensional distributions, computing the sliced Wasserstein distance requires a large number of linear projections. To address this issue, more recent literature considered the **M**ax-**S**liced (**MS**) Wasserstein distance that seeks the *optimal* projection direction such that the Wasserstein distance between projected distributions is maximized [97, 199, 201, 245, 306]. Later, Wang et al. [307] modified the MS Wasserstein distance by seeking an optimal *nonlinear* projection belonging to a ball of reproducing kernel Hilbert space (RKHS), which we call the **K***ernel* **M***ax*-**S***liced (***KMS***) Wasserstein distance*. The motivation is that a nonlinear projector can be more flexible in capturing the differences between two high-dimensional distributions; it is worth noting that KMS Wasserstein reduces to MS Wasserstein when specifying a dot product kernel.

Despite promising applications of the KMS Wasserstein distance, its statistical and computational results have not yet been fully developed. From a statistical perspective, Wang et al. [307] built concentration properties of the empirical KMS Wasserstein distance for distribution that satisfies the Poincar'e inequality projection and the Poincar'e inequality [189], which could be difficult to verify in practice. From a computational perspective, the authors therein developed a gradient-based algorithm to approximately compute the empirical KMS Wasserstein distance. However, there is no theoretical guarantee on the quality of the local optimum solution obtained. In numerical experiments, the quality of the solution obtained is highly sensitive to the initialization.

To address the aforementioned limitations, this paper provides new statistical and computational guarantees for the KMS Wasserstein distance. Our key contributions are summarized as follows.

- We provide a non-asymptotic estimate on the KMS p-Wasserstein distance between two empirical distributions based on n samples, referred to as the *finite-sample guarantees*. Our result shows that when the samples are drawn from identical populations, the rate of convergence is $n^{-1/(2p)}$, which is dimension-free and optimal in the worst case scenario.

- We analyze the computation of KMS 2-Wasserstein distance between two empirical distributions based on n samples. First, we show that computing this distance exactly is NP-hard. Consequently, we are prompted to propose a semidefinite relaxation (SDR) as an approximate heuristic with various guarantees.

  - We develop an efficient first-order method with biased gradient oracles to solve the SDR, the complexity of which for finding a δ-optimal solution is $\widetilde{\mathcal{O}}\left(n^3\delta^{-3}\right)$. In comparison, the complexity of the interior point method for solving SDR is $\widetilde{\mathcal{O}}(n^{6.5})$ [25].

  - We derive theoretical guarantees for the optimal solutions from the SDR. We show that there exists an optimal solution from SDR that is at most rank-k, where $k \triangleq 1 + \lfloor\sqrt{2n + 9/4} - 3/2\rfloor$, whereas computing the KMS distance exactly requires a rank-1 solution. We also provide a corresponding rank reduction algorithm designed to identify such low-rank solutions from the pool of optimal solutions of SDR.

- We exemplify our theoretical results in non-parametric two-sample testing, human activity detection, and generative modeling. Our numerical results showcase the stable performance and quick computational time of our SDR formulation, as well as the desired sample complexity rate of the empirical KMS Wasserstein distance.

In the following, we compare our work with the most closely related literature, and defer the detailed comparision of KMS Wasserstein distance with other variants of OT divergences in Appendix B.1.

Literature Review. The study on the statistical and computational results of MS and KMS Wasserstein distances is popular in the existing literature. From a statistical perspective, existing results on the rate of empirical MS/KMS Wasserstein are either dimension-dependent, suboptimal or require regularity assumptions (e.g., log-concavity, Poincaré inequality, projection Bernstein tail condition) on the population distributions [19, 199, 238, 306], except for the very recent literature [50] that provides a sharp, dimension-free rate for MS Wasserstein with data distributions supported on a compact subspace but without regularity assumptions. From the computational perspective, there are two main approaches to compute such distances. One is to apply gradient-based algorithms to find local optimal solutions or stationary points, see, e.g., [162, 163, 168, 199, 307]. Unfortunately, due to the highly non-convex nature of the optimization problem, the quality of the estimated solution is unstable and highly depends on the choice of initial guess. The other is to consider solving its SDR instead [245], yet theoretical guarantees on the solution from convex relaxation are missing. Inspired by existing reference [21, 102, 198, 243] that studied the rank bound of SDR for various applications, we adopt their proof techniques to provide similar guarantees for computing KMS in Theorem 10. Besides, all listed references add entropic regularization to the inner OT problem and solve the regularized version of MS/KMS Wasserstein distances instead, while the gap between the solutions from regularized and original problems could be non-negligible.

3.2 Background

We first introduce the definition of Wasserstein and KMS Wasserstein distances below.

Definition 6 (Wasserstein Distance)**.** *Let $p \in [1, \infty)$. Given a normed space $(\mathcal{V}, \|\cdot\|)$, the p-Wasserstein distance between two probability measures μ, ν on $\mathcal{V}$ is defined as*

$$W_p(\mu, \nu) = \left(\min_{\pi \in \Gamma(\mu,\nu)} \int \|x - y\|^p \, \mathrm{d}\pi(x, y) \right)^{1/p},$$

where $\Gamma(\mu, \nu)$ denotes the set of all probability measures on $\mathcal{V} \times \mathcal{V}$ with marginal distributions μ and ν.

Definition 7 (RKHS)**.** *Consider a symmetric and positive definite kernel $K : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$, where $\mathcal{B} \subseteq \mathbb{R}^d$. Given such a kernel, there exists a unique Hilbert space $\mathcal{H}$, called RKHS, associated with the reproducing kernel K. Denote by K_x the kernel section at $x \in \mathcal{B}$ defined by $K_x(y) = K(x, y), \forall y \in \mathcal{B}$. Any function $f \in \mathcal{H}$ satisfies the reproducing property $f(x) = \langle f, K_x \rangle_{\mathcal{H}}, \forall x \in \mathcal{B}$. For $x, y \in \mathcal{B}$, it holds that $K(x, y) = \langle K_x, K_y \rangle_{\mathcal{H}}$.*

Definition 8 (KMS Wasserstein Distance)**.** *Let $p \in [1, \infty)$. Given two distributions μ and ν, define the p-KMS Wasserstein distance as*

$$\mathcal{KMS}_p(\mu, \nu) = \max_{f \in \mathcal{H},\ \|f\|_{\mathcal{H}} \leq 1} W_p(f_{\#}\mu, f_{\#}\nu),$$

where $f_{\#}\mu$ and $f_{\#}\nu$ are the pushforward measures of μ and ν by $f : \mathcal{B} \to \mathbb{R}$, respectively.

In particular, for dot product kernel $K(x, y) = x^{\mathrm{T}}y$, the RKHS $\mathcal{H} = \{f : f(x) = x^{\mathrm{T}}\beta, \exists \beta \in \mathbb{R}^d\}$. In this case, the KMS Wasserstein distance reduces to the max-sliced Wasserstein distance [97]. A more flexible choice is the Gaussian kernel $K(x, y) = \exp(-\frac{1}{2\sigma^2}\|x - y\|_2^2)$, where $\sigma > 0$ denotes the temperature hyper-parameter. In this case, the function class $\mathcal{H}$ satisfies the *universal property* as it is dense in the continuous function class with respect to the ∞-type functional norm. It is easy to see the following theorem holds.

Theorem 5 (Metric Property of KMS)**.** *Let $\mathcal{H}$ be a universal RKHS. Then $\mathcal{KMS}_p(\mu, \nu) = 0$ if and only if $\mu = \nu$.*

Example 1. *We present a toy example that highlights the flexibility of the KMS Wasserstein distance. Fig.3.1(a) displays a scatter plot of the circle dataset, which consists of two groups of samples distributed along inner and outer circles, perturbed by Gaussian noise. Since the data exhibit a nonlinear structure, distinguishing these groups using a linear*

projection is challenging. As shown in Fig.3.1(b), the density plot of the projected samples using the MS Wasserstein distance with a linear projector is not sufficiently discriminative. In contrast, Fig. 3.1(c) demonstrates that the KMS Wasserstein distance is better suited for distinguishing these two groups. Intuitively, the optimal nonlinear projector should take the form $f^*(x) \propto \|x\|_2^c$ *for some scalar* $c > 0$*, as illustrated in Fig.3.1(d). We plot the nonlinear projector by computing the empirical KMS Wasserstein distance, shown in Fig.3.1(e), which closely resembles the circular landscape depicted in Fig. 3.1(d). This result demonstrates that the KMS Wasserstein distance provides a data-driven, non-parametric nonlinear projector capable of effectively distinguishing distinct data groups.*

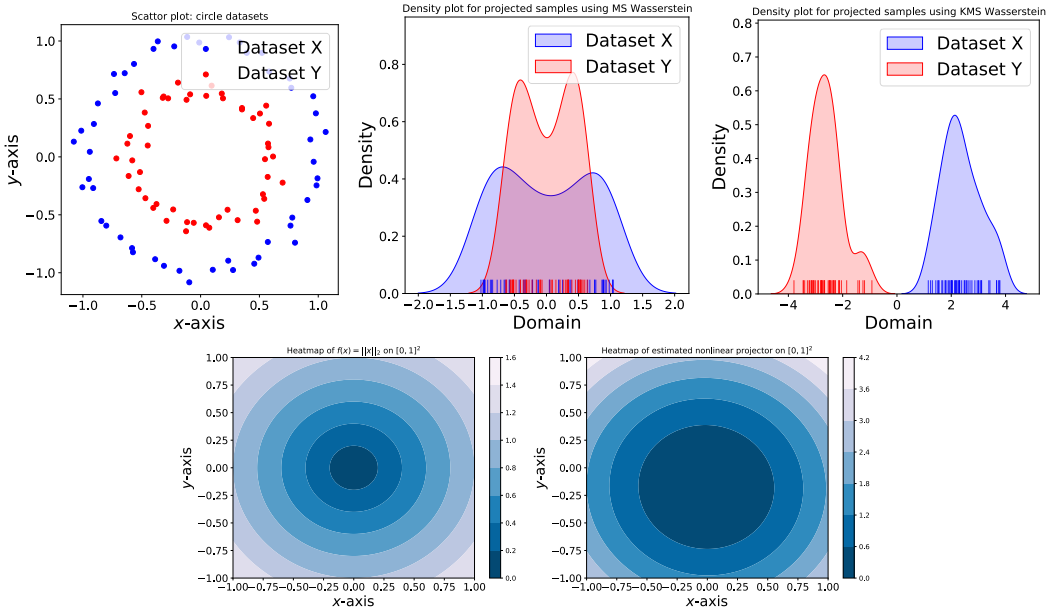

Figure 3.1: Results on a 2-dimensional toy example. (a) Scatter plot of circle dataset; (b) Density plot using MS Wasserstein; (c) Density plot using KMS Wasserstein; (d) Plot of $f(x) = \|x\|_2, x \in [0,1]^2$; (e) Plot of estimated projector using KMS Wasserstein.

Given the RKHS $\mathcal{H}$, let the *canonical feature map* that embeds data to $\mathcal{H}$ as

$$\Phi : \mathcal{B} \to \mathcal{H}, \qquad x \mapsto \Phi(x) = K_x. \tag{3.1}$$

Define the functional $u_f : \mathcal{H} \to \mathbb{R}$ by $u_f(g) = \langle f, g \rangle_{\mathcal{H}}$ for any $g \in \mathcal{H}$, which can be viewed as a linear projector that maps data from the Hilbert space $\mathcal{H}$ to the real line. In light of this,

for two probability measures μ and ν on $\mathcal{H}$, we define the MS p-Wasserstein distance

$$\mathcal{MS}_p(\mu,\nu) = \sup_{f\in\mathcal{H}:\ \|f\|_{\mathcal{H}}\le 1} W_p\Big((u_f)_{\#}\mu, (u_f)_{\#}\nu\Big), \tag{3.2}$$

where $(u_f)_{\#}\mu$ denotes the pushforward measure of μ by the map u_f, i.e., if μ is the distribution of a random element X of $\mathcal{H}$, then $(u_f)_{\#}\mu$ is the distribution of the random variable $u_f(X) = \langle f, X\rangle_{\mathcal{H}}$, and $(u_f)_{\#}\nu$ is defined likewise. In the following, we show that the KMS Wasserstein distance in Definition 8 can be reformulated as the MS Wasserstein distance between two distributions on (infinite-dimensional) Hilbert space.

Remark 4 (Reformulation of KMS Wasserstein)**.** *By the reproducing property, we can see that $f(x) = \langle f, K_x\rangle_{\mathcal{H}} = u_f(\Phi(x))$, which implies $f = u_f \circ \Phi$. As a consequence,*

$$\begin{aligned} &\mathcal{KMS}_p(\mu,\nu) \\ =&\sup_{f\in\mathcal{H}:\ \|f\|_{\mathcal{H}}\le 1} W_p\Big((u_f)_{\#}\big(\Phi_{\#}\mu\big), (u_f)_{\#}\big(\Phi_{\#}\nu\big)\Big) \\ =&\mathcal{MS}_p\Big(\Phi_{\#}\mu, \Phi_{\#}\nu\Big). \end{aligned} \tag{3.3}$$

In other words, the KMS Wasserstein distance first maps data points into the infinite-dimensional Hilbert space $\mathcal{H}$ through the canonical feature map Φ, and next finds the linear projector to maximally distinguish data from two populations. Compared with the traditional MS Wasserstein distance [97] that performs linear projection in $\mathbb{R}^d$, KMS Wasserstein distance is a more flexible notion.

Remark 5 (Connections with Kernel PCA)**.** *Given data points $x_1, \ldots, x_n$ on $\mathcal{B}$, denote by $\widehat{\mu}_n$ the corresponding empirical distribution. Assume $\frac{1}{n}\sum_{i\in[n]}\Phi(x_i) = 0$, since otherwise one can center those data points as a preprocessing step. Kernel PCA [221] is a popular tool for nonlinear dimensionality reduction. When seeking the first principal nonlinear projection function f, Mairal and Vert [215] presents the following reformulation of kernel*

PCA:

$$\underset{f\in\mathcal{H}:\ \|f\|_{\mathcal{H}}\leq 1}{\arg\max}\ \mathit{Var}\Big((u_f)_{\#}(\Phi_{\#}\widehat{\mu}_n)\Big), \tag{3.4}$$

where Var$(\cdot)$ *denotes the variance of a given probability measure. In comparison, the KMS Wasserstein distance aims to find the optimal nonlinear projection that distinguishes two populations and replaces the variance objective in* (3.4) *with the Wasserstein distance between two projected distributions in* (3.3)*. Also, kernel PCA is a special case of KMS Wasserstein by taking* $p = 2, \mu \equiv \widehat{\mu}_n, \nu \equiv \delta_0$ *in* (3.3)*.*

Notations. Let $\langle\cdot,\cdot\rangle$ denote the inner product operator. For any positive integer n, denote $[n] = \{1, 2, \ldots, n\}$. Define Γ_n as the set

$$\Big\{\pi \in \mathbb{R}_+^{n\times n} :\ \sum_{i=1}^{n}\pi_{i,j} = \frac{1}{n}, \sum_{j=1}^{n}\pi_{i,j} = \frac{1}{n}, \forall i, j \in [n]\Big\}. \tag{3.5}$$

Let Conv(P) denote a convex hull of the set P, and $\mathbb{S}_n^+$ denote the set of positive semidefinite matrices of size $n \times n$. We use $\widetilde{\mathcal{O}}(\cdot)$ as a variant of $\mathcal{O}(\cdot)$ to hide logarithmic factors.

3.3 Statistical Guarantees

Suppose samples $x^n := \{x_i\}_{i\in[n]}$ and $y^n := \{y_i\}_{i\in[n]}$ are given and follow distributions μ, ν, respectively. Denote by $\widehat{\mu}_n$ and $\widehat{\nu}_n$ the corresponding empirical distributions from samples x^n and y^n. In this section, we provide a finite-sample guarantee on the p-KMS Wasserstein distance between $\widehat{\mu}_n$ and $\widehat{\nu}_n$ with $p \in [1, \infty)$. This guarantee can be helpful for KMS Wasserstein distance-based hypothesis testing that has been studied in [307]: Suppose one aims to build a non-parametric test to distinguish two hypotheses $H_0 :\ \mu = \nu$ and $H_1 :\ \mu \neq \nu$. Thus, it is crucial to control the high-probability upper bound of $\mathcal{KMS}_p(\widehat{\mu}_n, \widehat{\nu}_n)$ under H_0 as it serves as the critical value to determine whether H_0 is rejected or not. We first make the following assumption on the kernel.

Assumption 2. *There exists some constant* $A > 0$ *such that the kernel* $K(\cdot,\cdot)$ *satisfies*

$\sqrt{K(x,x)} \leq A, \forall x \in \mathcal{B}$.

Assumption 2 is standard in the literature (see, e.g., [138]), and is quite mild: Gaussian kernel $K(x,y) = \exp(-\|x-y\|_2^2/\sigma^2)$ naturally fits into this assumption. For dot product kernel $K(x,y) = x^{\mathrm{T}}y$, if we assume the support $\mathcal{B}$ has a finite diameter, this assumption can also be satisfied. Define the critical value

$$\Delta(n,\alpha) = 4A\Big(C + 4\sqrt{\log\frac{2}{\alpha}}\Big)^{1/p} \cdot n^{-1/(2p)},$$

where $C \geq 1$ is a universal constant. We show the following finite-sample guarantees on KMS p-Wassersrein distance.

Theorem 6 (Finite-Sample Guarantees)**.** *Fix $p \in [1,\infty)$, level $\alpha \in (0,1)$, and suppose Assumption 2 holds.*

(I) *(One-Sample Guarantee) With probability at least $1-\alpha$, it holds that*

$$\mathcal{KMS}_p(\widehat{\mu}_n, \mu) \leq \frac{1}{2}\Delta(n,\alpha).$$

(II) *(Two-Sample Guarantee) With probability at least $1-\alpha$, it holds that*

$$\mathcal{KMS}_p(\widehat{\mu}_n, \widehat{\nu}_n) \leq \mathcal{KMS}_p(\mu,\nu) + \Delta(n,\alpha).$$

The dimension-free upper bound $\Delta(n,\alpha) = O(n^{-1(/2p)})$ is optimal in the worst case. Indeed, in the one-dimension case $\mathcal{B} = [0,1]$ and $K(x,y) = xy$, the kernel max-sliced Wasserstein distance $\mathcal{KMS}_p$ coincides with the classical Wasserstein distance W_p. In this case, it is easy to see that if $\mu = (\delta_0 + \delta_1)/2$ is supported on the two points 0 and 1, the expectation of $\mathcal{KMS}(\widehat{\mu}_n, \widehat{\nu}_n)$ is of order $n^{-1/(2p)}$ [116]. We also compare this bound with other OT divergences in Appendix B.1.

We design a two-sample test $\mathcal{T}_{\mathrm{KMS}}$ such that H_0 is rejected if $\mathcal{KMS}_p(\widehat{\mu}_n, \widehat{\nu}_n) > \Delta(n, \alpha)$. By Theorem 6, we have the following performance guarantees of $\mathcal{T}_{\mathrm{KMS}}$.

Corollary 1 (Testing Power of $\mathcal{T}_{\mathrm{KMS}}$)**.** *Fix a level $\alpha \in (0, 1/2)$, $p \in [1, \infty)$, and suppose Assumption 2 holds. Then the following result holds:*

(I) *(Risk): The type-I risk of $\mathcal{T}_{\mathrm{KMS}}$ is at most α;*

(II) *(Power): Under $H_1 : \mu \neq \nu$, suppose the sample size n is sufficiently large such that $\varrho_n := \mathcal{KMS}_p(\mu, \nu) - \Delta(n, \alpha) > 0$, the power of $\mathcal{T}_{\mathrm{KMS}}$ is at least $1 - c \cdot n^{-1/(2p)}$, where c is a constant depending on A, C, p, ϱ_n.*

Remark 6 (Comparision with Maximum Mean Discrepancy (MMD))**.** *MMD has been a popular kernel-based tool to quantify the discrepancy between two probability measures (see, e.g., [28, 118, 138, 175, 202, 227, 270, 271, 291, 304]), which, for any two probability distributions μ and ν, is defined as*

$$\begin{aligned} \mathit{MMD}(\mu, \nu) &= \max_{\substack{f \in \mathcal{H}, \\ \|f\|_{\mathcal{H}} \leq 1}} \mathbb{E}_\mu[f] - \mathbb{E}_\nu[f] \\ &= \max_{\substack{f \in \mathcal{H}, \\ \|f\|_{\mathcal{H}} \leq 1}} \overline{(u_f)_{\#}\big(\Phi_{\#}\mu\big)} - \overline{(u_f)_{\#}\big(\Phi_{\#}\nu\big)}, \end{aligned} \tag{3.6}$$

where $\overline{\xi}$ denotes the mean of a given probability measure ξ. The empirical (biased) MMD estimator also exhibits dimension-free finite-sample guarantee as in Theorem 6: it decays in the order of $\mathcal{O}(n^{-1/2})$, where n is the number of samples. However, the KMS Wasserstein distance is more flexible as it replaces the mean difference objective in (3.6) *by the Wasserstein distance, which naturally incorporates the geometry of the sample space and is suitable for hedging against adversarial data perturbations [124].*

3.4 Computing 2-KMS Wasserstein Distance

Let $\widehat{\mu}_n$ and $\widehat{\nu}_n$ be two empirical distributions supported on n points, i.e., $\widehat{\mu}_n = \frac{1}{n}\sum_i \delta_{x_i}, \widehat{\nu}_n = \frac{1}{n}\sum_j \delta_{y_j}$, where $\{x_i\}_i, \{y_j\}_j$ are data points in $\mathbb{R}^d$. This section focuses on the computation

of 2-KMS Wasserstein distance between these two distributions. By Definition 8 and monotonicity of square root function, it holds that

$$\begin{aligned}&\mathcal{KMS}_2(\widehat{\mu}_n, \widehat{\nu}_n)\\ &= \left(\max_{f\in\mathcal{H},\ \|f\|_{\mathcal{H}}^2\le 1} \left\{ \min_{\pi\in\Gamma_n} \sum_{i,j\in[n]} \pi_{i,j} |f(x_i) - f(y_j)|^2 \right\} \right)^{1/2}, \end{aligned} \tag{KMS}$$

where Γ_n is defined in (3.5).

Although the outer maximization problem is a *functional optimization* that contains uncountably many parameters, one can apply the representer theorem (see below) to reformulate Problem (KMS) as a finite-dimensional optimization.

Theorem 7 (Theorem 1 in [307])**.** *There exists an optimal solution to* (KMS)*, denoted as* $\widehat{f}$*, such that for any* z*,*

$$\widehat{f}(z) = \sum_{i=1}^{n} a_{x,i} K(z, x_i) - \sum_{i=1}^{n} a_{y,i} K(z, y_i), \tag{3.7}$$

where $a_x = (a_{x,i})_{i\in[n]}, a_y = (a_{y,i})_{i\in[n]}$ *are coefficients to be determined.*

Define gram matrix $K(x^n, x^n) = (K(x_i, x_j))_{i,j\in[n]}$ and other gram matrices $K(x^n, y^n)$, $K(y^n, x^n), K(y^n, y^n)$ likewise, then define the concatenation of gram matrics

$$G = \begin{pmatrix} K(x^n, x^n) & -K(x^n, y^n) \\ -K(y^n, x^n) & K(y^n, y^n) \end{pmatrix} \in \mathbb{R}^{2n\times 2n}. \tag{3.8}$$

Assume G is positive definite [1] such that it admits the Cholesky decomposition $G^{-1} = UU^{\mathrm{T}}$. By substituting the expression (3.7) into (KMS) and calculation (see Appendix B.4), we

[1] In Appendix B.3, we provide its sufficient condition.

obtain the exact reformulation of (KMS):

$$\max_{\omega\in\mathbb{R}^{2n}:\ \|\omega\|_2=1} \left\{ \min_{\pi\in\Gamma_n} \sum_{i,j} \pi_{i,j} (M_{i,j}^{\mathrm{T}}\omega)^2 \right\}. \tag{3.9}$$

Here, we omit taking the square root of the optimal value of the max-min optimization problem for simplicity of presentation and define the vector $M_{i,j} = U^{\mathrm{T}} M'_{i,j}$, where

$$M'_{i,j} = \begin{pmatrix} (K(x_i, x_l) - K(y_j, x_l))_{l\in[n]} \\ (K(y_j, y_l) - K(x_i, y_l))_{l\in[n]} \end{pmatrix} \in \mathbb{R}^{2n}.$$

Since Problem (3.9) is a non-convex program, it is natural to question its computational hardness. The following gives an *affirmative* answer, whose proof is provided in Appendix B.5.

Theorem 8 (NP-hardness)**.** *Problem* (3.9) *is NP-hard for the worst-case instances of* $\{M_{i,j}\}_{i,j}$.

The proof idea of Theorem 8 is to find an instance of $\{M_{i,j}\}_{i,j}$ that depends on a generic collection of n vectors $\{A_i\}_i$ such that solving (3.9) is at least as difficult as solving the fair-PCA problem [263] with rank-1 matrices (or fair beamforming problem [283]) and has been proved to be NP-hard [283]. Interestingly, the computational hardness of the MS Wasserstein distance arises from both the data dimension d and the sample size n, whereas that of the KMS Wasserstein distance arises from the sample size n only.

To tackle the computational challenge of solving (3.9), in the subsequent subsections, we present an SDR formula and propose an efficient first-order algorithm to solve it. Next, we analyze the computational complexity of our proposed algorithm and the theoretical guarantees on SDR.

3.4.1 Semidefinite Relaxation with Efficient Algorithms

We observe the simple reformulation of the objective in (3.9):

$$\sum_{i,j} \pi_{i,j} (M_{i,j}^{\mathrm{T}} \omega)^2 = \sum_{i,j} \pi_{i,j} \langle M_{i,j} M_{i,j}^{\mathrm{T}}, \omega\omega^{\mathrm{T}} \rangle.$$

Inspired by this relation, we use the change of variable approach to optimize the rank-1 matrix $S = \omega\omega^{\mathrm{T}}$, i.e., it suffices to consider the equivalent reformulation of (3.9):

$$\max_{\substack{S \in \mathbb{S}_+^{2n}, \\ \mathrm{Trace}(S)=1, \mathrm{rank}(S)=1}} \left\{ F(S) = \min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} \langle M_{i,j} M_{i,j}^{\mathrm{T}}, S \rangle \right\}. \tag{3.10}$$

An efficient SDR is to drop the rank-1 constraint to consider the semidefinite program (SDP):

$$\max_{S \in \mathcal{S}_{2n}} F(S), \quad \text{where} \quad \mathcal{S}_{2n} = \Big\{ S \in \mathbb{S}_+^{2n} : \ \mathrm{Trace}(S) = 1 \Big\}. \tag{SDR}$$

Remark 7 (Connection with [322])**.** *We highlight that Xie and Xie [322] considered the same SDR heuristic to compute the MS* 1*-Wasserstein distance. However, the authors therein apply the interior point method to solve a large-scale SDP, which has expansive complexity* $\mathcal{O}(n^{6.5} \mathrm{polylog}(\frac{1}{\delta}))$ *(up to* δ*-accuracy) [25]. In the following, we present a first-order method that exhibits much smaller complexity* $\widetilde{\mathcal{O}}\left(n^2 \delta^{-3}\right)$ *in terms of the problem size* n *(see Theorem 9). Besides, theoretical guarantees on the solution from SDR have not been explored in [322], and we are the first literature to provide these results.*

The constraint set $\mathcal{S}_{2n}$ is called the *Spectrahedron* and admits closed-form Bregman projection. Inspired by this, we propose an inexact mirror ascent algorithm to solve (SDR). Its high-level idea is to iteratively construct an inexact gradient estimator and next perform the mirror ascent on iteration points. By properly balancing the trade-off between the bias and cost of querying gradient oracles, this type of algorithm is guaranteed to find a near-optimal solution [155, 157, 158, 159].

We first discuss how to construct supgradient estimators of F. By Danskin's theorem [29],

$$\partial F(S) = \text{Conv}\Big\{ \sum_{i,j} \pi^*_{i,j}(S) M_{i,j} M_{i,j}^{\mathrm{T}} : \ \ \pi^*(S) \in \Pi(S) \Big\},$$

where $\Pi(S)$ denotes the set of optimal solutions to the following OT problem:

$$\min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} \langle M_{i,j} M_{i,j}^{\mathrm{T}}, S \rangle. \tag{3.11}$$

The main challenge of constructing a supgradient estimator is to compute an optimal solution $\pi^*(S) \in \Gamma(S)$. Since computing an exactly optimal solution is too expensive, we derive its near-optimal estimator, denoted as $\widehat{\pi}$, and practically use the following supgradient estimator:

$$v(S) = \sum_{i,j} \widehat{\pi}_{i,j} M_{i,j} M_{i,j}^{\mathrm{T}}. \tag{3.12}$$

We adopt the stochastic gradient-based algorithm with Katyusha momentum in [325] to compute a ϵ-optimal solution $\widehat{\pi}$ to (3.11). It achieves the state-of-the-art complexity $\widetilde{\mathcal{O}}\left(n^2 \epsilon^{-1}\right)$. See the detailed algorithm in Appendix B.6. Next, we describe the main algorithm for solving (SDR). Define the (negative) von Neumann entropy $h(S) = \sum_{i \in [2n]} \lambda_i(S) \log \lambda_i(S)$, where $\{\lambda_i(S)\}_i$ are the eignevalues of S, and define the von Neumann Bregman divergence

$$\begin{aligned} V(S_1, S_2) &= h(S_1) - h(S_2) - \langle S_1 - S_2, \nabla h(S_2)^{\mathrm{T}} \rangle \\ &= \text{Trace}(S_1 \log S_1 - S_1 \log S_2). \end{aligned}$$

Iteratively, we update S_{k+1} by performing mirror ascent with constant stepsize $\gamma > 0$:

$$S_{k+1} = \arg\max_{S \in \mathcal{S}_{2n}} \gamma \langle v(S_k), S \rangle + V(S, S_k),$$

Algorithm 3 Inexact Mirror Ascent for solving (SDR)

1: **Input:** Max iterations T, initial guess S_1, tolerance ϵ, constant stepsize γ.
2: **for** $k = 1, \ldots, T-1$ **do**
3: Obtain a ϵ-optimal solution (denoted as $\widehat{\pi}$) to (3.11)
4: Construct inexact supgradient $v(S_k)$ by (3.12)
5: Perform mirror ascent by (3.13)
6: **end for**
7: **Return** $\widehat{S}_{1:T} = \frac{1}{T}\sum_{k=1}^{T} S_k$

which admits the following closed-form update:

$$\widetilde{S}_{k+1} = \exp\left(\log S_k + \gamma v(S_k)\right), \quad S_{k+1} = \frac{\widetilde{S}_{k+1}}{\text{Trace}(\widetilde{S}_{k+1})}. \tag{3.13}$$

The general procedure for solving (SDR) is summarized in Algorithm 3.

3.4.2 Theoretical Analysis

In this subsection, we establish the complexity and performance guarantees for solving (SDR). Since the constraint set $\mathcal{S}_{2n}$ is compact and the objective in (SDR) is continuous, an optimal solution, denoted by S^*, is guaranteed to exist with a finite optimal value. A feasible solution $\widehat{S} \in \mathcal{S}_{2n}$ is said to be δ-optimal if it satisfies the condition $F(\widehat{S}) - F(S^*) \leq \delta$. Define the constant $C = \max_{i,j} \|M_{i,j}\|_2^2$.

Theorem 9 (Complexity Bound)**.** *Fix the precision $\delta > 0$ and specify hyper-parameters*

$$T = \left\lceil \frac{16C^2 \log(2n)}{\delta^2} \right\rceil, \qquad \epsilon = \frac{\delta}{4}, \qquad \gamma = \frac{\log(2n)}{C\sqrt{T}}.$$

Then, the complexity of Algorithm 3 for finding δ-optimal solution to (SDR) *is*

$$\mathcal{O}\left(C^3 n^3 (\log n)^{3/2} \delta^{-3}\right).$$

Next, we analyze the quality of the solution to (SDR). Recall the exact reformulation (3.9) requires that the optimal solution to be rank-1 while the tractable relaxation (SDR) does not

enforce such a constraint. Therefore, it is of interest to provide theoretical guarantees on the low-rank solution of (SDR), i.e., we aim to find the smallest integer $k \geq 1$ such that there exists an optimal solution to (SDR) that is at most rank-k. The integer k is called a rank bound on (SDR), which is characterized in the following theorem.

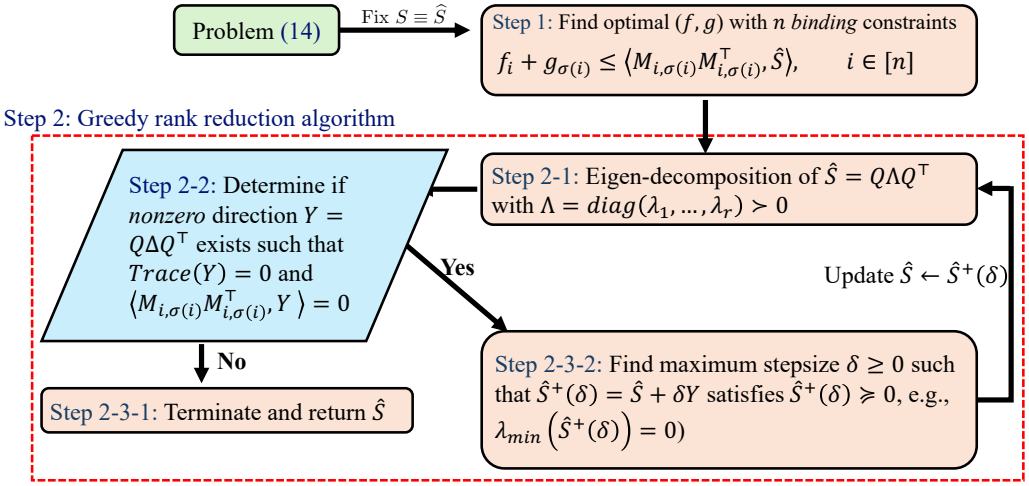

Figure 3.2: Diagram of the rank reduction algorithm. Here $\sigma(\cdot)$ denotes the permutation operator on $[n]$, Step 1 can be implemented using the Hungarian algorithm [184], and Step 2-2 finds a direction that lies in the null space of the constraint of Problem (3.14).

Theorem 10 (Rank Bound on (SDR))**.** *There exists an optimal solution to* (SDR) *of rank at most* $k \triangleq 1 + \left\lfloor \sqrt{2n + \frac{9}{4}} - \frac{3}{2} \right\rfloor$*. As a result,*

$$\begin{aligned} \mathit{Optval}(3.9) &= \max_{S \in \mathbb{S}_+^{2n}, \mathit{Trace}(S)=1, \mathit{rank}(S)=1} F(S) \\ &\leq \mathit{Optval}(\mathrm{SDR}) \leq \max_{S \in \mathbb{S}_+^{2n}, \mathit{Trace}(S)=1, \mathit{rank}(S)=k} F(S). \end{aligned}$$

The trivial rank bound on (SDR) should be $2n$, as the matrix S is of size $2n \times 2n$. Theorem 10 provides a novel rank bound that is significantly smaller than the trivial one.

Proof. Proof Sketch of Theorem 10. We first reformulate (SDR) by taking the dual of the

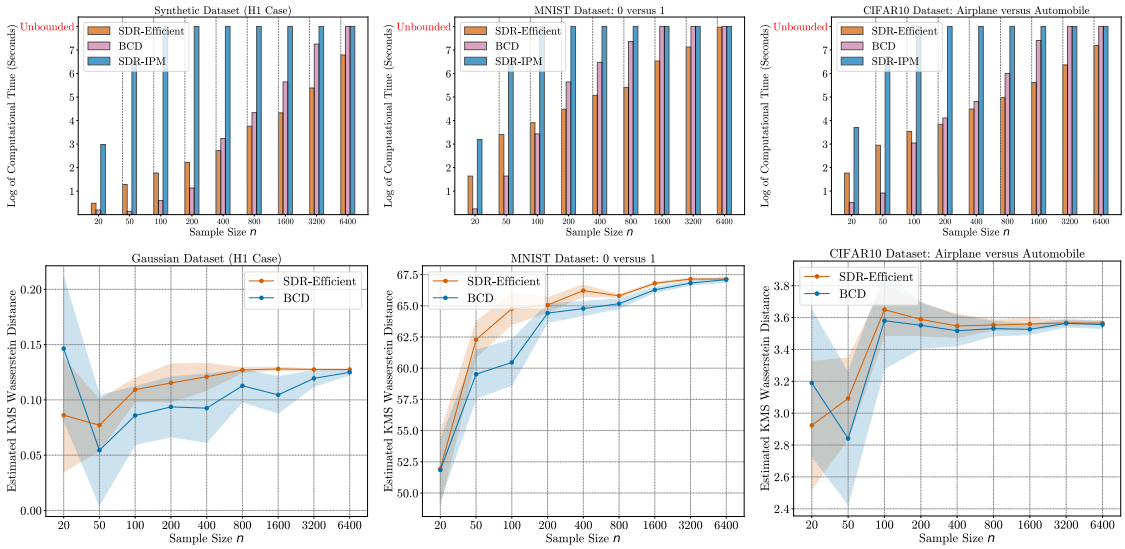

Figure 3.3: Comparison of `SDR-Efficient` with the baseline methods `SDR-IPM` and `BCD` in terms of computational time and solution quality. The columns, from left to right, correspond to the synthetic Gaussian dataset (100-dimensional), `MNIST`, and `CIFAR-10`. The top plots display the computational time, where the y-axis is labeled as "unbounded" if the running time exceeds the 1-hour time limit. The bottom plots present the estimated KMS 2-Wasserstein distance for each method.

inner OT problem:

$$\max_{\substack{S\in\mathcal{S}_{2n}\\ f,g\in\mathbb{R}^n}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(f_i+g_i) :\ f_i+g_j \le \langle M_{i,j}M_{i,j}^{\mathrm{T}}, S\rangle, \forall i,j \right\}. \tag{3.14}$$

By Birkhoff's theorem [37] and complementary slackness of OT, there exists an optimal solution of (f,g) such that at most n constraints of (3.14) are binding, and with such an optimal choice, one can adopt the convex geometry analysis from [195, 198] to derive the desired rank bound for any feasible extreme point of variable S. As the set of optimal solutions of (SDR) has a non-empty intersection with the set of feasible extreme points, the desired result holds. □ □

It is noteworthy that Algorithm 3 only finds a near-optimal solution $\widehat{S}_{1:T}$ of (SDR), which is not guaranteed to satisfy the rank bound in Theorem 10. To fill the gap, we develop a rank-reduction algorithm that further converts $\widehat{S}_{1:T}$ to the feasible solution that maintains the desired rank bound. See the general diagram that outlines this algorithm in Fig. 3.2 and

the detailed description in Appendix B.10. We also provide its complexity analysis in the following theorem, though in numerical study the complexity is considerably smaller than the theoretical one.

Theorem 11. *The rank reduction algorithm in Fig. 3.2 satisfies that* (I) *for a δ-optimal solution to (SDR), it outputs another δ-optimal solution with rank at most k;* (II) *its worst-case complexity is $\mathcal{O}(n^5)$.*

Additionally, by adopting the proof from Luo et al. [210], we show the optimality gap guarantee in Theorem 12. Although the approximation ratio seems overly conservative, we find (SDR) has good numerical performance.

Theorem 12 (Relaxation Gap of (SDR))**.** *Denote by $\varepsilon = 4 \cdot (0.33)^3$ an universal constant. Then*

$$\varepsilon n^{-4} \cdot \mathrm{Optval}(\mathrm{SDR}) \leq \mathrm{Optval}(\mathbf{KMS}) \leq \mathrm{Optval}(\mathrm{SDR}).$$

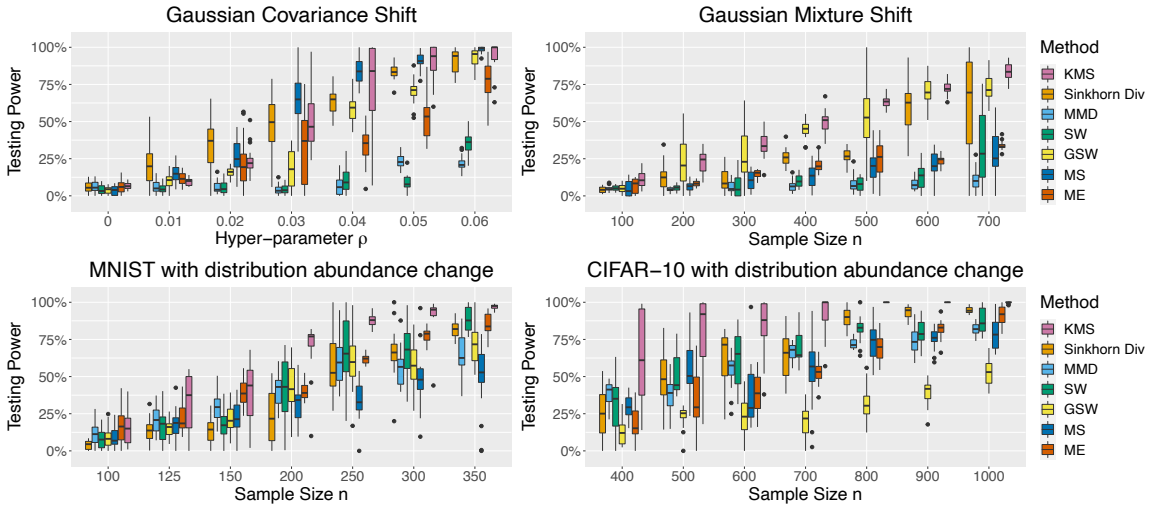

Figure 3.4: Testing power with a controlled type-I error rate of 0.05 across four datasets. Figures from left to right correspond to (a) Gaussian covariance shift, (b) Gaussian mixture distribution shift, (c) MNIST, and (d) CIFAR-10 with distribution abundance changes.

3.5 Numerical Study

This section presents experiment results for KMS 2-Wasserstein distance that is solved using SDR with first-order algorithm and rank reduction (denoted as `SDR-Efficient`). Baseline approaches include the block coordinate descent (`BCD`) algorithm [307], which finds stationary points of KMS 2-Wasserstein, and using interior point method (IPM) by off-the-shelf solver `cvxpy` [103] for solving SDR relaxation (denoted as `SDR-IPM`). Each instance is allocated a maximum time budget of one hour. All experiments were conducted on a MacBook Pro with an Intel Core i9 2.4GHz and 16GB memory. Unless otherwise stated, error bars are reproduced using 20 independent trials. Throughout the experiments, we specify the kernel as Gaussian, with the bandwidth being the median of pairwise distances between data points. Other details and extra numerical studies can be found in Appendices B.11 and B.12.

Computational Time and Solution Quality. We first compare our approach to baseline methods in terms of running time and solution quality. The quality of a given nonlinear projector is assessed by projecting testing data points from two groups and calculating their 2-Wasserstein distance. The experimental results, shown in the top of Fig. 3.3, indicate that for small sample size instances, `SDR-IPM` requires significantly more time than the other two approaches. Additionally, for small sample size instances, the running time of `BCD` is slightly shorter than that of `SDR-Efficient`. However, for larger instances, `BCD` outperforms `SDR-Efficient` in terms of running time. This observation aligns with our theoretical analysis, which shows that the complexity of `SDR` is $\tilde{\mathcal{O}}(n^2\delta^{-3})$, lower than the complexity of `BCD` [307], which is $\tilde{\mathcal{O}}(n^3\delta^{-3})$. The plots in the bottom of Fig. 3.3 show the quality of methods `SDR-Efficient` and `BCD`. We find the performance of solving SDR outperforms `BCD`, as indicated by its larger means and smaller variations. One possible explanation is that `BCD` is designed to find a local optimum solution for the original non-convex problem, making it highly sensitive to the initial guess and potentially less effective

in achieving optimal performance.

High-dimensional Hypothesis Testing. Then we validate the performance of KMS 2-Wasserstein distance for high-dimensional two-sample testing using both synthetic and real datasets. Baseline approaches include the two-sample testing with other statistical divergences, such as (i) Sinkhorn divergence (`Sinkhorn Div`) [130], (ii) maximum mean discrepancy with Gaussian kernel and median bandwidth heuristic (`MMD`) [138], (iii) sliced Wasserstein distance (`SW`) [55], (iv) generalized-sliced Wasserstein distance (`GSW`) [177], (v) max-sliced Wasserstein distance (`MS`) [97], and (vi) optimized mean-embedding test (`ME`) [170]. The baselines `KMS, MS, ME` partition data into training and testing sets, learn parameters from the training set, and evaluate the testing power on the testing set. Other baselines utilize both sets for evaluation. Synthetic datasets include the high-dimensional Gaussian distributions with covariance shift, or Gaussian mixture distributions. Real datasets include the MNIST [96] and CIFAR-10 [179] with changes in distribution abundance. The type-I error is controlled within 0.05 for all methods.

We report the testing power for all these approaches and datasets in Fig. 3.4. For the Gaussian covariance shift scenario the MS method achieves the best performance, which can be explained by the fact that linear mapping is optimal to separate the high-dimensional Gaussian distributions. For the other three scenarios, our approach has the superior performance compared with those baselines.

Table 3.1: Detection delay of various methods with controlled false alarm rate $\alpha = 0.05$. Mean and standard deviation (std) are calculated based on data from 10 different users.

Method	KMS	Sinkhorn Div	MMD	SW	GSW	MS	ME
Mean	**11.4**	16.5	50.6	17.2	12.9	17.8	65.4
Std	5.56	**4.4**	39.5	8.7	6.4	9.2	25.7

Human Activity Detection. We evaluate the performance of the KMS Wasserstein distance in detecting human activity transitions as quickly as possible using MSRC-12 Kinect gesture dataset [115]. After preprocessing, the dataset consists of 10 users, each with 80 attributes,

performing the action throwing/lifting before/after the change-point at time index 600. We employ a sliding window approach [322] with a false alarm rate of 0.01 to construct the test statistic at each time index, which increases significantly when a change-point is detected. Figure 3.5 illustrates the test statistics generated by our method compared to baseline approaches. The experimental results, summarized in Table 3.1, show that our method achieves superior performance.

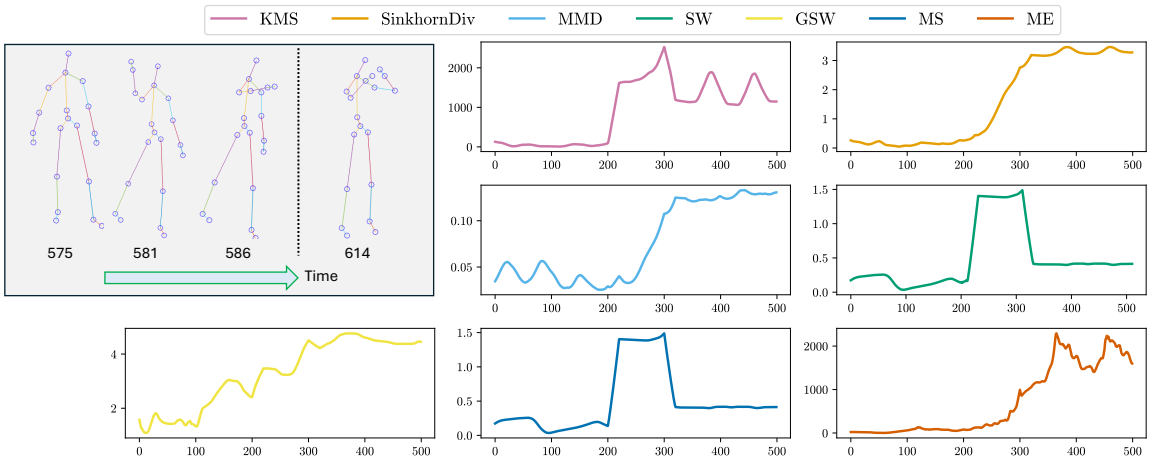

Figure 3.5: Top Left: Illustration of sequential data before and after the change-point. Remaining: Testing statistics computed from our and baseline approaches.

Generative Modeling. Finally, we replace the MS Wasserstein distance with our proposed KMS Wasserstein distance for the generative modeling task on the MNIST dataset, following a similar procedure as in [97]. Figure 3.6 shows the generated samples produced by generative models based on either the sliced Wasserstein distance or the KMS Wasserstein distance. To evaluate the performance of these models, we use the Fréchet Inception Distance (FID) score [147], calculated by extracting image features with a three-layer convolutional neural network. A lower FID score indicates better generative performance. Compared to the baseline, the KMS Wasserstein-based generative model achieves a lower FID score and produces higher-quality images, suggesting that the KMS Wasserstein distance is effective for learning high-dimensional data distributions.

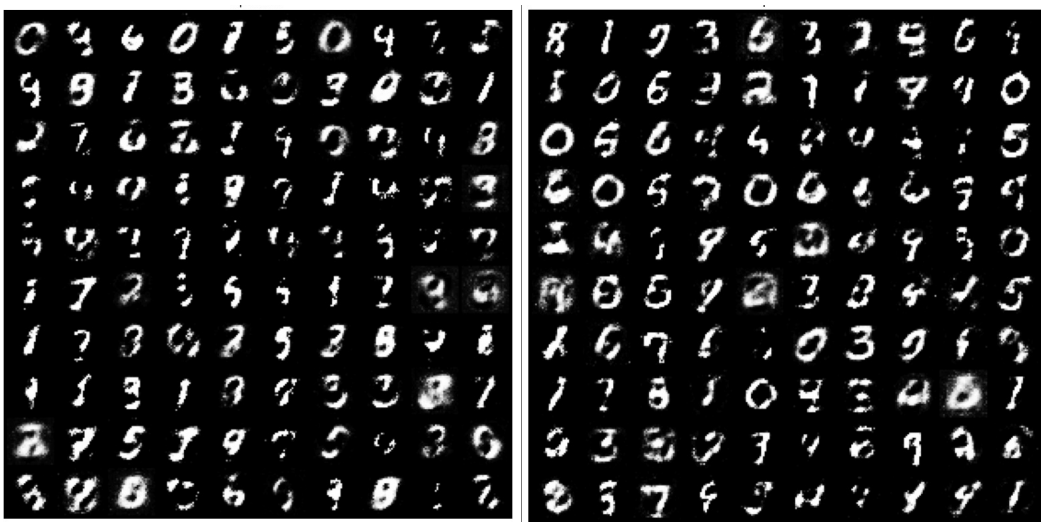

Figure 3.6: Left: samples generated from sliced Wasserstein-based generative models with FID score `5.37e-2`; Right: samples generated from KMS Wasserstein-based generative models with FID score `3.35e-2`. Models are trained with feed-forward neural-nets and 30 epoches.

3.6 Concluding Remarks

In this paper, we presented both statistical and computational guarantees for the KMS Wasserstein distance. From a statistical perspective, we derived finite-sample guarantees under mild conditions on the kernel function, yielding a dimension-free concentration error bound. Moreover, our analysis relaxes the compactness assumption on the sample space typically required by the MS Wasserstein distance. From a computational perspective, we investigated the SDR approach for solving the empirical KMS 2-Wasserstein distance. This method is not only computationally efficient using first-order optimization techniques but also enjoys strong performance guarantees for the obtained solutions. Our numerical experiments validated the theoretical results and demonstrated the superior performance of the KMS Wasserstein distance in high-dimensional hypothesis testing, human activity detection, and generative modeling applications.

CHAPTER 4

VARIABLE SELECTION FOR KERNEL TWO-SAMPLE TESTS

In this chapter, we consider the variable selection problem for two-sample tests, aiming to select the most informative variables to determine whether two collections of samples follow the same distribution. To address this, we propose a novel framework based on the kernel maximum mean discrepancy (MMD). Our approach seeks a subset of variables with a pre-specified size that maximizes the variance-regularized kernel MMD statistic. We focus on three commonly used types of kernels: linear, quadratic, and Gaussian. From a computational perspective, we derive mixed-integer programming formulations and propose exact and approximation algorithms with performance guarantees to solve these formulations. From a statistical viewpoint, we derive the rate of testing power of our framework under appropriate conditions. These results show that the sample size requirements for the three kernels depend crucially on the number of selected variables, rather than the data dimension. Experimental results on synthetic and real datasets demonstrate the superior performance of our method, compared to other variable selection frameworks, particularly in high-dimensional settings. This work is mainly summarized in [304].

4.1 Introduction

Two-sample test is a classic problem in statistics and an important tool for scientific discovery. Given two sets of observations $\mathbf{x}^n := \{x_i\}_{i=1}^n$ and $\mathbf{y}^n = \{y_i\}_{i=1}^n$[1], which represent n independent and identically distributed (i.i.d.) D-dimensional samples from distributions μ and ν, respectively. Using these samples, we aim to decide whether μ and ν are distinct. Two-sample test has wide applications: for example, in clinical trials to evaluate the effectiveness

[1] In this paper, we consider the simplified setting where the two sets of samples are the same size; it can be generalized to the setting when the two sets of samples do not have the same size.

of two distinct treatments on patient outcomes; in finance, to compare the performance of two different investment strategies; and in machine learning, to investigate whether the source domain and target domain have significant differences.

Recently, kernel two-sample tests have become a popular approach for modern high-dimensional data (see, e.g., [76, 138]). Despite the vast literature on kernel two-sample tests, studying variable selection for two-sample testing remains relatively limited. In this context, variable selection seeks the most informative d variables from a pool of D variables (usually $d \ll D$) to differentiate distributions μ and ν. On the one hand, finding interpretable variables is crucial for understanding population differences for domains such as gene expression analysis, where only a small subset of variables elucidates disparities between normal and abnormal populations [334]. On the other hand, the dissimilarities between high-dimensional datasets often exhibit a low-dimensional structure [303], and thus, extracting a small set of crucial variables as a pre-processing step may enhance the efficacy of high-dimensional two-sample testing.

Variable selection in the kernel two-sample testing is very different from the widely studied variable selection problem in linear models (notably, Lasso) and generalized linear models (see, e.g., [144]) because we face completely different objective functions in the optimization formulation. An interesting aspect of the problem is that depending on the choice of the kernel, the nature of the optimization objective can range from simple to hard. Moreover, the original formulation of the variable selection problem will also lead to the so-called "subset selection problem" [222], which leads to an integer program and can be hard to solve directly; it remains a question of how to develop computationally efficient procedures and approximate algorithms. On the other hand, for analyzing statistical performance variable selection for kernel two-sample tests facing finite samples, we need to study the statistical performance for tests (such as the false-detection and detection power), which is characteristic of testing problems that are different from regression type of prediction problems.

In this paper, we provide a novel variable selection framework for two-sample testing by choosing key variables that maximize the variance-regularized kernel maximum mean discrepancy (MMD) statistic, which in turn (approximately) maximizes the corresponding testing power. Our focus is on three types of kernels: linear, quadratic, and Gaussian. The contributions are summarized as follows.

- From the computational perspective, we leverage mixed-integer programming techniques to solve the MMD optimization problem for variable selection. For linear kernel, we reformulate the optimization as an *inhomogeneous* quadratic maximization with ℓ_2 and ℓ_0 norm constraints (see Section 4.4.1), called **S**parse **T**rust **R**egion **S**ubproblem (STRS). Despite its NP-hardness, we provide an exact mixed-integer semi-definite programming formulation together with exact and approximation algorithms for solving this problem. To the best of our knowledge, this study is new in the literature. For quadratic and Gaussian kernels, the MMD optimization becomes a sparse maximization of a non-concave function (see Section 4.4.1), which is intractable in general. We propose a heuristic algorithm that iteratively optimizes a quadratic approximation of the objective function, which is also a special case of STRS.

- From the statistical perspective, we derive the rate of testing power of our framework under appropriate conditions. We demonstrate that when the training sample size n_{Tr} is sufficiently large, the type-II error decays in the order of $n_{\mathrm{Te}}^{-1/2}$, where n_{Te} denotes the testing sample size. For the three focused types of kernels, the training sample size requirement is almost independent of the data dimension D but dependent on the number of selected variables d: For linear, quadratic, and Gaussian kernels, to achieve satisfactory performance, the training sample sizes are at least $\Omega(d^2 \log \frac{D}{d}), \Omega(d^4 \log \frac{D}{d})$, and $\Omega(d \log \frac{D}{d})$, respectively.

- By combining both viewpoints, it becomes evident that there exists a balance between

computational tractability and statistical guarantees. While the Gaussian kernel requires a smaller sample size to be statistically powerful, the corresponding MMD optimization is challenging. Conversely, the linear or quadratic kernel may require more samples to be statistically powerful, but the optimization is easier to solve.

- Finally, we perform numerical experiments using synthetic and real datasets to showcase the effectiveness of our proposed framework compared to other baseline methods. We utilize synthetic datasets to evaluate the testing power and variable selection performance. Subsequently, we test our approach on real data, including the standard MNIST handwritten digits and more specialized sepsis detection datasets.

Notations. Given a positive integer n, define $[n] = \{1, \ldots, n\}$. Let $\mathbb{F} = \{0, 1\}$, and $\mathbb{S}_n^+$ denote the collection of $n \times n$ symmetric positive semi-definite matrices. Given a vector $z \in \mathbb{R}^D$ and a set $S \subseteq [D]$, we use $z^{(k)}$ denote the k-th entry in z, and $z^{(S)}$ to denote the subvector with entries indexed by S. Given an $m \times n$ matrix A and two sets $S \subseteq [m], t \subseteq [n]$, denote $A^{(i,j)}$ the (i, j)-th entry in A and denote $A^{(S,T)}$ as the submatrix with rows and columns indexed by S and T. Given a vector $z \in \mathbb{R}^D$ and a distribution μ in $\mathbb{R}^D$, denote $z \circ \mu$ as the distribution of the random variable $\sum_{k\in[D]} z^{(k)}x^{(k)}$ provided that $x \sim \mu$. Define the norm $\|z\|_{(d)} = \max_{S:\,|S|\le d} \left\|z^{(S)}\right\|_2$. For a D-dimensional distribution μ and $s \in [D]$, let $\mathrm{Proj}_{s\#}\mu$ be the s-th marginal distributions of μ.

Related work

Variable selection. Classical variable selection approaches seek to extract the most valuable features from a group of high-dimensional data points. To name a few, sparse PCA seeks to select crucial variables that maximize the sample covariance based on sample sets [100, 101, 194]; truncated SVD aims to formulate a low-rank data matrix with minimum approximation error [197], and the maximum entropy sampling or experiment design aims to select a subgroup of samples that reserve information as much as possible [193, 196]. However, existing literature has paid less attention to variable selection for identifying differences

between the two groups except [52, 142, 165, 166]. The references therein mainly rely on parametric assumptions regarding the data-generating distributions. In particular, Taguchi and Rajesh [293] assume target distributions are Gaussian and find important variables such that the difference between mean and covariance among two groups is maximized. The works [142, 165, 166] further model distributions as Gaussian graphical models and detect the difference between distributions in correlation and partial correlation. However, it is undesirable to restrict the analysis to parametric distributions, especially for real-world applications. Bonferroni method [52] has been proposed in the two-sample testing context to compare every single feature using statistical tests to obtain representative variables. Still, it may not perform well when correlations exist between variables.

Kernel two-sample tests. A popular approach for non-parametric two-sample testing is based on kernel methods [269]: such tests quantify the difference of probability distributions by measuring the difference in *kernel mean embedding* [28, 227], which is also called the maximum mean discrepancy (MMD) [118, 138, 175, 270, 271]. The follow-up works [202, 291] further improve the performance of kernel-based two-sample tests by selecting kernels that maximize the variance-normalized empirical MMD. We adopt this idea in our variable selection framework. However, we observe that using this criterion for variable selection results in a fractional program subject to sparsity and norm constraints, which is highly challenging to solve. Hence, we are inspired to consider optimizing the variance-regularized empirical MMD statistic as a surrogate (see (4.12)).

Other two-sample tests. Many widely-used frameworks employ classification techniques for two-sample testing (see, e.g., [75, 181, 182]). It is worth noting that our approach adopts a distinct framework compared to those references: these aforementioned testing methods may not effectively identify interpretable variables capable of distinguishing between two distributions. One potential alternative is to employ a classifier based on sparse logistic regression [33] to construct a two-sample test. However, this approach may not yield satisfactory performance due to the limited flexibility of the parametric form of the classifier,

as we will demonstrate in Section 4.3.2. Recently, Mueller and Jaakkola [228] proposed to find the optimal subset of features such that the Wasserstein distance between projected distributions in dimension $d = 1$ is maximized. Later Wang et al. [306, 307] modified the projection function as the linear mapping with general dimension $d > 1$ and nonlinear mapping, respectively, thus improving the flexibility of dimensionality reduction and power of two-sample testing. Nevertheless, these references do not impose sparsity constraints when performing dimensionality reduction. Therefore, they cannot select a subset of variables that differentiate the differences between the two groups.

4.2 Background

We first present some background information about maximum mean discrepancy (MMD). It measures the discrepancy between two probability distributions by employing test functions within a reproducing kernel Hilbert space (RKHS), which has been commonly used in two-sample testing area [80, 138, 139, 140, 170, 202].

Definition 9 (Maximum Mean Discrepancy)**.** *A kernel function $K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is called a positive semi-definite kernel if for any finite set of n samples $\{x_i\}_{i=1}^n$ in $\mathbb{R}^D$ and $\{c_i\}_{i=1}^n$ in $\mathbb{R}$, it holds that $\sum_{i\in[n]}\sum_{j\in[n]} c_i c_j K(x_i, x_j) \geq 0$. A positive semi-definite kernel K induces a unique RKHS $\mathcal{H}$. Given $\mathcal{H}$ containing a class of candidate testing functions and two distributions μ, ν, define the corresponding MMD statistic as*

$$\mathrm{MMD}(\mu, \nu; K) \triangleq \sup_{f\in\mathcal{H}, \|f\|_{\mathcal{H}}\leq 1} \Big\{\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\Big\}. \tag{4.1}$$

Leveraging reproducing properties of the RKHS, the MMD statistic can be equivalently written as

$$\mathrm{MMD}^2(\mu, \nu; K) = \mathbb{E}_{x,x'\sim\mu}[K(x, x')] + \mathbb{E}_{y,y'\sim\nu}[K(y, y')] - 2\mathbb{E}_{x\sim\mu, y\sim\nu}[K(x, y)],$$

which enables convenient computation and sample estimation. When the distributions μ and ν are not available, one can formulate an estimate of $\mathrm{MMD}^2(\mu, \nu; K)$ based on samples $\mathbf{x}^n$ and $\mathbf{y}^n$ using the following statistic [138]:

$$\widehat{\mathrm{MMD}}^2(\mathbf{x}^n, \mathbf{y}^n; K) \triangleq \frac{1}{n(n-1)} \sum_{i \in [n], j \in [n], i \neq j} H_{i,j}, \tag{4.2}$$

with

$$H_{i,j} := K(x_i, x_j) + K(y_i, y_j) - K(x_i, y_j) - K(y_i, x_j). \tag{4.3}$$

The choice of kernel function largely influences the performance of variable selection for two-sample tests. To achieve satisfactory performance, we consider the following types of kernel functions, denoted as $K_z(\cdot, \cdot)$. Here, the coefficient vector $z = (z^{(s)})_{s \in [D]}$ involved in the kernel functions determines which variables to be selected, which is in the domain set

$$\mathcal{Z} := \{z \in \mathbb{R}^D : \|z\|_2 = 1, \|z\|_0 \leq d\}. \tag{4.4}$$

- **Linear Kernel**: For each coordinate $s \in [D]$, we specify the scalar-input kernel $k_s : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and then construct

$$K_z(x, y) = \sum_{s \in [D]} z^{(s)} k_s\big(x^{(s)}, y^{(s)}\big). \tag{4.5}$$

  Those scalar-input kernels $k_s(\cdot, \cdot), s \in [D]$ defined above are used to compare the difference of distributions among each coordinate, which can be chosen as the Gaussian kernel with certain bandwidth hyper-parameter τ_s^2, i.e., $k_s(x, y) = e^{-(x-y)^2/(2\tau_s^2)}$.

- **Quadratic Kernel**: For each coordinate $s \in [D]$, we specify the scalar-input kernel

$k_s : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and then construct

$$K_z(x, y) = \left(\sum_{s \in [D]} z^{(s)} k_s\big(x^{(s)}, y^{(s)}\big) + c \right)^2 . \tag{4.6}$$

Here $c \geq 0$ is a bandwidth hyper-parameter of the quadratic kernel, and scalar-input kernels $k_s(\cdot, \cdot), s \in [D]$ can be chosen in the same way as defined in the linear kernel case.

- **(Isotropic) Gaussian Kernel**: We first specify the bandwidth hyper-parameter $\sigma^2 > 0$ and then construct isotropic Gaussian-type kernel

$$K_z(x, y) = \exp\left(-\frac{\sum_{s \in [D]} \big(z^{(s)}(x^{(s)} - y^{(s)})\big)^2}{2\sigma^2} \right). \tag{4.7}$$

We use isotropic Gaussian kernel, a common choice in the literature on two-sample testing, primarily because it only involves a bandwidth hyper-parameter that is easy to tune.

4.3 Formulation

A natural criterion of variable selection is to pick the coefficient vector z such that the empirical MMD statistic is optimized, i.e., it suffices to solve the formulation

$$\max_{z \in \mathcal{Z}} \ \widehat{\mathrm{MMD}}^2(\mathbf{x}^n, \mathbf{y}^n; K_z). \tag{4.8}$$

To motivate this formulation, we start with an example showcasing the nature of the problem: the complexity of the problem depends on the choice of the kernel, and the simplest linear kernel leads to the analytical solution. Despite that, the linear kernel is known to have limited testing power, and one usually prefers to use a non-linear kernel, whose solution is not analytical. Moreover, the test statistic may require normalization by standard deviation,

for which case an analytical solution is not available; all these require further algorithm development of algorithm as presented in Section 4.4.

Consider Problem (4.8) when the kernel function K_z is a linear kernel in (4.5). By straightforward calculation, it can be reformulated as the following linear optimization on the domain set $\mathcal{Z}$:

$$\max_{z \in \mathcal{Z}} \ z^{\mathrm{T}} a, \tag{4.9}$$

where for $s \in [D]$, the s-th entry of the vector $a \in \mathbb{R}^D$ is

$$a^{(s)} = \widehat{\mathrm{MMD}}^2 \Big(\{x_i^{(s)}\}_{i \in [n]}, \{y_i^{(s)}\}_{i \in [n]}, k_s \Big).$$

It is easy to check that the optimal solution to (4.9) is obtained by taking the non-zero indexes of z to be the indexes of the d largest absolute values of the vector a. In other words, based on the linear kernel and the criterion (4.8), our framework selects those variables whose MMD discrepancy between two distributions at the corresponding coordinates is as large as possible. Since this idea only utilizes the information of marginal distributions of a high-dimensional distribution in each coordinate, the linear kernel has limited testing power in practice.

4.3.1 Variance regularized MMD optimization

Although the idea behind the formulation (4.8) is simple, as pointed out in existing literature [140, 182, 291], directly optimizing the MMD statistic does not result in a powerful two-sample test in practice. Inspired by these references, we incorporate the variance statistic of MMD in the formulation to achieve more competitive performance.

Specifically, we pick the sparse selection vector z to achieve the most powerful test. Since the kernel function leading to the most powerful two-sample test approximately maximizes the MMD testing statistic normalized by its standard deviation [291], the natural idea is to

pick the selection vector z that solves the following fractional optimization problem:

$$\max_{z\in\mathcal{Z}} \frac{\widehat{\mathrm{MMD}}^2(\mathbf{x}^n,\mathbf{y}^n;K_z)}{\left(\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z)\right)^{1/2}}, \tag{4.10}$$

where $\widehat{\mathrm{MMD}}^2(\mathbf{x}^n,\mathbf{y}^n;K_z)$ and $\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z)$ are unbiased empirical estimators of the population testing statistic and the variance of testing statistic under alternative hypothesis $\mathcal{H}_1:\ \mu\neq\nu$, respectively. For fixed samples $\mathbf{x}^n,\mathbf{y}^n$ and kernel function $K(\cdot,\cdot)$, by [291], the variance estimator

$$\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K)=\frac{4}{n^3}\sum_{i\in[n]}\left(\sum_{j\in[n]}H_{i,j}\right)^2-\frac{4}{n^4}\left(\sum_{i\in[n]}\sum_{j\in[n]}H_{i,j}\right)^2, \tag{4.11}$$

where $H_{i,j},i,j\in[n]$ are defined in (4.3). Since the optimization over a fraction in (4.10) is difficult to handle, we introduce a regularization hyper-parameter $\lambda>0$ (which can be tuned by cross-validation in practice) and propose to solve the new optimization problem instead:

$$\max_{z\in\mathcal{Z}} \quad \left\{\widehat{F}(z;\mathbf{x}^n,\mathbf{y}^n):=\widehat{\mathrm{MMD}}^2(\mathbf{x}^n,\mathbf{y}^n;K_z)-\lambda\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z)\right\}. \tag{4.12}$$

The rationale behind problem (4.12) is that, by properly tuning the regularizer $\lambda>0$, we balance the trade-off between maximizing the testing statistic and minimizing its variance, which amounts to approximately optimizing the testing power criteria in (4.10).

The hyper-parameter λ is tuned using the following cross-validation procedure: We take a set of candidate choices of λ, denoted as $\{0.1,0.5,1,2,5\}$, and split the dataset into 50% training and 50% validation datasets. For each choice of hyper-parameters, we use the training dataset to obtain the optimal coefficient vector 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 collections of samples in the validation dataset. We finally

Algorithm 4 A permutation two-sample test using MMD with variable selection

Require: cardinality d, type-I error threshold α_{level}, bootstrap size N_p, collected samples $\mathbf{x}^n$ and $\mathbf{y}^n$.

1: Split data as $\mathbf{x}^n = \mathbf{x}^{\text{Tr}} \cup \mathbf{x}^{\text{Te}}$ and $\mathbf{y}^n = \mathbf{y}^{\text{Tr}} \cup \mathbf{y}^{\text{Te}}$.
2: Solve (4.12) with input data $(\mathbf{x}^{\text{Tr}}, \mathbf{y}^{\text{Tr}})$ to obtain optimal sparse selection vector z^*.
3: Compute test statistic $T = \widehat{\text{MMD}}^2(\mathbf{x}^{\text{Te}}, \mathbf{y}^{\text{Te}}; K_{z^*})$.
4: **for** $i = 1, \ldots, N_p$ **do** {} Step 4-8: Decide threshold using bootstrap
5: Shuffle $\mathbf{x}^{\text{Te}} \cup \mathbf{y}^{\text{Te}}$ to obtain $\mathbf{x}^{\text{Te}}_{(i)}$ and $\mathbf{y}^{\text{Te}}_{(i)}$.
6: Compute test statistic for bootstrap samples $T_i = \widehat{\text{MMD}}^2(\mathbf{x}^{\text{Te}}_{(i)}, \mathbf{y}^{\text{Te}}_{(i)}; K_{z^*})$.
7: **end for**
8: $t_{\text{thres}} \leftarrow (1 - \alpha_{\text{level}})$-quantile of $\{T_i\}_{i \in [N_p]}$.
9: Reject $\mathcal{H}_0$ (i.e., decide the two sample distributions are different) if $T > t_{\text{thres}}$.

specify the hyper-parameter as the one with the highest value of hold-out performance.

Using the proposed variable selection framework, we present a kernel two-sample test as follows. The data points are divided into training and testing datasets. Initially, the training set is utilized to obtain the selection coefficient that optimally identifies the differences between the two groups. Next, a permutation test is performed on the testing data points, projected based on the trained selection coefficient. The threshold for this permutation test is calibrated by bootstrapping following [138, Section 5] and [8]. The detailed algorithm is presented in Algorithm 4. This test is guaranteed to control the type-I error [136] because we evaluate the p-value of the test via the permutation approach. In the following sections, we discuss how to solve the optimization problem (4.12) with linear and quadratic kernels, respectively. In the subsequent sections, we develop algorithms for solving the MMD optimization problem and then establish statistical testing power guarantees for our proposed framework.

4.3.2 Connections with classification-based testing

It is worth mentioning that any method for classification can be applied to two-sample testing: Given training samples $\mathbf{x}^{\text{Tr}} = \{x_i\}_{i \in [|\mathbf{x}^{\text{Tr}}|]}$ and $\mathbf{y}^{\text{Tr}} = \{y_i\}_{i \in [|\mathbf{y}^{\text{Tr}}|]}$, we formulate the feature-label pairs as $\mathcal{D}_{\text{Tr}} = \{(x_i, 0)\}_{i \in [|\mathbf{x}^{\text{Tr}}|]} \cup \{(y_i, 1)\}_{i \in [|\mathbf{y}^{\text{Tr}}|]}$. One can use

any classification method to obtain a classifier $\widehat{f}$ based on training dataset $\mathcal{D}_{\mathrm{Tr}}$. After that, the testing statistic based on testing samples $\mathbf{x}^{\mathrm{Te}}$ and $\mathbf{y}^{\mathrm{Te}}$ can be computed as $T = \frac{1}{|\mathbf{x}^{\mathrm{Te}}|}\sum_{x\in\mathbf{x}^{\mathrm{Te}}}\widehat{f}(x) - \frac{1}{|\mathbf{y}^{\mathrm{Te}}|}\sum_{y\in\mathbf{y}^{\mathrm{Te}}}\widehat{f}(y)$. If the testing statistic T is greater than a certain threshold, the null hypothesis $\mathcal{H}_0$ is rejected, and otherwise, it is accepted.

A notable existing variable selection approach for classification is the *sparse logistic regression* (SLR) [33], which uses the classifier $\widehat{f}(\cdot) = \langle\cdot, \beta\rangle$ for some sparse vector β. The coefficient vector β can be obtained by solving a sparse optimization problem using training dataset $\mathcal{D}_{\mathrm{Tr}}$, and its non-zero entries correspond to the selected variables that distinguish the differences between groups $\mathbf{x}^{\mathrm{Tr}}$ and $\mathbf{y}^{\mathrm{Tr}}$. Based on samples $\mathbf{x}^{\mathrm{Te}}$ and $\mathbf{y}^{\mathrm{Te}}$, SLR formulates the following testing statistic and rejects the null hypothesis if it exceeds a certain threshold: $T_{\mathrm{SLR}} = \frac{1}{|\mathbf{x}^{\mathrm{Te}}|}\sum_{x\in\mathbf{x}^{\mathrm{Te}}}\ \beta^{\mathrm{T}}x - \frac{1}{|\mathbf{y}^{\mathrm{Te}}|}\sum_{y\in\mathbf{y}^{\mathrm{Te}}}\ \beta^{\mathrm{T}}y$. Such an approach assumes a parametric assumption that the data distributions μ and ν are *linearly separable* since otherwise, the linear predictor may not achieve satisfactory performance.

In contrast, our proposed method can be viewed as a generalized classification-based testing, which consists of two phases:

(I) At the first phase, we choose a suitable kernel function $K(\cdot,\cdot)$ based on training data $\mathbf{x}^{\mathrm{Tr}}$ and $\mathbf{y}^{\mathrm{Tr}}$ that depends only on a small group of variables leading to satisfactory two-sample testing performance. Such a variable selection procedure makes our classification model more interpretable.

(II) At the second phase, we obtain the classifier (also called the witness function in [138, Section 2.3]), denoted as $\widehat{f}$, based on validation data $\mathbf{x}^{\mathrm{Te}}$ and $\mathbf{y}^{\mathrm{Te}}$:

$$\widehat{f}(z) \propto \frac{1}{|\mathbf{x}^{\mathrm{Te}}|}\sum_{x\in\mathbf{x}^{\mathrm{Te}}} K(x,z) - \frac{1}{|\mathbf{y}^{\mathrm{Te}}|}\sum_{y\in\mathbf{y}^{\mathrm{Te}}} K(y,z). \tag{4.13}$$

In comparison with the SLR framework, we replace the linear classifier with the kernel-based classifier, which is a more flexible and powerful choice. In the following, we provide an example demonstrating that our proposed framework can successfully select useful variables

to distinguish the difference between two groups, while the sparse logistic regression cannot finish this task.

Example 2 (Example when sparse logistic regression cannot identify variables)**.** *Consider the example where* $\mu = \mathcal{N}(0, I_D)$ *and* $\nu = \mathcal{N}(0, \mathrm{diag}((1+\epsilon)^2, 1, \ldots, 1))$ *with* $\epsilon > 0$*. Here, only the first coordinate can differentiate between* μ *and* ν*. When using the sparse logistic regression, it is clear that for any* β *satisfying* $\|\beta\|_0 \leq 1$*, it holds that the population version of testing statistic* $\mathbb{E}[T_{\mathrm{SLR}}] = 0$*. This indicates that sparse logistic regression may not achieve satisfactory performance in hypothesis testing or classification. In contrast, consider our proposed MMD framework with the linear kernel. For any* z *such that* $\|z\|_2 = 1, \|z\|_0 \leq 1$*, it holds that the population version of the objective in* (4.12) *achieves the unique optimal solution* $\widehat{z}$ *with* $\widehat{z}^{(1)} = 1$ *if the variance regularization* λ *is selected properly. Specifically, when* λ *is chosen to be smaller than a constant* $\bar{\lambda} > 0$*, our proposed MMD framework can always select the true useful variable.*[2]

[2]Here we take $\bar{\lambda} = \frac{\mathrm{MMD}^2(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1)}{\sigma^2_{\mathcal{H}_1}(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1)}$ to satisfy the desired result. Specifically, we provide closed-form expressions on those statistics in the following (see the proof in Appendix C.2):

$$
\begin{aligned}
A &\triangleq \mathrm{MMD}^2(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1) = \sqrt{\frac{\tau_1^2}{\tau_1^2+2}} + \sqrt{\frac{\tau_1^2}{\tau_1^2+2(1+\epsilon)^2}} - 2\sqrt{\frac{\tau_1^2}{\tau_1^2+1+(1+\epsilon)^2}},\\
B &\triangleq \sigma^2_{\mathcal{H}_1}(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1) = 4C - 4A^2,\\
C &\triangleq \sqrt{\frac{\tau_1^4}{(\tau_1^2+1)(3+\tau_1^2)}} + \sqrt{\frac{4\tau_1^4}{(\tau_1^2+2)(\tau_1^2+2(1+\epsilon)^2)}}\\
&\quad - \sqrt{\frac{16\tau_1^4}{2\tau_1^2+1+(1+\epsilon)^2+(1+\tau_1^2)((1+\epsilon)^+\tau_1^2)}} - \sqrt{\frac{16\tau_1^4}{(\tau_1^2+1+(1+\epsilon)^2)(\tau_1^2+2(1+\epsilon)^2)}}\\
&\quad + \sqrt{\frac{16\tau_1^4}{(\tau_1^2+(1+\epsilon)^2)(\tau_1^2+(1+\epsilon)^2+2)}} + \sqrt{\frac{\tau_1^4}{(\tau_1^2+(1+\epsilon)^2)(\tau_1^2+3(1+\epsilon)^2)}}.
\end{aligned}
$$

4.4 Algorithm

To prepare for solving the MMD optimization problem (4.12), we first introduce the following Sparse Trust Region Subproblem (STRS):

$$\max_{z\in\mathcal{Z}} \Big\{ z^{\mathrm{T}} A z + z^{\mathrm{T}} a \Big\}, \tag{STRS}$$

where the set $\mathcal{Z}$ is defined in (4.4) and (A, a) are problem coefficients. This problem extends the standard Trust Region Subproblem [84], since we require the decision variable to satisfy an extra sparse constraint. In Section 4.4.1, we will show that Problem (4.12) can be reformulated as a special case of (STRS) for linear kernels. For generic kernels, the optimization procedure for solving (4.12) involves solving the subproblem (STRS). Unfortunately, the set $\mathcal{Z} = \{z \in \mathbb{R}^D : \ \|z\|_2 = 1, \|z\|_0 \leq d\}$ in (STRS) involves ℓ_0-norm constraint, which typically leads to a mixed-integer program reformulation that is NP-hard to solve. This motivates us to provide efficient optimization algorithms to tackle this challenge in Section 4.4.2.

Without loss of generality, we assume $A \succeq 0$ in (STRS), since otherwise, we can rewrite the problem as $\max_{z\in\mathcal{Z}} \Big\{ z^{\mathrm{T}}(A - \lambda_{\min}(A) I_D) z + z^{\mathrm{T}} a \Big\} + \lambda_{\min}(A)$, where the shifted matrix $A - \lambda_{\min}(A) I_D \succeq 0$, where $\lambda_{\min}$ denotes the smallest eigenvalue of a matrix. It is worth mentioning that the problem (STRS) reduces to sparse PCA formulation when the coefficient vector $a = 0$ (that is, the linear term is zero), which has been studied extensively in the literature [27, 119, 194, 225]. However, the study for general vector a for the problem (STRS) is new. In Section 4.4.2, we discuss the exact and approximation algorithms for solving (STRS) with generic data matrix A and vector a.

4.4.1 MMD optimization with different kernels

In the following, we provide detailed algorithms for solving the MMD optimization problem (4.12) for various kernels considered in (4.5)-(4.7).

Linear kernel

For linear kernel defined in (4.5), one can verify that $H_{i,j} \in \mathbb{R}$ defined in (4.3) can be written as a linear function in terms of z:

$$H_{i,j} = \sum_{s\in[D]} z^{(s)} \Big[k_s(x_i^{(s)}, x_j^{(s)}) + k_s(y_i^{(s)}, y_j^{(s)}) - k_s(x_i^{(s)}, y_j^{(s)}) - k_s(y_i^{(s)}, x_j^{(s)}) \Big] = z^{\mathrm{T}} h_{i,j},$$

where we denote the D-dimensional vector

$$h_{i,j} = \Big\{ k_s(x_i^{(s)}, x_j^{(s)}) + k_s(y_i^{(s)}, y_j^{(s)}) - k_s(x_i^{(s)}, y_j^{(s)}) - k_s(y_i^{(s)}, x_j^{(s)}) \Big\}_{s\in[D]}. \tag{4.14}$$

Since the empirical MMD estimator $\widehat{\mathrm{MMD}}^2(\mathbf{x}^n, \mathbf{y}^n; K_z)$ is a linear combination of $\{H_{i,j}\}_{i,j}$ and the empirical variance estimator $\hat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n, \mathbf{y}^n; K_z)$ is a quadratic function in terms of $\{H_{i,j}\}_{i,j}$, it is clear that the MMD optimization problem (4.12) can be reformulated as the mixed-integer quadratic optimization problem (STRS), where the data matrix $A \in \mathbb{R}^{D\times D}$ and vector $a \in \mathbb{R}^D$ have the following expressions:

$$\begin{aligned} A^{(s_1,s_2)} = {} & \frac{4\lambda}{n^3} \sum_{i\in[n]} \left(\sum_{j\in[n]} h_{i,j}^{(s_1)} \right) \left(\sum_{j\in[n]} h_{i,j}^{(s_2)} \right) \\ & - \frac{4\lambda}{n^4} \left(\sum_{i,j\in[n]} h_{i,j}^{(s_1)} \right) \left(\sum_{i,j\in[n]} h_{i,j}^{(s_1)} \right), \ \forall s_1, s_2 \in [D], \\ a = {} & \frac{1}{n(n-1)} \sum_{i\in[n], j\in[n], i\neq j} h_{i,j}. \end{aligned}$$

Therefore, one can query either the exact or approximation algorithm to solve problem (4.12) with strong optimization guarantees for this linear kernel case. In the following remark, we discuss under which conditions will linear kernel MMD may or may not achieve satisfactory performance on the variable selection task.

Remark 8 (Limitation of Linear Kernel)**.** *Under the linear kernel choice, it can be shown*

that the population MMD statistic becomes

$$\mathrm{MMD}^2(\mu, \nu; K_z) = \sum_{s \in [D]} z^{(s)} \mathrm{MMD}^2(\textit{Proj}_{s\#}\mu, \textit{Proj}_{s\#}\nu; k_s),$$

where $\textit{Proj}_{s\#}\mu, \textit{Proj}_{s\#}\nu$ are the s-th marginal distributions of μ, ν, respectively. In other words, the selection coefficient z aims to find a direction to identify the difference between marginal distributions of μ and ν. However, under the case where marginal distributions of μ and ν are the same, the linear kernel MMD does not have enough power to find informative variables to distinguish those two distributions.

Other kernel choices

For other kernel choices, such as the quadratic kernel in (4.6) and Gaussian kernel in (4.7), the objective for MMD optimization is a nonlinear non-concave function with respect to z. This, together with the sparse constraint of the domain set $\mathcal{Z}$, makes this type of problem very challenging to solve. In this part, we provide a heuristic algorithm that incorporates simulated annealing (SA) [35] and STRS that tries to find a feasible solution of (4.12) with high solution quality. Such a heuristic can also be naturally extended for generic kernel choices.

Here, we outline our SA and STRS-based heuristics. For notational simplicity, we denote the objective of (4.12) as $F(z)$ instead. Our proposed algorithm is an iterative method that generates a trajectory of feasible solutions $z_1, \ldots, z_{i_{\max}}$. At the iteration point z_i, we generate a candidate solution $\tilde{z}_i$ by optimizing a second-order approximation of the objective $F(z)$ with quadratic penalty regularization around z_i:

$$\begin{aligned} \tilde{z}_i = \arg\max_{z \in \mathcal{Z}} \Big\{ & F(z_i) + \nabla F(z_i)^{\mathrm{T}}(z - z_i) \\ & + \frac{1}{2}(z - z_i)^{\mathrm{T}} \nabla^2 F(z_i)(z - z_i) - \frac{\tau_i}{2}\|z - z_i\|_2^2 \Big\}, \end{aligned} \tag{4.15}$$

where τ_i denotes the quadratic regularization value. Such a problem is a special case

Algorithm 5 Heuristic algorithm for solving (4.12) with generic kernel

Input: Max iterations $i_{\max}$, initial guess z_1, initial temperature $\texttt{Tem}$, cooling parameter α, and a set of regularization values $\mathcal{G}$
for $i = 1, \ldots, i_{\max} - 1$ **do**
  Randomly pick the regularization value τ_i from $\mathcal{G}$.
  Obtain $\tilde{z}_i$ by solving a STRS in (4.15).
  Compute residual level $\Delta_i = F(\tilde{z}_i) - F(z_i)$ and probability $p_i = e^{\Delta_i/\texttt{Tem}}$
  if $\mathrm{rand}(0, 1) < p_i$ **then**
    $z_{i+1} = \tilde{z}_i$
  else
    $z_{i+1} = z_i$
  end if
  $\texttt{Tem} = \alpha \cdot \texttt{Tem}$
end for
Return $z_{i_{\max}}$

of (STRS), where the data matrix $A \in \mathbb{R}^{D\times D}$ and vector $a \in \mathbb{R}^D$ have the following expressions:

$$A = \frac{1}{2}\nabla^2 F(z_i) - \frac{\tau_i}{2} I_D, \quad a = \nabla^2 F(z_i) z_i - \tau_i z_i + \nabla F(z_i).$$

Hence, Problem (4.15) can be solved by querying the exact or approximation algorithm described in Section 4.4. Let $\Delta_i = F(\tilde{z}_i) - F(z_i)$ denote the residual value for moving from z_i to $\tilde{z}_i$. The central idea of SA is always to accept moves with positive residual values while not forbidding moves with negative residual values. Specifically, we assign a certain temperature $\texttt{Tem}$, and update z_{i+1} as $\tilde{z}_i$ according to the probability

$$p_i = \begin{cases} 1, & \text{if } \Delta_i \geq 0 \\ e^{\Delta_i/\texttt{Tem}}, & \text{if } \Delta_i < 0. \end{cases}$$

If the candidate solution $\tilde{z}_i$ is not accepted, we update z_{i+1} as z_i. The temperature parameter $\texttt{Tem}$ is a critical hyper-parameter in this algorithm. We assign an initial value of $\texttt{Tem}$ and iteratively decrease it such that in the last iterations, the moves with worse objective values are less and less likely to be accepted. See our detailed algorithm procedure in Algorithm 5.

Finally, we add remarks regarding the tractability and flexibility of quadratic and Gaussian kernels.

Remark 9 (Quadratic kernel)**.** *For quadratic kernel defined in* (4.6)*, it can be shown that the population MMD is given by* $\mathrm{MMD}^2(\mu,\nu;K_z) = z^{\mathrm{T}}\mathcal{A}(\mu,\nu)z + z^{\mathrm{T}}\mathcal{T}(\mu,\nu)$*, where* $\mathcal{A}(\mu,\nu)$ *is a* $\mathbb{R}^{D\times D}$*-valued mapping such that*

$$\begin{aligned}(\mathcal{A}(\mu,\nu))^{(s_1,s_2)} = \mathbb{E}_{x,x'\sim\mu}[k_{s_1}(x^{(s_1)},x'^{(s_1)})k_{s_2}(x^{(s_2)},x'^{(s_2)})] \\ + \mathbb{E}_{y,y'\sim\nu}[k_{s_1}(y^{(s_1)},y'^{(s_1)})k_{s_2}(y^{(s_2)},y'^{(s_2)})] \\ - 2\mathbb{E}_{x\sim\mu,y\sim\nu}[k_{s_1}(x^{(s_1)},y^{(s_1)})k_{s_2}(x^{(s_2)},y^{(s_2)})],\end{aligned}$$

and $\mathcal{T}(\mu,\nu)$ *is a* $\mathbb{R}^D$*-valued mapping such that*

$$(\mathcal{T}(\mu,\nu))^{(s)} = 2c\mathrm{MMD}^2(\mathit{Proj}_{s\#}\mu, \mathit{Proj}_{s\#}\nu; k_s).$$

Given two multivariate distributions, the quadratic MMD aims to find a direction z *to distinguish the difference in each coordinate and the correlation between the two coordinates the most. Compared with the linear MMD, which only identifies the difference in each coordinate, the quadratic MMD is a more flexible choice. However, it can be shown that the objective in* (4.12) *with the quadratic kernel is a* 4*-th order non-concave monomial with respect to* z*, which is computationally intractable to optimize. In practical experiments, we use the heuristic algorithm in Algorithm 5 to obtain a reasonably high-quality solution.*

Remark 10 (Gaussian kernel)**.** *One can also re-write the population testing statistic for the Gaussian kernel defined in* (4.7)*. For notational simplicity, let* $K(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right)$ *be a standard Gaussian kernel with* d*-dimensional data, and define* $z_{\#}\nu$ *as a* d*-dimensional distribution such that*

$$z_{\#}\nu = \left(z^{(s)}x^{(s)}\right)_{s\in\mathrm{supp}(z)}, \quad \textit{where } x\sim\nu.$$

With these notations, it can be shown that the population MMD statistic becomes

$$\mathrm{MMD}^2(\mu, \nu; K_z) = \mathrm{MMD}^2(z_{\#}\mu, z_{\#}\nu; K).$$

Since the kernel K satisfies the universal property [220], our proposed Gaussian kernel distinguishes the difference between μ and ν as long as there exists a d-size sub-group of coordinates of μ and ν that cause the difference. Compared with linear and quadratic kernels, the Gaussian kernel is a more flexible choice. Unfortunately, the computation burden of the Gaussian kernel is heavier than the other two simple kernels because the objective in (4.12) *can be viewed as a non-concave ∞-degree monomial with respect to z, whereas the second-order approximation scheme in* (4.15) *may not provide reliable performance for optimization.*

4.4.2 Tackling the challenge of solving (STRS)

There are two challenges in solving (STRS) in particular for large-scale problems. First, since the objective function is non-concave in z, it is difficult to develop exact algorithms directly for solving (STRS). Instead, we provide a mixed-integer *convex* programming reformulation, which motivates us to develop exact algorithms in Section 4.4.2. Second, this problem is NP-hard even if the coefficient vector $a = 0$, as pointed out in [213]. When the problem is large-scale, we provide approximation algorithms with provable performance guarantees.

Exact mixed-integer SDP (MISDP) reformulation

We first provide an exact MISDP reformulation of (STRS). When the coefficient vector $a = 0$, similar reformulation results have been developed in the sparse PCA literature [30, 194]. However, such a reformulation for $a \neq 0$ is new in the literature. For notational simplicity,

we define the following block matrix of size $(D+1)\times(D+1)$:

$$\tilde{A} = \begin{pmatrix} 0 & \frac{1}{2}a^{\mathrm{T}} \\ \frac{1}{2}a & A \end{pmatrix}.$$

Theorem 13 (MISDP Reformulation of (STRS))**.** *Problem* (STRS) *can be equivalently formulated as the following MISDP*

$$\max_{Z\in\mathbb{S}^{+}_{D+1},q\in\mathcal{Q}} \quad \langle \tilde{A}, Z\rangle \tag{4.16a}$$

$$\textit{s.t.} \quad Z^{(i,i)} \le q^{(i)}, \quad i\in[D], \tag{4.16b}$$

$$Z^{(0,0)} = 1, \textit{Tr}(Z) = 2, \tag{4.16c}$$

where the set

$$\mathcal{Q} = \left\{ q\in\mathbb{F}^{D} : \sum_{k\in[D]} q^{(k)} \le d \right\}, \tag{4.17}$$

and we assume the indices of $Z, \tilde{A}\in\mathbb{S}^{+}_{D+1}$ *are both over* $[0:D]\times[0:D]$*. The continuous relaxation value of* (4.16) *equals* $w_{\mathrm{rel}} = \max_{z:\ \|z\|_2=1} \left\{ z^{\mathrm{T}}Az + z^{\mathrm{T}}t \right\}$.

The proof idea of Theorem 13 is to express the problem (STRS) as a *rank*-1 *constrained SDP* problem. Leveraging well-known results on rank-constrained optimization (see, e.g., [99, 195, 251]), one can remove the rank constraint without changing the optimal value of the original SDP problem. Although (4.16) is equivalent to (STRS), the fact that its continuous relaxation value is equal to w_{rel} suggests that it may be a weak formulation. Inspired from [30, 194], we propose the additional two valid inequalities to strengthen the formulation (4.16) in Corollary 2.

Corollary 2 (Stronger MISDP Reformulation of (STRS))**.** *The problem* (STRS) *reduces to*

the following stronger MISDP formulation:

$$\max_{Z\in\mathbb{S}^+_{D+1},q\in\mathcal{Q}} \quad \langle \tilde{A}, Z\rangle \tag{4.18a}$$

$$\text{s.t.} \quad \text{(4.16c)}, \sum_{j\in[D]}(Z^{(i,j)})^2 \le Z^{(i,i)}q^{(i)}, \left(\sum_{j\in[D]}|Z^{(i,j)}|\right)^2 \le dZ^{(i,i)}q^{(i)}, \quad \forall i\in[D]. \tag{4.18b}$$

It is worth noting that two distinct references [30, 194] have independently introduced two valid inequalities to enhance the performance of solving the sparse PCA problem, which is a special instance of (STRS) for $a = 0$. However, one of the valid inequalities in (4.18b) proposed in [30] is dominated by a valid inequality proposed in [194]. In contrast, the other valid inequality has been proposed simultaneously in these two references. This motivates us to incorporate two valid inequalities from [194] into our formulation, as outlined in Corollary 2. On the one hand, the resulting formulation (4.18) can be directly solved via some exact MISDP solvers such as YALMIP [205]. On the other hand, it enables us to develop a customized, exact algorithm to solve this formulation based on Benders decomposition since the binary vector q can be separated from other decision variables.

To develop the exact algorithm, we first reformulate the problem (4.18) as a max-min saddle point problem so that it can be solved based on the outer approximation technique [51, 114].

Theorem 14 (Saddle Point Reformulation of (4.18))**.** *Problem* (4.18) *shares the same optimal value as the following problem:*

$$\max_{q\in\mathcal{Q}} \left\{ f(q) \triangleq \max_{Z\in\mathbb{S}^+_{D+1}} \left\{ \langle \tilde{A}, Z\rangle : s.t.\ \text{(4.18b)} \right\} \right\}. \tag{4.19}$$

Here the function $f(q)$ is concave in q over the domain $\overline{\mathcal{Q}} := \mathrm{conv}(\mathcal{Q})$, and equivalently, is

Algorithm 6 Exact Algorithm for solving (STRS)

1: **Input:** Max iterations $i_{\max}$, initial guess q_1, tolerance ϵ.
2: **for** $i = 1, \ldots, i_{\max} - 1$ **do**
3: Compute q_{i+1} as the optimal solution from

$$\max_{q \in \mathcal{Q}} \left\{ \bar{f}^i(q) \triangleq \min_{1 \le j \le i} \bar{f}(q; q_j) \right\}$$

4: Compute $f(q_{i+1})$ and $g_{q_{i+1}} \in \partial f(q_{i+1})$
5: **Break** if $f(q_{i+1}) - \bar{f}^i(q_{i+1}) < \epsilon$
6: **end for**
7: **Return** $q_{i_{\max}}$

the optimal value to the following problem:

$$\begin{aligned}
\min_{\lambda, \lambda_0, \nu_1, \nu_2, \Lambda, \beta, \mu, W_1, W_2} \quad & \lambda_0 + 2\lambda + q^{\mathrm{T}} \left[\frac{d}{2}(\nu_1 - \nu_2) + \frac{1}{2}(\mu - \operatorname{diag}(\Lambda)) \right] \\
\textit{s.t.} \quad & \begin{pmatrix} -\lambda_0 & \frac{1}{2} t^{\mathrm{T}} \\ \frac{1}{2} t & A - \lambda I_D + W_1 - W_2 + \Lambda + \frac{1}{2} \operatorname{diag}(\nu_1 + \nu_2) \end{pmatrix} \preceq 0, \\
& W_1 + W_2 - \operatorname{diag}(\beta) \le 0, \\
& \sum_j (\Lambda^{(i,j)})^2 \le (\mu^{(i)})^2, \quad (\beta^{(i)})^2 + (\nu_2^{(i)})^2 \le (\nu_1^{(i)})^2, \quad i \in [D], \\
& \nu_1, \beta, \mu \in \mathbb{R}^D_+, \quad W_1, W_2 \in \mathbb{R}^{D \times D}_+, \\
& \lambda, \lambda_0 \in \mathbb{R}, \quad \nu_2 \in \mathbb{R}^D, \Lambda \in \mathbb{R}^{D \times D}.
\end{aligned} \tag{4.20}$$

For fixed q, the sup-gradient of f with respect to q can be computed as

$$\partial f(q) = \frac{d}{2}(\nu_1^* - \nu_2^*) + \frac{1}{2}(\mu^* - \operatorname{diag}(\Lambda^*)),$$

where $(\nu_1^*, \nu_2^*, \mu^*, \Lambda^*)$ is an optimal solution to the optimization problem above.

By Theorem 14, we find that given a reference direction $\hat{q}$, $f(q) \le \bar{f}(q; \hat{q}) \triangleq f(\hat{q}) + g_{\hat{q}}^{\mathrm{T}}(q - \hat{q})$, where $g_{\hat{q}}$ is a sup-gradient of f at $\hat{q}$. Based on this observation, we use the common outer-approximation technique, which is widely used for general mixed-integer

nonlinear programs [51, 114], to solve the problem: at iterations $i = 1, 2, \ldots, i_{\max} - 1$, we maximize and refine a piecewise linear upper-bound of $f(q)$: $\bar{f}^i(q) = \min_{1 \le j \le i} \ \bar{f}(q; q_j)$. The algorithm is summarized in Algorithm 6.

By the reference [114], it can be shown that this algorithm yields a non-increasing sequence of overestimators $\{\bar{f}^i(q)\}_{i=1}^{i_{\max}}$, which converge to the optimal value of $f(q)$ within a finite number of iterations $i_{\max} \le \binom{D}{1} + \cdots + \binom{D}{d}$.

Convex relaxation algorithm

Inspired by Theorem 14, a natural idea of approximately solving the problem (4.16) is to consider the following problem, in which we replace the nonconvex constraint $q \in \mathcal{Q}$ by a set of linear constraints, which forms its convex hull:

$$\max_{q \in \overline{\mathcal{Q}}} \ f(q), \quad \text{where } \overline{\mathcal{Q}} = \mathrm{conv}(\mathcal{Q}) = \left\{ q \in [0,1]^D : \ \sum_i q^{(i)} \le d \right\}. \tag{4.21}$$

Since the problem (4.21) is a convex program, it can be solved in polynomial time. Besides, one can obtain a high-quality feasible solution to the problem (4.16), using a greedy rounding scheme: We first solve (4.21) to obtain its optimal solution $\tilde{q}$, and then project it onto $\overline{\mathcal{Q}}$ to obtain q. Next, we solve the problem (4.16) by fixing the variable q and optimizing Z only.

In the following theorem, we provide the approximation ratio regarding the SDP formulation above. The proof adopts similar techniques as in [194, Theorem 5], but we extend the analysis for inhomogeneous quadratic maximization formulation.

Theorem 15 (Approximation Gap for Convex Relaxation)**.** *Denote by* optval(4.21) *and* optval(4.16) *the optimal values of problem* (4.21) *and* (4.16)*, respectively. Then, it holds that*

$$\begin{aligned} \textit{optval}(4.16) &\le \textit{optval}(4.21) \\ &\le \|a\|_2 + \min \left\{ D/d \cdot \textit{optval}(4.16), d \cdot \textit{optval}(4.16) - \min_k |a^{(k)}| \right\}. \end{aligned}$$

Despite the convexity of problem (4.21), it is challenging to solve, especially for high-dimensional scenarios. References [30, 194] solved a special case of problem (4.21) when $a = 0$ based on the interior point method (see, e.g., [5, 58, 295]). Unfortunately, since the constraint set of (4.21) involves the intersection of a semidefinite cone and a large number of second-order cones, re-writing it as a standard conic program and using off-the-shelf solvers to solve this problem spends lots of time. The work [211] designed a novel variable-splitting technique and proposed a first-order Alternating Direction Method of Multipliers [57] (ADMM) algorithm to solve a special convex relaxation of sparse PCA. Unlike this reference that only considers the simplest convex relaxation of sparse PCA without adding strong inequalities, our problem (4.21) has considerably complicated constraints.

Inspired by Ma [211], we use a similar variable-splitting technique to split the second-order conic constraints and all the other constraints in two blocks of variables and then propose an ADMM algorithm to optimize the augmented Lagrangian function. The advantage is that each subproblem in iteration update involves only second-order conic constraints or other constraints that are easy to deal with, which results in considerably fast computational speed. We provide a detailed implementation of the proposed algorithm for solving (4.21) in Appendix C.1.

Truncation algorithms with tighter approximation gap

Unfortunately, the SDP relaxation formulation is still challenging to solve for extremely high-dimension scenarios, which motivates us to develop the following computationally cheap truncation approximation algorithms. Compared with the approximation ratio of relaxed SDP formulation in Theorem 15 (i.e., $\min(D/d, d) + \mathcal{O}(1)$), the ratio for our proposed algorithm is tighter (i.e., $\min(D/d, \sqrt{d}) + \mathcal{O}(1)$). First, we introduce the definition of a normalized sparse truncation operator.

Definition 10 (Normalized Sparse Truncation)**.** *For a vector $z \in \mathbb{R}^D$ and an integer $d \in [D]$,*

we say $\bar{z}$ is a d-sparse truncation of z if

$$\bar{z}^{(i)} = \begin{cases} z^{(i)}, & \text{if } |z^{(i)}| \text{ is one of the } d \text{ largest (in absolute value) entries in } z \\ 0, & \text{otherwise}. \end{cases}$$

Besides, the vector $\widehat{z} = \bar{z}/\|\bar{z}\|_2$ is said to be the normalized d-sparse truncation of z.

Now, we introduce the following two truncation algorithms:

Truncation Algorithm (I): Let $A^{(:,i)}$ be the i-th column of A for $i \in [D]$, and denote by $\hat{z}_i$ the normalized d-sparse truncation of $A^{(:,i)}$. Then return the estimated optimal solution as the best over all $\hat{z}_i$'s and e_i's for $i \in [D]$, where e_i denotes the i-th standard basis vector.

Truncation Algorithm (II): Relax the ℓ_0-norm constraint in (STRS) and solve the trust region problem $\max_{z:\ \|z\|_2 \le 1} \{z^{\mathrm{T}} A z + z^{\mathrm{T}} a\}$ to obtain the optimal primal solution v. Then, return the estimated optimal solution z as the normalized d-sparse truncation of v.

We summarize the approximation ratios of these two truncation algorithms in Theorem 16. Its proof technique is adopted from [65]. The difference is that the authors consider the approximation ratio under the case $a = 0$, while we adopt the structure of inhomogeneous quadratic function maximization to extend the case for the general coefficient vector a.

Theorem 16 (Approximation Gap for Truncation Algorithm)**.** (I) *Truncation Algorithm* (I) *returns a feasible solution of* (STRS) *with objective value $V_{(\mathrm{I})}$ such that*

$$\textit{optval}(\text{STRS}) \ge V_{(\mathrm{I})} \ge \frac{1}{\sqrt{d}} \textit{optval}(\text{STRS}) - 2\|a\|_{(d+1)}.$$

(II) *Truncation Algorithm* (II) *returns a feasible solution of* (STRS) *with objective value $V_{(\mathrm{II})}$ such that*

$$\textit{optval}(\text{STRS}) \ge V_{(\mathrm{II})} \ge \frac{d}{D} \cdot \textit{optval}(\text{STRS}) - \frac{d}{D} \cdot \|a\|_2 - \left(1 + \sqrt{\frac{d}{D}}\right) \cdot \|a\|_{(d)}.$$

We return the best over the output from Truncation Algorithm (I) and (II) as the estimated optimal solution. By Theorem 16, we find the returned solution approximates the optimal solution up to approximation ratio $\min(D/d, \sqrt{d}) + \mathcal{O}(1)$. It has been shown in Chan et al. [65] that it is NP-hard to implement any algorithm with *constant* approximation ratio. Therefore, it is of research interest to explore polynomial-time approximation algorithms with approximation ratio that has milder dependence on D and d. Instead of trying this direction, in the next subsection, we propose another approximation algorithm such that, though NP-hard to solve, it achieves a higher approximation ratio.

Approximation algorithm via convex integer programming

In this part, we propose an approximation algorithm based on convex integer programming. We first consider the following ℓ_1-norm relaxation of the problem (STRS), which plays a key role in developing our algorithm:

$$\max \left\{ z^{\mathrm{T}} A z + z^{\mathrm{T}} a : \ \|z\|_2 \leq 1, \|z\|_1 \leq \sqrt{d} \right\}. \tag{4.22}$$

This problem is a relaxation of problem (STRS) because constraints $\|z\|_2 \leq 1, \|z\|_0 \leq d$ imply $\|z\|_1 \leq \sqrt{d}$. Following the similar proof technique as in [100, Theorem 1], we show that solving this new problem results in a constant approximation ratio. The difference is that the authors therein only consider the special case of (STRS) with $a = 0$, while we extend their analysis for general inhomogeneous quadratic objective functions.

Theorem 17 (Approximation Gap for ℓ_1-Norm Relaxation)**.** *Let* $\rho = 1 + \sqrt{d/(d+1)}$. *Then we have that* ***optval***(STRS) $\leq$ ***optval***(4.22) $\leq \rho^2$ ***optval***(STRS).

Although the problem (4.22) is a relaxation of (STRS), it is still intractable to solve due to the non-concavity of the objective function (recall that $A \succeq 0$). We adopt techniques from [100, Section 2.2] to derive a further convex integer program that serves as a further relaxation of the relaxation problem (4.22). Before proceeding, we define the following

notations. For $i \in [D]$, denote by (λ_i, v_i) the i-th eigen-pair of the matrix A, denote $\theta_i := \max\{z^{\mathrm{T}} v_i : \ \|z\|_2 \leq 1, \|z\|_1 \leq \sqrt{d}\}$, and let $\gamma_i^{[-N:N]}$ be the set of partition points of the domain $[-\theta_i, \theta_i]$, i.e., $\gamma_i^j = \frac{j}{N}\theta_i, \quad j = -N, \ldots, N$. Let $\lambda_0 \in \mathbb{R}_+$ be a fixed number such that $\lambda_0 \leq$ optval(STRS).

Proposition 2 (Convex Integer Programming Relaxation of (4.22))**.** *Consider the convex integer program:*

$$\textit{Maximize} \qquad \lambda_0 + \textstyle\sum_{i:\ \lambda_i > \lambda_0} (\lambda_i - \lambda_0)\xi_i - s \tag{4.23a}$$

that is subject to the following constraints:

$$\begin{cases} g_i = z^{\mathrm{T}} v_i, \\ |g_i| \leq \theta_i, \end{cases} \quad i \in [D], \qquad \begin{cases} \displaystyle\sum_{i \in [D]} y_i \leq \sqrt{d}, \\ y_i \geq \left| z^{(i)} \right|, i \in [D], \end{cases}$$

$$\begin{cases} g_i = \displaystyle\sum_{j \in [-N,N]} \gamma_i^j \eta_i^j, \\ \xi_i = \displaystyle\sum_{j \in [-N,N]} (\gamma_i^j)^2 \eta_i^j, \quad i \in \{i :\ \lambda_i > \lambda_0\}, \\ \eta_i^{[-N,N]} \in \text{SOS-2}, \end{cases} \qquad \begin{cases} \displaystyle\sum_{i \in [D]} (z^{(i)})^2 \leq 1, \\ \displaystyle\sum_{i:\ \lambda_i > \lambda_0} \left(\xi_i - \frac{\theta_i^2}{4N^2} \right) + \sum_{i:\ \lambda_i \leq \lambda_0} g_i^2 \leq 1, \\ \displaystyle\sum_{i:\ \lambda_i < \lambda_0} -(\lambda_i - \lambda_0) g_i^2 - z^{\mathrm{T}} a \leq s, \end{cases}$$

with SOS-2 *denoting the special ordered set of type-2 [23], and involves the following decision variables:*

$$\{g_i\}_{i=1}^D \in \mathbb{R}^D, \quad \{\xi_i\}_{i \in \{i:\ \lambda_i > \lambda_0\}} \in \mathbb{R}^{|\{i:\ \lambda_i > \lambda_0\}|},$$
$$\{\eta_i^j\}_{i \in \{i:\ \lambda_i > \lambda_0\}, j \in [-N:N]} \in \mathbb{R}^{(2N+1)|\{i:\ \lambda_i > \lambda_0\}|}, \quad \{y_i\}_{i=1}^D \in \mathbb{R}^D, \quad s \in \mathbb{R}, \quad z \in \mathbb{R}^D.$$

This problem is a relaxation of the ℓ_1-norm relaxed problem (4.22)*. Besides, it holds that* optval(STRS) $\leq$ optval(4.23) $\leq \rho^2$ optval(STRS) $+ \frac{1}{4N^2} \sum_{i:\ \lambda_i > \lambda_0} (\lambda_i - \lambda_0)\theta_i^2$*, where the constant $\rho > 0$ is defined in Theorem 17.*

The convex integer program (4.23) seems appealing because it only requires solving

$O((2N)^{|\{i:\ \lambda_i > \lambda_0\}|})$ number of finite-dimensional convex optimization problems to obtain its optimal solution. In practice, the choice of λ_0 influences the computational traceability of problem (4.23), and the choice of N influences the quality of the approximation. We follow the heuristic described in [100, Section 4.3.1] to select λ_0 and N. After solving the problem (4.23), one obtains the decision variable z that may not be feasible in $\mathcal{Z}$. Then, one can use the greedy rounding scheme to project z onto $\mathcal{Z}$ to obtain a primal feasible solution.

Finally, we acknowledge a recent concurrent study [224] that employed ℓ_1-norm relaxation as a heuristic for MMD-based variable selection. In contrast to this literature, which presented a heuristic approach, our work is the first study that provides both ℓ_1 relaxation methodology and its theoretical performance guarantees.

4.5 Statistical Properties

In this section, we provide statistical performance guarantees for the variance-regularized MMD statistics in (4.12), specialized for our proposed kernels in (4.5), (4.6), and (4.7). In addition, we develop the guarantees for a generic kernel in Appendix C.4.

We first show that empirical estimators $S^2(\mathbf{x}^n, \mathbf{y}^n; K_z)$ and $\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n, \mathbf{y}^n; K_z)$ uniformly converge to their population version as the sample size n increases. Such a property is useful for showing the testing consistency and the rate of testing power of our MMD framework.

Proposition 3 (Non-asymptotic Concentration Properties)**.** *For Gaussian kernel in* (4.7)*, we assume the sample space* $\Omega \subseteq \{x \in \mathbb{R}^D :\ \|x\|_\infty \leq R\}$ *for some constant* $R > 0$*. With probability at least* $1 - \delta$*, (i) the bias approximation error can be bounded as*

$$\sup_{z \in \mathcal{Z}} \left| S^2(\mathbf{x}^n, \mathbf{y}^n; K_z) - \mathrm{MMD}^2(\mu, \nu; K_z) \right| \leq \epsilon^1_{n,\delta} = \begin{cases} \widetilde{\mathcal{O}}(dn^{-1/2}), & \textit{for linear kernel}, \\ \widetilde{\mathcal{O}}(d^{3/2}n^{-1/2}), & \textit{for quadratic kernel}, \\ \widetilde{\mathcal{O}}(d^{1/2}n^{-1/2}), & \textit{for Gaussian kernel}, \end{cases}$$

where $\widetilde{\mathcal{O}}(\cdot)$ *hides a multiplicative factor* $(\log n + \log(D/d) + \log \frac{1}{\delta})^{1/2}$ *and other constants that are independent to parameters* D, d, n*. (ii) and the variance approximation error can*

be bounded as

$$\sup_{z\in\mathcal{Z}}\left|\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z)-\mathbb{E}_{\mathbf{x}^n\sim\mu,\mathbf{y}^n\sim\nu}[\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z)]\right|\leq\epsilon^2_{n,\delta}$$

where $\epsilon^2_{n,\delta}$ *shares the same order of decaying rate as* $\epsilon^1_{n,\delta}$.

Based on the concentration properties above, we are ready to derive the asymptotic distribution of the testing statistic. Furthermore, we impose specific assumptions on data distributions when defining linear, quadratic, or Gaussian kernels. The technical assumption regarding the Gaussian kernel is the most lenient, followed by the quadratic kernel, which is less lenient, whereas the assumption for the linear kernel is the most restrictive, reflecting a gradual decrease in flexibility across these kernels.

Assumption 3 (Structure of Data)**.** *Assume the following conditions hold:*

(I) *For linear kernel, there exists* $s^*\in[D]$ *such that* $\mathrm{Proj}_{s^\#}\mu\neq\mathrm{Proj}_{s^\#}\nu$.

(II) *For quadratic kernel, there exists* $s^*\in[D]$ *such that*

$$\mathrm{Proj}_{s^\#}\mu\neq\mathrm{Proj}_{s^\#}\nu \text{ or } (\mathcal{A}(\mu,\nu))^{(s^*,s^*)}>0$$

(III) *For Gaussian kernel, the sample space* $\Omega\subseteq\{x\in\mathbb{R}^D:\ \|x\|_\infty\leq R\}$ *for some constant* $R>0$*, and there exists* $S\subseteq[D]$ *with* $|S|\leq d$ *such that* $\mathrm{Proj}_{S\#}\mu\neq\mathrm{Proj}_{S\#}\nu$.

Proposition 4 (Asymptotic Distribution of Testing Statistic)**.** *Under Assumption 3, suppose the hyper-parameter* $\lambda>0$ *is properly selected (see its detailed range in Proposition 15), and let* $\hat{z}_{\mathrm{Tr}}$ *be the obtained sparse coefficient by solving* (4.12) *from training dataset* $(\mathbf{x}^{\mathrm{Tr}},\mathbf{y}^{\mathrm{Tr}})$ *with* $|\mathbf{x}^{\mathrm{Tr}}|=|\mathbf{y}^{\mathrm{Tr}}|=n_{\mathrm{Tr}}$*,* $T_{n_{\mathrm{Te}}}$ *be the testing statistic evaluated on testing dataset* $(\mathbf{x}^{\mathrm{Te}},\mathbf{y}^{\mathrm{Te}})$ *with* $|\mathbf{x}^{\mathrm{Te}}|=|\mathbf{y}^{\mathrm{Te}}|=n_{\mathrm{Te}}$*. Then, it holds that*

(I) *Under alternative hypothesis* $\mathcal{H}_1 : \mu \neq \nu$*, for training sample size*

$$n_{\mathrm{Tr}} = \begin{cases} \Omega\left(d^2 \log(D/d)\right), & \textit{for linear kernel}, \\ \Omega\left(d^4 \log(D/d)\right), & \textit{for quadratic kernel}, \\ \Omega\left(d \log(D/d)\right), & \textit{for Gaussian kernel}, \end{cases} \tag{4.24}$$

it holds that $\mathbb{E}[T_{n_{\mathrm{Te}}} \mid \mathcal{H}_1] > 0$ *with high probability.*

(II) *Under null hypothesis* $\mathcal{H}_0 : \mu = \nu$*, it holds that* $n_{\mathrm{Te}} T_n \to \sum_i \sigma_i (Z_i^2 - 2)$*, with* σ_i *denoting the eigenvalues of the* μ*-covariance operator of the centered kernel; under alternative hypothesis* $\mathcal{H}_1 : \mu \neq \nu$*, it holds that* $\sqrt{n_{\mathrm{Te}}}(T_n - \mathbb{E}[T_{n_{\mathrm{Te}}} \mid \mathcal{H}_1]) \to \mathcal{N}(0, \sigma^2_{\mathcal{H}_1}(\mu, \nu; K_{\hat{z}_{\mathrm{Tr}}}))$.

It is worth mentioning that the order of training sample size in (4.24) depends only on factors poly(d) and $\log(D/d)$, indicating the statistical guarantees of our proposed variable selection framework do not suffer from the curse of dimensionality. Based on Proposition 4, we finally present the consistency and rate of testing power of our framework.

Theorem 18 (Consistency)**.** *Under the same assumption as in Proposition 4, specify the training sample size* n_{Tr} *as in* (4.24)*. Let* $\alpha \in (0, 1)$ *denote the level of two-sample test and take* τ *as the* $(1 - \alpha)$*-quantile of the limiting distribution* $\sum_i \sigma_i (Z_i^2 - 2)$ *defined in Proposition 4*(II)*, and let the threshold of the test be* $t_{\mathrm{thres}} := \frac{\tau}{n_{\mathrm{Te}}}$*. As a consequence,* $\mathbb{P}\big(T_{n_{\mathrm{Te}}} > t_{\mathrm{thres}} \mid \mathcal{H}_0\big) \to \alpha$ *and* $\mathbb{P}\big(T_{n_{\mathrm{Te}}} \leq t_{\mathrm{thres}} \mid \mathcal{H}_1\big) \to 0$.

Theorem 19 (Testing Power)**.** *Under the same assumption as in Proposition 4, and the same choices of training sample size and threshold as in Theorem 18, we additionally assume* $\mathbb{E}[|T_1|^3 \mid \mathcal{H}_1] < \infty$*. When the testing sample size* n_{Te} *is sufficiently large so that*

$$t_{\mathrm{thres}} + \frac{\Phi^{-1}(1 - n_{\mathrm{Te}}^{-1/2})}{\sqrt{n_{\mathrm{Te}}}} = \frac{\tau}{n_{\mathrm{Te}}} + \sqrt{\frac{\ln \frac{n_{\mathrm{Te}}}{2\pi} - \ln\ln \frac{n_{\mathrm{Te}}}{2\pi}}{n_{\mathrm{Te}}}}(1 + o(1))$$

is sufficiently small, where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} \, \mathrm{d}t$ *denotes the error function, it holds*

that

$$\mathbb{P}\big(T_{n_{\mathrm{Te}}} > t_{\mathrm{thres}} \mid \mathcal{H}_0\big) \leq \alpha + \mathcal{O}(n_{\mathrm{Te}}^{-1/2}) \ \textit{and} \ \mathbb{P}\big(T_{n_{\mathrm{Te}}} \leq t_{\mathrm{thres}} \mid \mathcal{H}_1\big) \leq \mathcal{O}(n_{\mathrm{Te}}^{-1/2}),$$

where $\mathcal{O}(\cdot)$ *hides constant related to parameters* $\mathbb{E}[|T_1|^3 \mid \mathcal{H}_1]$ *and*

$$\sigma^2_{\mathcal{H}_1} := \mathbb{E}_{\mathbf{x}^n \sim \mu, \mathbf{y}^n \sim \nu}[\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n, \mathbf{y}^n; K_{\hat{z}_{\mathrm{Tr}}})].$$

Theorem 19 indicates that under alternative hypothesis $\mu \neq \nu$, as long as the testing sample size n_{Te} is sufficiently large, and the training sample size is chosen according to (4.24), the testing power approaches 1 with error rate $\mathcal{O}(n_{\mathrm{Te}}^{-1/2})$.

4.6 Simulated numerical examples

We first consider synthesized data sets to examine the performance of our proposed variable selection framework. We consider the following four cases:

(I) (Gaussian Mean Shift): Data distribution $\mu = \mathcal{N}(0, \Sigma)$ with the covariance matrix $\Sigma^{(s_1,s_2)} = \rho^{|s_1 - s_2|}$ for some correlation level $\rho \in (0, 1)$. Data distribution $\nu = \mathcal{N}(\mu, \Sigma)$ with the mean vector $\mu^{(s)} = \tau/s, \forall s \in [d_{\mathrm{true}}]$ for some scalar $\tau > 0$ and otherwise $\mu^{(s)} = 0$.

(II) (Gaussian Covariance Shift): Data distribution $\mu = \mathcal{N}(0, \Sigma)$ with Σ specified the same as in Part (I), and $\nu = \mathcal{N}(0, \tilde{\Sigma})$, with $\tilde{\Sigma}^{(s_1,s_2)} = \tau\Sigma^{(s_1,s_2)}, \forall s_1, s_2 \in [d_{\mathrm{true}}]$ for some scalar $\tau > 1$ and otherwise $\tilde{\Sigma}^{(s_1,s_2)} = \Sigma^{(s_1,s_2)}$.

(III) (Gaussian versus Laplacian): Data distribution $\mu = \mathcal{N}(0, I_D)$. The first d_{true} coordinates of ν are independent Laplace distributions with zero mean and standard deviation 0.8. The remaining coordinates of ν_Y are independent Gaussian distributions $\mathcal{N}(0, 1)$.

(IV) (Gaussian versus Gaussian Mixture): Data distribution $\mu = \mathcal{N}(0, I_D)$. The first d

coordinates of ν are Gaussian mixture distribution $\frac{1}{2}\mathcal{N}(-\mu, I_{d_{\text{true}}}) + \frac{1}{2}\mathcal{N}(\mu, I_{d_{\text{true}}})$ while the remaining coordinates are independent Gaussian distributions $\mathcal{N}(0, 1)$. Here the mean vector $\mu^{(s)} = \tau/s, \forall s \in [d_{\text{true}}]$ for some scalar $\tau > 0$.

Unless otherwise specified, we take hyper-parameters in Case (I) as $\tau = 1, \rho = 0.5$, in Case (II) as $\tau = 2, \rho = 0.5$, and in Case (IV) as $\tau = 2$. We quantify the performance in terms of hypothesis testing metrics rather than the prediction accuracy metrics used in the literature. Besides, we also measure the quality of variable selection using *false-discovery proportion* (FDP) and the *non-discovery proportion* (NDP) defined in [16]:

$$\text{FDP}(I) = \frac{|I \setminus I^*|}{|I|}, \quad \text{NDP}(I) = \frac{|I^* \setminus I|}{|I^*|}, \tag{4.25}$$

where I^* denotes the ground truth feature set and I denotes the set obtained by variable selection algorithms. The smaller the FDP or NDP is, the better performance the obtained feature set has.

For simplicity of implementation, we chose the bandwidth hyper-parameter τ_s^2 for the kernel $k_s(x, y)$ using the median heuristic, i.e., we specify it as the median among all pairwise distances for data points in the s-th coordinate. Similarly, we take bandwidth σ^2 of Gaussian kernel as the median among all pairwise distances for data points (over all coordinates), and bandwidth of quadratic kernel as $c = \sqrt{\sigma^2}$. Users are also recommended to tune those hyper-parameters based on the cross-validation technique, which tends to return near-optimal hyper-parameter choices for large sample sizes.

4.6.1 Numerical performance for solving (STRS)

We first examine the numerical performance of various approximation algorithms for solving (STRS), by taking the MMD optimization with linear kernel (see the reformulation in Section 4.4.1) as a numerical example. For each of the four synthetic datasets, we try various

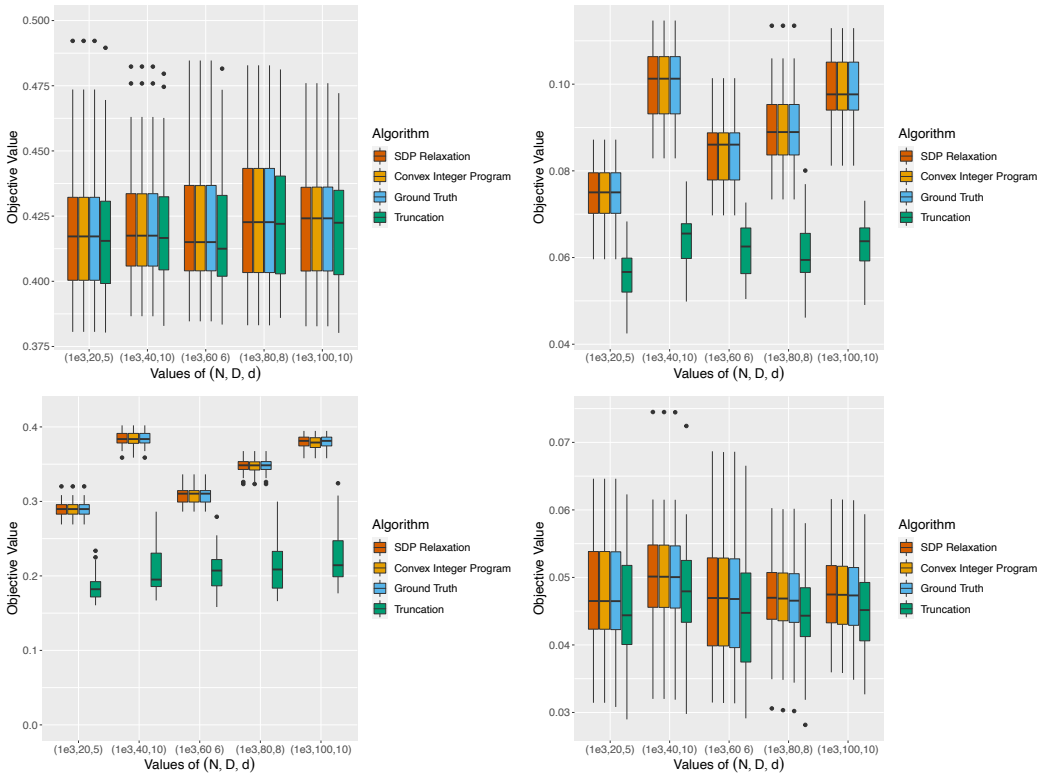

Figure 4.1: Box plots on the performance of various approximation algorithms for solving (STRS). The x-axis corresponds to various choices of (N, D, d), and y-axis corresponds to the estimated objective value of (STRS). Plots from top to bottom correspond to four types of synthetic datasets.

choices of parameters (N, D, d) from the set

$$\{(\textsf{1e3}, 20, 5), (\textsf{1e3}, 40, 10), (\textsf{1e3}, 60, 6), (\textsf{1e3}, 80, 8), (\textsf{1e3}, 100, 10)\}.$$

We also specify different hyper-parameters $\lambda \in \{0.8, 0.7, 0.6, 0.5\}$ when using these four different datasets, respectively. Since those approximation algorithms may return a solution that is infeasible to the constraint $\mathcal{Z}$, we estimate the corresponding feasible solution by performing the normalized sparse truncation (see Definition 10).

Figure 4.1 reports the objective value obtained from the feasible solution based on those approximation algorithms, where the error bars are generated using 100 independent trials. The larger the objective value is, the better performance the designed algorithm has. From the plot, we can see that semidefinite relaxation and convex integer programming algorithms

perform nearly optimally compared with the ground truth. In contrast, the performance truncation algorithm is slightly worse compared with those approaches. Table 4.1 reports the corresponding computation time of those approximation algorithms, from which we identify that the truncation algorithm has the fastest computational speed. In contrast, SDP relaxation has the slowest speed. Since the convex integer programming algorithm has satisfactory performance with relatively fast computational speed, we recommend using this approximation algorithm when solving (STRS).

Table 4.1: Averaged computational time of various approximation algorithms for solving (STRS).

Data Type	Parameters			Averaged Computational Time(s) of Approximation Algorithms		
	n	D	d	Truncation Algorithm	SDP Relaxation	Convex Integer Programming
Gaussian	1e3	20	5	2.13e-3	1.18	1.25e-1
Mean Shift	1e3	40	10	6.08e-3	2.55	2.29e-1
	1e3	60	6	1.30e-2	4.80	4.47e-1
	1e3	80	8	3.08e-2	6.19	6.87e-1
	1e3	100	10	6.87e-2	9.47	8.77e-1
Gaussian	1e3	20	5	2.33e-3	1.18	1.24e-1
Covariance Shift	1e3	40	10	5.86e-3	2.57	3.11e-1
	1e3	60	6	1.32e-2	4.80	4.02e-1
	1e3	80	8	3.07e-2	6.46	9.38e-1
	1e3	100	10	6.76e-2	9.73	1.23
Gaussian	1e3	20	5	2.28e-3	1.29	1.69e-1
versus Laplacian	1e3	40	10	6.39e-3	2.85	5.65e-1
	1e3	60	6	1.44e-2	5.20	5.14e-1
	1e3	80	8	3.31e-2	6.79	1.15
	1e3	100	10	6.95e-2	1.02e+1	2.10
Gaussian	1e3	20	5	2.17e-3	1.16	1.17e-1
versus	1e3	40	10	6.38e-3	2.57	2.14e-1
Gaussian Mixture	1e3	60	6	1.42e-2	4.72	3.94e-1
	1e3	80	8	3.31e-2	6.07	5.91e-1
	1e3	100	10	7.25e-2	9.44	8.67e-1

4.6.2 Impact of sample size and data dimension

In this subsection, we compare the performance of variable selection based on the following approaches: (I) Linear kernel MMD; (II) Quadratic kernel MMD; (III) Gaussian kernel MMD; (IV) Sparse Logistic Regression: a framework that trains the projection vector with ℓ_0-norm constraint to minimize the logistic loss [33]; and (V) Projected Wasserstein: variable selection framework using projected Wasserstein distance [228]. For baselines (I)-(III), we also compare the performance of standard MMD testing without the variable selection technique. We quantify the performance using the testing power metric with controlled type-I error $\alpha_{\text{level}} = 0.05$, and take the training/testing sample sizes as $n_{\text{Tr}} = n_{\text{Te}} = n$.

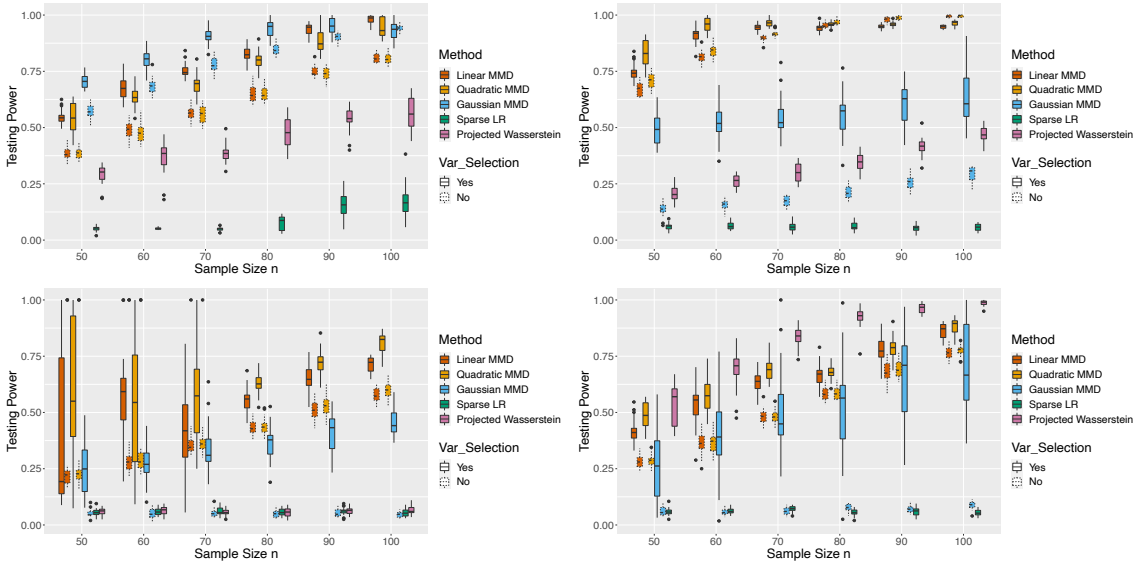

Figure 4.2: Testing power of various two-sample tests with different choices of sample size n. Here we fix parameters $D = 100,\ d_{\text{true}} = 20,\ d = 20$ and control the type-I error $\alpha_{\text{level}} = 0.05$. Plots from top to bottom correspond to four different types of synthetic datasets.

Figure 4.2 reports a numerical study on the impact of sample size n with data dimension $D = 100$, number of different variables $d_{\text{true}} = 20$ and sparsity level $d = 20$. The error bars are generated using 20 independent trials. From these plots, we find that the sparse logistic regression does not have a competitive performance in general. The explanation is that a linear classifier is not flexible enough to distinguish the distributions from two

groups. Following the similar argument from Example 2, one can check the testing statistic of this baseline always equals zero as long as the mean vectors of two distributions are the same, which explains why this baseline has nearly zero power for the synthetic dataset of case (II)-(IV). The testing power for the other two-sample testing methods increases with respect to the sample size. We can see the variable selection technique improves the performance of the standard MMD framework. For the first three synthetic datasets, the linear or quadratic MMD testing with variable selection achieves superior performance than other baselines. At the same time, for the last example, the projected Wasserstein distance has the best performance. One possible explanation is that the MMD testing framework may not be good at detecting distribution changes for Gaussian mixture distributions.

Next, we examine the impact of the data dimension D with fixed $n = 50, d_{\text{true}} = 20, d = 20$ in Figure 4.3. We omit the performance of the sparse logistic regression baseline because it does not achieve satisfactory testing performance as studied before. From those plots, we find that as the data dimension increases, all methods tend to have decreasing testing power. However, the decaying rate of MMD testing with the variable selection procedure seems slower than that of standard MMD testing. For the synthetic dataset of case (I), the Gaussian kernel has the best performance, while for case (II)-(III), the linear or quadratic kernel has the best performance. A possible explanation is that one can optimize the linear kernel with strong performance guarantees, whereas we only use quadratic approximation heuristics to optimize other types of kernel functions. Since the quadratic approximation of the objective for the quadratic kernel seems to be tight, it is intuitive to see the performance of the quadratic kernel is also consistently good.

4.6.3 Results on support recovery

In this subsection, we demonstrate the performance of support recovery for various variable selection methods, evaluated using the FDP and NDP metrics defined in (4.25). Since our goal is to assess consistent performance across both metrics, we also compute the average

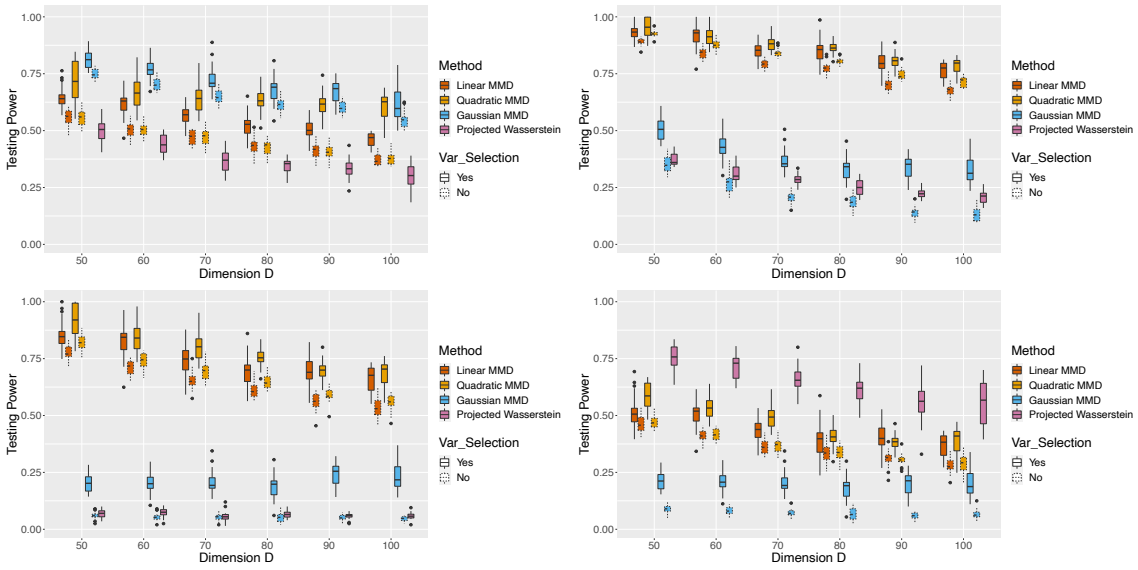

Figure 4.3: Testing power of various two-sample tests with different choices of data dimension D. Here we fix parameters $n = 50,\ d_{\text{true}} = 20,\ d = 20$ and control the type-I error $\alpha_{\text{level}} = 0.05$. Plots from top to bottom correspond to four different types of synthetic datasets.

of the two as a combined measure. We set the parameters to $D = 100$ and $d_{\text{true}} = 20$, while varying the sparsity level $d = 1, \ldots, 20$. The sample size $n = 150$ remains fixed for Cases (II) and (III), while for Case (I) and Case (IV), we adjust the parameters to $(\tau, n) = (2, 200)$ and $(\tau, n) = (4, 200)$, respectively, to ensure reliable support recovery.

Figure 4.4 presents the numerical results for support recovery: each row corresponds to a different dataset, from Case (I) to (IV), while each column represents a different performance measure (FDP, NDP, or their average). Each subplot illustrates the performance of various variable selection methods at different sparsity levels. From these plots, we observe that our proposed variable selection framework, whether using a linear or quadratic kernel, consistently outperforms the alternatives across all four cases, as reflected by the lowest FDP and NDP values. This finding aligns with the testing power performance discussed in the previous subsection. Additionally, it is noteworthy that for Cases (II) and (III), our methods achieve near-optimal performance, as indicated by the nearly horizontal FDP lines and the almost straight NDP lines.

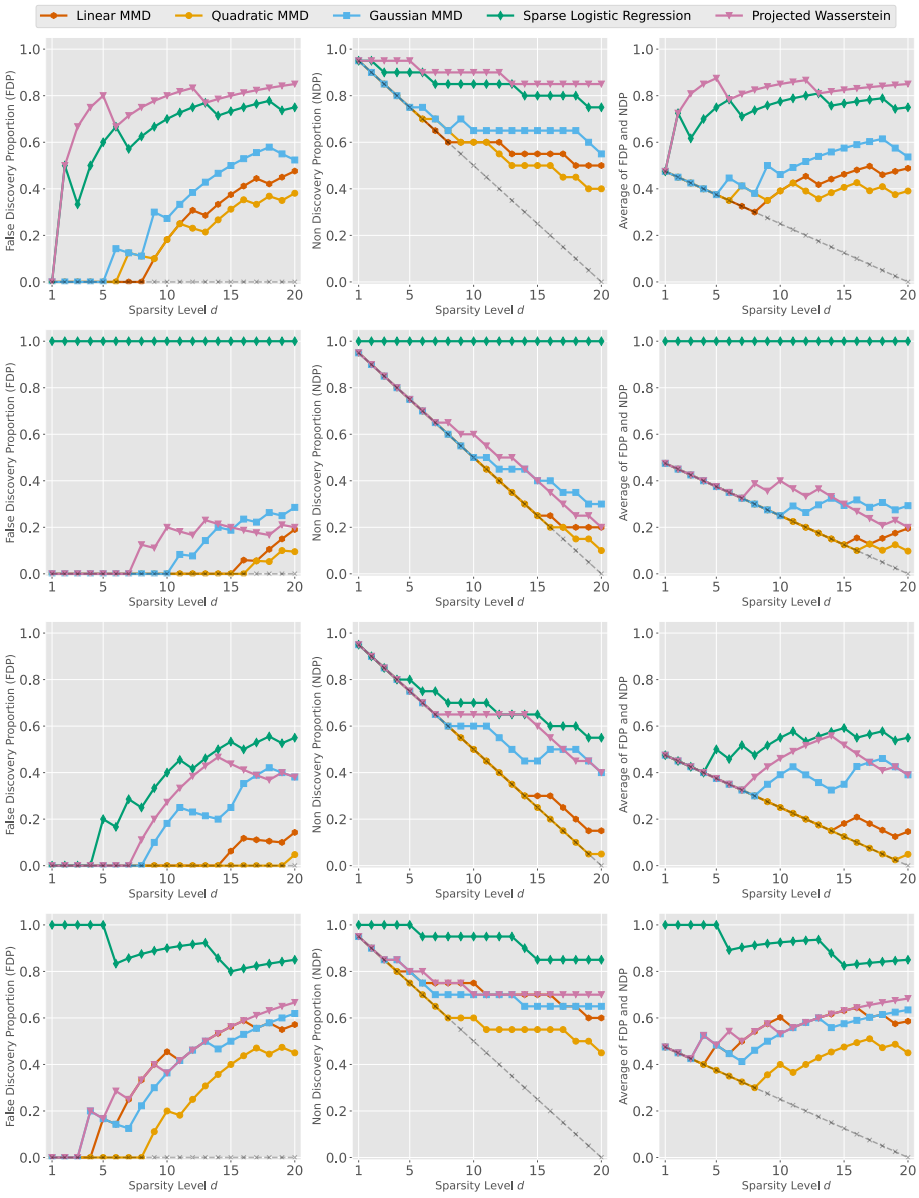

Figure 4.4: FDP and NDP metrics obtained by various approaches for different choices of sparsity level d. The dimension of the problem is $D = 100$; the true sparsity level is 20. Figures for different columns correspond to different synthetic datasets. For each subplot, the dashed gray line denotes the performance for the ideal case where ground truth variables are selected successfully; the closer to the dashed gray line, the better.

4.7 Real data examples

In this section, we present additional numerical studies with real-world datasets. Specifically, we demonstrate a visualization of variable selection based on the MNIST handwritten digits

image dataset in Section 4.7.1. Next, we show that the variable selection approach can help identify key variables for disease diagnosis in Section 4.7.2.

4.7.1 Visualization on MNIST image datasets

In this part, we demonstrate a visualization of our variable selection framework by taking the classification of MNIST image datasets, consisting of 28×28 gray-scale handwritten images for digits from 0 to 9, as toy examples. We take the training sample size $n_{\mathrm{Tr}} = 20$ and testing sample size $n_{\mathrm{Te}} = 5$. We pre-process the MNIST images by performing a 2d convolutional operator using the kernel of size 9×9. The pre-processed samples have dimension $D = 169$, and we take the number of selected variables $d = 20$. We construct four types of data distributions (μ, ν) for two-sample testing: μ and ν are distributions of images corresponding to digits 0 and 6, 8 and 9, 3 and 8, or 7 and 9, respectively. We show the visualization results in Figure 4.5. Specifically, different rows correspond to different data distributions for two-sample testing. Plots in the left two columns visualize the selected pixels (highlighted with red square markers) on two different image samples based on our linear kernel variable selection framework, from which we can see that our proposed method identifies the difference between two digits correctly. Plots in the third column report the MMD statistic compared with the empirical distribution under H_0 via test-only bootstrap, where the green circle markers correspond to the bootstrap threshold for rejecting H_0 and red star markers correspond to the testing statistics. From these plots, we find our proposed framework has satisfactory testing power even with small training and testing sample sizes. Plots on the fourth column report the visualization of the distribution of the MNIST dataset after variable selection embedded in 2D generated by tSNE [297], which is estimated based on 1000 testing samples. In comparison, we plot the estimated witness function (defined in (4.13)) as a color field over those samples in the fifth column. From those plots, we can see that the estimated witness function identifies the region of the distribution change for all of these four two-sample testing tasks.

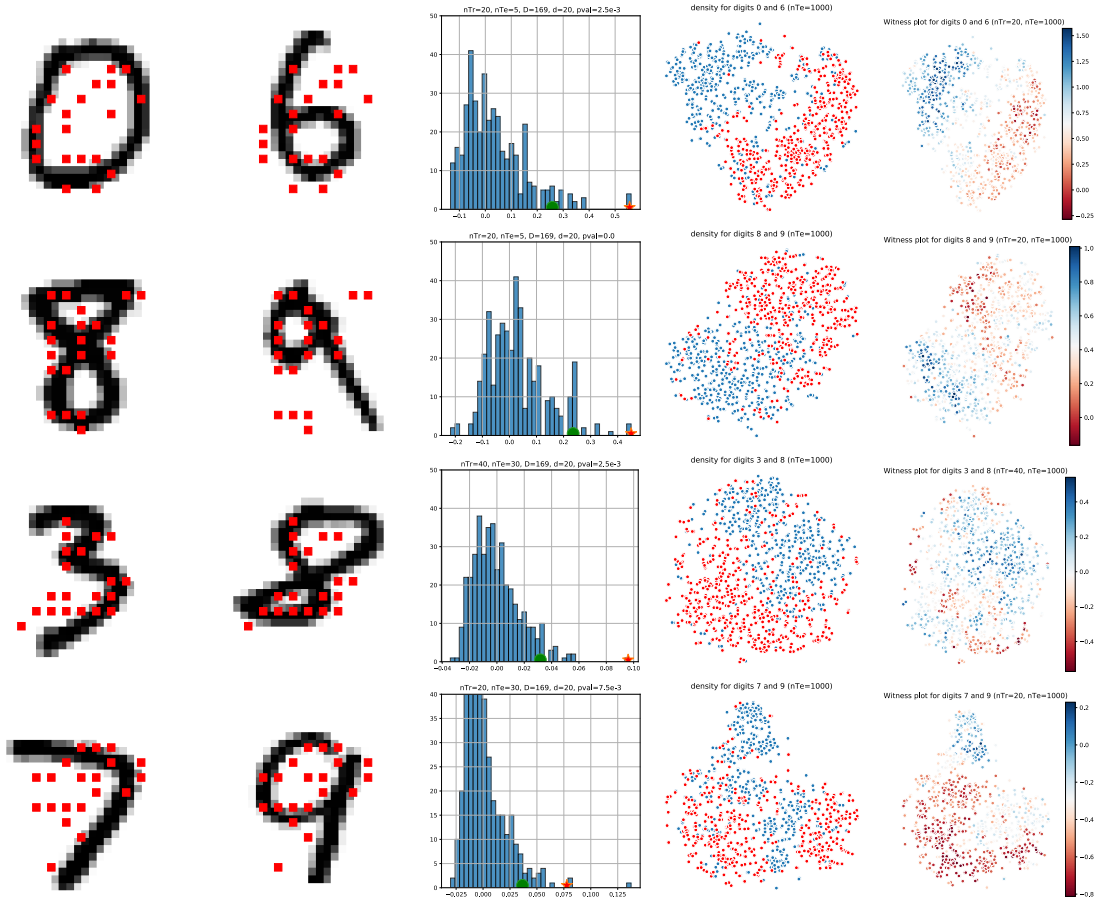

Figure 4.5: Different rows correspond to two-sample testing with different MNIST digits. The first two column plots visualize the selected pixels based on the variable selection framework. The third column plots visualize the MMD statistic together with the empirical distribution under H_0 that is estimated via bootstrapping (the green circle markers correspond to the bootstrap threshold for rejecting H_0, and red star markers correspond to the testing statistics). The fourth column plots visualize the distribution of MNIST digits after variable selection embedded in 2D. The fifth column plots visualize the estimated witness function (defined in (4.13)) for MMD.

4.7.2 Healthcare datasets

Finally, we study the performance of variable selection on a healthcare dataset [313] that records information for healthy people and Sepsis patients. This dataset consists of $D = 39$ features from $m = 20771$ healthy people and $n = 2891$ Sepsis patients. We take training samples with sample sizes $m_{\text{Tr}} = 20000, n_{\text{Tr}} = 2000$ and specify the remaining as validation samples. We quantify the performance of variable selection as the testing power on testing

samples with sample size $m_{\text{Te}} = n_{\text{Te}} = 100$. The testing power is computed based on randomly selected samples and their associated labels from the validation sample sets, with a significance level of 0.05. We repeat the testing procedure for 2000 independent trials and report the average testing power in Table 4.2.

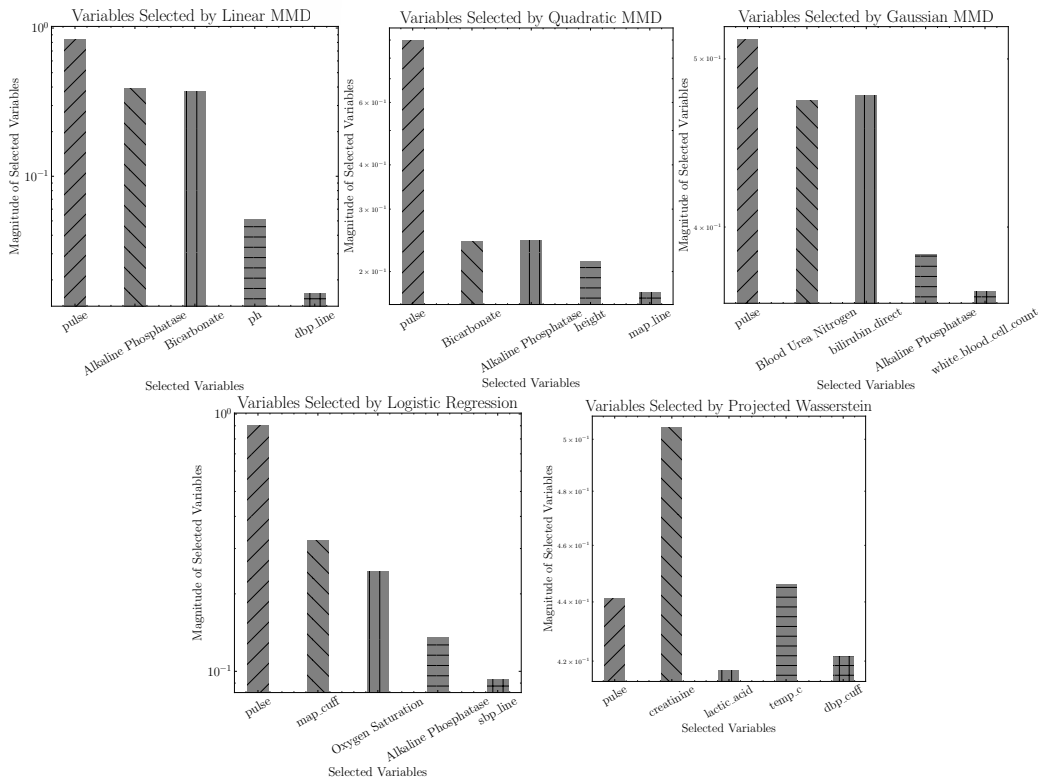

Figure 4.6: Top 5 variables selected by various approaches in the healthcare dataset.

We report the top 5 features selected by various approaches based on the training samples in Figure 4.6. From the Table, we can see that methods Quadratic MMD and Linear MMD perform the best, and the intersection of those selected features are pulse, Bicarbonate, Alkaline Phosphatase.

4.8 Conclusion

We studied variable selection for the kernel-based two-sample testing problem, which can be formulated as mixed-integer programming problems. We developed exact and approximate algorithms with performance guarantees to solve those formulations. Theoretical properties

Table 4.2: Averaged testing power for the sepsis prediction at a significance level $\alpha = 0.05$ (i.e., the threshold is set such that the Type-I error when the two distributions are the same is set to 0.05.)

Linear MMD	Quadratic MMD	Gaussian MMD	Logistic Regression	Projected Wasserstein
0.835	0.915	0.784	0.771	0.749

for the proposed frameworks are provided. Finally, we validated the power of this approach in synthetic and real datasets. In the meantime, several interesting research topics are left for future work. For example, providing theoretical analysis on the optimal choice of kernel hyper-parameters and support recovery for variable selection is of future research interest. Additionally, it holds great significance in developing more efficient algorithms for variable selection when working with different types of kernels.

CHAPTER 5
SINKHORN DISTRIBUTIONALLY ROBUST OPTIMIZATION

In this chapter, we study distributionally robust optimization (DRO) with Sinkhorn distance—a variant of Wasserstein distance based on entropic regularization. We derive a convex programming dual reformulation for general nominal distributions, transport costs, and loss functions. To solve the dual reformulation, we develop a stochastic mirror descent algorithm with biased subgradient estimators and derive its computational complexity guarantees. Finally, we provide numerical examples using synthetic and real data to demonstrate its superior performance. This work is mainly summarized in [305].

5.1 Introduction

Decision-making problems under uncertainty arise in various fields such as operations research, machine learning, engineering, and economics. In these scenarios, uncertainty in the data arises from factors like measurement error, limited sample size, contamination, anomalies, or model misspecification. Addressing this uncertainty is crucial to obtain reliable and robust solutions. In recent years, Distributionally Robust Optimization (DRO) has emerged as a promising data-driven approach to tackle these challenges. DRO aims to find a minimax robust optimal decision that minimizes the expected loss under the most adverse distribution within a predefined set of relevant distributions, known as an ambiguity set. This approach provides a principled framework to handle uncertainty and obtain solutions that are resilient to distributional variations. It goes beyond the traditional sample average approximation (SAA) method used in stochastic programming and offers improved out-of-sample performance. For a comprehensive overview of DRO, we refer interested readers to the recent survey by [183].

At the core of distributionally robust optimization lies the crucial task of selecting an

appropriate ambiguity set. An ambiguity set should strike a balance between computational efficiency and practical interpretability while being rich enough to encompass relevant distributions and avoiding unnecessary ones that may lead to overly conservative decisions. In the literature, various formulations of DRO have been proposed, among which the ambiguity set based on Wasserstein distance has gained significant attention in recent years [44, 125, 226, 319]. The Wasserstein distance incorporates the geometry of the sample space, making it suitable for comparing distributions with non-overlapping supports and hedging against data perturbations [125]. The Wasserstein ambiguity set has received substantial theoretical attention, with provable performance guarantees [43, 45, 46, 120, 274]. Empirical success has also been demonstrated across a wide range of applications, including operations research [40, 78, 237, 284, 285], machine learning [41, 69, 209, 236, 275], stochastic control [311, 327, 328], and more.

However, the current Wasserstein DRO framework has its limitations. First, the computational efficiency of Wasserstein DRO is achieved under somewhat stringent conditions, as its dual formulation involves a subproblem that requires the global supremum of some regularized loss function over the sample space. Let $\min_{\theta\in\Theta}\max_{\mathbb{P}\in\mathfrak{M}}\ \mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)]$ denote a typical Wasserstein DRO formulation, where the loss function $f_\theta(z)$ is convex in θ belonging to a closed and convex feasible region Θ, and the ambiguity set $\mathfrak{M}$ is centered around a nominal distribution $\widehat{\mathbb{P}}$ and contains distributions supported on a space $\mathcal{Z}$. Table 5.1 summarizes the known cases where solving Wasserstein DRO is computationally efficient. One general approach to solving it is to use a finite and discrete grid of scenarios to approximate the entire sample space. This involves solving the formulation restricted to the approximated sample space [72, 203, 249], but suffers from the curse of dimensionality. Simplified convex reformulations are known when the loss function can be expressed as a pointwise maximum of finitely many concave functions [111, 125, 273], or when the loss is the generalized linear model [272, 273, 275, 332]. In addition, efficient first-order algorithms have been developed for Wasserstein DRO with strongly convex transport cost, smooth loss functions,

and sufficiently small radius (or equivalently, sufficiently large Lagrangian multiplier) so that the involved subproblem becomes strongly convex [47, 286]. However, beyond these conditions on the loss function and the transport cost, solving Wasserstein DRO becomes a computationally challenging task. Second, from a *modeling* perspective, in data-driven Wasserstein DRO, where the nominal distribution is finitely supported (usually the empirical distribution), the worst-case distribution is shown to be a discrete distribution [125] (which is unique when the regularized loss function has a unique maximizer). This is the case even though the underlying true distribution in many practical applications may be continuous. Consequently, concerns arise regarding whether Wasserstein DRO hedges the right family of distributions and whether it induces overly conservative solutions.

Table 5.1: Known cases of Wasserstein DRO where it is computationally efficient to solve

References	Loss function $f_\theta(z)$	Transport cost	Nominal distribution $\widehat{\mathbb{P}}$	Support $\mathcal{Z}$
[72, 203, 249]	General	General	General	Discrete and finite set
[111, 125, 273]	Piecewise concave in z	General	Empirical distribution	General
[272, 273, 275, 332]	Generalized linear model in (z, θ)	General	General	Whole Euclidean space[1]
[47, 286]	$z \mapsto f_\theta(z) - \lambda^* c(x, z)$ is strongly concave[2]	Strongly convex function[3]	General	General

To address the aforementioned concerns while retaining the advantages of Wasserstein DRO, we propose a novel approach called Sinkhorn DRO. Sinkhorn DRO leverages the Sinkhorn distance [92], which hedges against distributions that are close to a given nominal distribution in Sinkhorn distance. The Sinkhorn distance can be viewed as a smoothed version of the Wasserstein distance and is defined as the minimum transport cost between two distributions associated with an optimal transport problem with entropic regularization (see Definition 11 in Section 5.2). To the best of our knowledge, this paper is the first to explore the DRO formulation using the Sinkhorn distance. Our work makes several contributions, which are summarized below:

(I) We derive a strong duality reformulation for Sinkhorn DRO (Theorem 20) in a highly general setting, where the loss function, transport cost, nominal distribution, and probability support are allowed to be arbitrary. The dual objective of Sinkhorn DRO smooths the dual objective of Wasserstein DRO, where the level of smoothness is controlled by the entropic regularization parameter (Remark 13).

(II) Our duality proof yields an insightful characterization of the worst-case distribution (Remark 14). Unlike Wasserstein DRO, where the worst-case distribution is typically discrete and finitely supported, the worst-case distribution in Sinkhorn DRO is absolutely continuous with respect to a pre-specified reference measure, such as the Lebesgue or counting measure. This characteristic of Sinkhorn DRO highlights its flexibility as a modeling choice and provides a more realistic representation of uncertainty that better aligns with the underlying true distribution in practical scenarios.

(III) The dual reformulation of Sinkhorn DRO can be viewed as a conditional stochastic optimization [154, 156, 160] involving an expectation (with respect to observed samples) of nonlinear transformation of a conditional expectation (with respect to a conditional distribution). In our work, we introduce and analyze an efficient stochastic mirror descent algorithm with biased subgradient estimators to solve this problem (Section 5.4). We quantify the computational cost using the number of generated samples from the outer expectation and the number of generated samples from the inner expectation. Our algorithm achieves both complexities of $\widetilde{\mathcal{O}}(\delta^{-2})$ for a fixed entropic regularization parameter ϵ.

(IV) To validate the effectiveness and efficiency of the proposed Sinkhorn DRO model, we conduct a series of experiments in Section 5.5, including the newsvendor problem, mean-risk portfolio optimization, and multi-class adversarial classification. Using synthetic and real datasets, we compare the Sinkhorn DRO model against benchmarks such as SAA, Wasserstein DRO, and KL-divergence DRO. The results demonstrate

that the Sinkhorn DRO model consistently outperforms the benchmarks in terms of out-of-sample performance and computational speed. We also provide a comprehensive set of experiment studies to show that there exists a large number of parameter choices under which Sinkhorn DRO outperforms Wasserstein DRO.

Related Literature

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

Feng and Schlögl [112] studied the Wasserstein DRO formulation with an additional differential entropy constraint on the optimal transport mapping, which is closely related to our Sinkhorn DRO formulation. They derived a weak dual formulation and characterized the worst-case distribution under the assumption that strong duality holds. It is important to note that such an assumption cannot be taken for granted for the considered infinite-dimensional problem. Instead, we provided a rigorous proof of strong duality for our Sinkhorn DRO formulation. Moreover, their results heavily depend on the assumption that the nominal distribution $\widehat{\mathbb{P}}$ is absolutely continuous with respect to the Lebesgue measure. This limits the applicability of their formulation in data-driven settings where $\widehat{\mathbb{P}}$ is discrete. Since the initial submission of our work, Azizian et al. [12] have presented a duality result similar to ours, but with different assumptions. Their results apply to more general regularization beyond entropic regularization, but they assume a continuous loss function and a compact probability space under the Slater condition. Song et al. [289] have recently explored the application of Sinkhorn DRO in reinforcement learning. Their duality proof rely on the boundedness of the loss function and the discreteness of the probability support. These three papers do not present numerical algorithms to solve the dual formulation. Blanchet and Kang [42, Section 3.2] solved a log-sum-exp approximation of the Wasserstein DRO dual formulation. This smooth approximation can be viewed as a special case of the dual reformulation of our Sinkhorn DRO model. However, their study did not specifically explore the primal

form of Sinkhorn DRO. Their algorithm employed unbiased subgradient estimators, even though the second-order moment could be unbounded. The paper did not provide explicit theoretical convergence guarantees for their algorithm. Additionally, numerical comparisons detailed in Appendix D.2.1 suggest that our proposed algorithm outperforms theirs in terms of empirical performance.

Next, we review papers on several related topics.

On DRO models. In the literature on DRO, there are two main approaches to constructing ambiguity sets. The first approach involves defining ambiguity sets based on descriptive statistics, such as support information [32], moment conditions [34, 74, 94, 135, 265, 318, 338], shape constraints [253, 298], marginal distributions [2, 105, 117, 232], etc. The second approach, which has gained popularity in recent years, involves considering distributions within a pre-specified statistical distance from a nominal distribution. Commonly used statistical distances in the literature include ϕ-divergence [22, 24, 108, 161, 315], Wasserstein distance [44, 73, 125, 226, 249, 319, 324, 336], and maximum mean discrepancy [290, 337]. Our proposed Sinkhorn DRO can be seen as a variant of the Wasserstein DRO. In the literature on Wasserstein DRO, researchers have also explored the regularization effects and statistical inference of the approach. In particular, it has been shown that Wasserstein DRO is asymptotically equivalent to a statistical learning problem with variation regularization [43, 122, 274]. When the radius is chosen properly, the worst-case loss of Wasserstein DRO serves as an upper confidence bound on the true loss [43, 45, 46, 120]. Variants of Wasserstein DRO have been proposed by combining it with other information, such as moment information [301] or marginal distributions [109] to enhance its modeling capabilities.

On Sinkhorn distance. Sinkhorn distance [92] was proposed to improve the computational complexity of Wasserstein distance, by regularizing the original mass transportation problem with relative entropy penalty on the transport plan. It has been demonstrated to be beneficial because of lower computational cost in various applications, including domain adaptations [85, 87, 88], generative modeling [131, 208, 244, 246], dimension

reduction [163, 199, 308], etc. In particular, this distance can be computed from its dual form by optimizing two blocks of decision variables alternatively, which only requires simple matrix-vector products and therefore significantly improves the computation speed [6, 200, 218, 247]. Such an approach first arises in economics and survey statistics [14, 95, 180, 333], and its convergence analysis is attributed to the mathematician Sinkhorn [288], which gives the name of Sinkhorn distance. Computing the Sinkhorn distance between a discrete distribution and an arbitrary distribution can be reformulated as a stochastic optimization problem with a log-sum-exp–type loss function [247, Section 5.4]. For Sinkhorn DRO, the dual objective takes the form of an expectation involving the logarithm of a conditional expectation of an exponential function. When the inner expectation is over a discrete distribution, the problem retains a similar structure and can be effectively addressed using standard stochastic optimization techniques. In the case of general (non-discrete) distributions, the formulation becomes more challenging to solve, primarily due to the difficulty of obtaining unbiased (sub)gradient estimators.

On algorithms for solving DRO models. In the introduction, we have elaborated on the literature that proposes efficient optimization algorithms for solving the Wasserstein DRO dual formulation [47, 72, 111, 125, 203, 272, 273, 275, 286, 332, 336], in which the computational efficiency is limited to a certain class of loss functions, transport costs, and nominal distributions. To solve the ϕ-divergence DRO, one common approach is to employ sample average approximation (SAA) to approximate the dual formulation. However, SAA requires storing the entire set of samples, making it inefficient in terms of storage usage. An alternative approach is to use first-order stochastic subgradient algorithms, which are more storage-efficient. These algorithms have the advantage of complexity that can be independent of the sample size of the nominal distribution [191, 231, 255]. Our derived dual reformulation of Sinkhorn DRO can be seen as an instance of the CSO problem [154, 156, 160]. In this context, we have developed stochastic mirror descent algorithms with biased subgradient oracles. Notably, our proof can be adjusted to show that the proposed

algorithm achieves near-optimal complexity for general CSO problems with both smooth and nonsmooth loss functions, marking an improvement over the state-of-the-art [160, Theorem 3.2] that is sub-optimal for nonsmooth loss functions.

The rest of the paper is organized as follows. In Section 5.2, we describe the main formulation for the Sinkhorn DRO model. In Section 5.3, we develop its strong dual reformulation. In Section 5.4, we propose a first-order optimization algorithm that solves the reformulation efficiently. We report several numerical results in Section 5.5, and conclude the paper in Section 5.6. All omitted proofs can be found in Appendices.

5.2 Model Setup

Notation. Assume the logarithm function $\log$ is taken with base e. For a positive integer N, we write $[N]$ for $\{1, 2, \ldots, N\}$. For a measurable set $\mathcal{Z}$, denote by $\mathcal{M}(\mathcal{Z})$ the set of measures (not necessarily probability measures) on $\mathcal{Z}$, and $\mathcal{P}(\mathcal{Z})$ the set of probability measures on $\mathcal{Z}$. Given a probability distribution $\mathbb{P}$ and a measure μ, we denote $\operatorname{supp}\mathbb{P}$ the support of $\mathbb{P}$, and write $\mathbb{P} \ll \mu$ if $\mathbb{P}$ is absolutely continuous with respect to μ. Given a measure $\mu \in \mathcal{M}(\mathcal{Z})$ and a measurable variable $f : \mathcal{Z} \to \mathbb{R}$, we write $\mathbb{E}_{z\sim\mu}[f(z)]$ for $\int f(z)\,\mathrm{d}\mu(z)$. For a given element x, denote by δ_x the one-point probability distribution supported on $\{x\}$. Denote $\mathbb{P} \otimes \mathbb{Q}$ as the product measure of two probability measures $\mathbb{P}$ and $\mathbb{Q}$. Denote by $\mathrm{Proj}_{1\#}\gamma$ and $\mathrm{Proj}_{2\#}\gamma$ the first and the second marginal distributions of γ, respectively. For a 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 11 (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 transport cost, and* $H(\gamma \mid \mu\otimes\nu)$ *denotes the relative entropy of* γ *with respect to the product measure* $\mu\otimes\nu$*:*

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

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

Remark 11 (Variants of Sinkhorn Distance)**.** *Sinkhorn distance in Definition 11 is based on general reference measures* μ *and* ν*. Special forms of distance have been investigated in the literature. For instance, the entropic regularized optimal transport distance* $\mathcal{W}_\epsilon^{\mathsf{Ent}}(\mathbb{P},\mathbb{Q})$ *[92, Equation (2)] chooses* μ *and* ν *as the Lebesgue measure when the corresponding* $\mathbb{P}$ *and* $\mathbb{Q}$ *are continuous, or counting measures if* $\mathbb{P}$ *and* $\mathbb{Q}$ *are discrete. For given* $\mathbb{P}$ *and* $\mathbb{Q}$*, one can check the two distances above are equivalent up to a constant:*

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

Another variant is to chose μ *and* ν *to be* $\mathbb{P},\mathbb{Q}$*, respectively [129, Section 2]. A hard-constrained variant of the relative entropy regularization has been discussed in [92, Definition 1] and [15]:*

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

where $R \ge 0$ *quantifies the upper bound for the relative entropy between distributions* γ

and $\mathbb{P} \otimes \mathbb{Q}$. ♣

Remark 12 (Choice of Reference Measures)**.** *We discuss below our choice of the two reference measures* μ *and* ν *in Definition 11. For the reference measure* μ*, observe from the definition of relative entropy and the law of probability, we can see that the regularization term in* $\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})$ *can be written as*

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

Therefore, any choice of the reference measure μ *satisfying* $\widehat{\mathbb{P}} \ll \mu$ *is equivalent up to a constant. For simplicity, in the sequel we will take* $\mu = \widehat{\mathbb{P}}$*. For the reference measure* ν*, observe that the worst-case solution* $\mathbb{P}$ *in* (Primal) *should satisfy that* $\mathbb{P} \ll \nu$ *since otherwise the entropic regularization in Definition 11 is undefined. As a consequence, we can choose* ν *which the underlying true distribution is absolutely continuous with respect to and is easy to sample from. For example, if we believe the underlying distribution is continuous, then we can choose* ν *to be the Lebesgue measure or Gaussian measure, or if we believe the underlying distribution is discrete, we can choose* ν *to be a counting measure. We refer to [250, Section 3.6] for the construction of a general reference measure.* ♣

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

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

where $\mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}}) := \left\{\mathbb{P} :\ \mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P}) \le \rho\right\}$ is the Sinkhorn ball of the radius ρ centered at the nominal distribution $\widehat{\mathbb{P}}$. Due to the convex entropic regularization in $\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})$ [89], the

Sinkhorn distance $\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})$ is convex in $\mathbb{P}$, i.e., it holds that $\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \lambda\mathbb{P}_1 + (1-\lambda)\mathbb{P}_2) \leq \lambda\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P}_1) + (1-\lambda)\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P}_2)$ for all probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ and all $0 \leq \lambda \leq 1$. Therefore, the Sinkhorn ball is a convex set, and the problem (`Primal`) is an (infinite-dimensional) convex program.

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

5.3 Strong Duality Reformulation

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

$$V_{\mathrm{D}} := \inf_{\lambda \geq 0} \left\{ \lambda\rho + \lambda\epsilon\, \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\log \mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))/(\lambda\epsilon)}\big]\Big] \right\}, \tag{5.1}$$

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

$$\overline{\rho} := \rho + \epsilon\, \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\log \mathbb{E}_{z\sim\nu}\big[e^{-c(x,z)/\epsilon}\big]\Big], \tag{5.2}$$

and the kernel probability distribution

$$\mathrm{d}\mathbb{Q}_{x,\epsilon}(z) := \frac{e^{-c(x,z)/\epsilon}}{\mathbb{E}_{u\sim\nu}\left[e^{-c(x,u)/\epsilon}\right]}\, \mathrm{d}\nu(z), \tag{5.3}$$

we have

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

The rest of this section is organized as follows. In Section 5.3.1, we summarize our main results on the strong duality reformulation of Sinkhorn DRO. Next, we provide detailed examples in Section 5.3.3 and discussions in Section 5.3.2. In Section 5.3.4, we provide a proof sketch of our main results.

5.3.1 Main Theorem

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

Assumption 4. (I) *The cost function $c(x,z)$ is $\widehat{\mathbb{P}}\otimes\nu$-measurable and satisfies $\nu\{z:\ 0\le c(x,z)<\infty\}=1$ for $\widehat{\mathbb{P}}$-almost every x;*

(II) $\mathbb{E}_{z\sim\nu}\left[e^{-c(x,z)/\epsilon}\right]<\infty$ *for $\widehat{\mathbb{P}}$-almost every x;*

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

(IV) *For every joint distribution γ on $\mathcal{Z}\times\mathcal{Z}$ with first marginal distribution $\widehat{\mathbb{P}}$, it has a regular conditional distribution γ_x given the value of the first marginal equals x.*

Assumption 4(I) implies that $0\le c(x,y)<\infty$ for $\widehat{\mathbb{P}}\otimes\nu$-almost every (x,y). By [247, Proposition 4.1], the Sinkhorn distance has an equivalent formulation

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

Therefore Assumption 4(I) ensures that the reference measure $\mathcal{K}$ is well-defined. Assumption 4(II) ensures the optimal transport mapping γ_* for Sinkhorn distance $\mathcal{W}_\epsilon(\widehat{\mathbb{P}},\mathbb{P})$ exists with density value $\frac{\mathrm{d}\gamma_*(x,y)}{\mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}\nu(y)}\propto e^{-c(x,y)/\epsilon}$. Hence, Assumptions 4(I) and 4(II) together ensure the Sinkhorn distance is well-defined. Assumption 4(III) ensures the expected loss $\mathbb{E}_{z\sim\mathbb{P}}[f(z)]$ to be well-defined for any distribution $\mathbb{P}$. Assumption 4(IV) ensures the joint distribution γ can be written as $\mathrm{d}\gamma(x,z)=\mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}\gamma_x(z)$ and the law of total expectation

holds; we refer to [173, Chapter 5] for the concept of the regular conditional distribution. Such an assumption is very mild; for instance, it holds if $\mathcal{Z}$ is a Polish space [38].

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

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

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

Theorem 20 (Strong Duality)**.** *Let $\widehat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$, and assume Assumption 4 holds. Then the following holds:*

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

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

(III) *If, in addition, Condition 1 holds and $\overline{\rho} > 0$, it holds that $V = V_{\mathrm{D}} < \infty$; otherwise $V = V_{\mathrm{D}} = \infty$.*

(IV) *Assume in addition that Condition 1 holds and $\overline{\rho} > 0$. Define the event $A := \{z : \ f(z) = \operatorname{ess\,sup}_{\nu}(f)\}$ with $\operatorname{ess\,sup}_{\nu}(f) := \inf\{t : \ \nu\{f(z) > t\} = 0\}$. The dual minimizer $\lambda^* = 0$ if and only if $\operatorname{ess\,sup}_{\nu}(f) < \infty$ and $\overline{\rho} \geq \epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}[\log(1/\mathbb{P}_{z\sim\mathbb{Q}_{x,\epsilon}}\{A\})]$.*

We remark that if $\overline{\rho} < 0$, by convention, $V = -\infty$ and $V_D = -\infty$ as well by Lemma 25 in Appendix D.5. Therefore, we have $V = V_D$ as long as Assumption 4 holds.

5.3.2 Discussions

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

Remark 13 (Comparison with Wasserstein DRO)**.** *As the regularization parameter $\epsilon \to 0$, the dual objective of the Sinkhorn DRO* (`Dual`) *converges to (see Appendix D.4 for details)*

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

which essentially follows from the fact that the log-sum-exp function is a smooth approximation of the supremum. Particularly, when $\operatorname{supp}\nu = \mathcal{Z}$*, the dual objective of the Sinkhorn DRO converges to the dual objective of the Wasserstein DRO [125, Theorem 1]. The main computational difficulty in Wasserstein DRO is solving the maximization problem inside the expectation above. All results in Table 5.1 ensure this inner maximization can be efficiently solved. On the one hand, as Sinkhorn DRO does not need to solve this maximization, it does not need stringent assumptions on* $f(\cdot)$ *and thus enables efficient implementation for a larger class of loss functions for a fixed regularization parameter* ϵ *(see detailed discussion in Section 5.4). On the other hand, when the stringent assumptions on* $f(\cdot)$ *are satisfied, (data-driven) Wasserstein DRO typically admits a finite-dimensional convex reformulation, which can be solved more efficiently than Sinkhorn DRO (see our numerical comparison of CPU times in Appendix D.2.3). The key reason is that, in these special cases, Wasserstein DRO yields a finitely supported worst-case distribution, whereas Sinkhorn DRO always results in a worst-case distribution supported over the entire sample space.*

We also remark that Sinkorn DRO and Wasserstein DRO result in different conditions for finite worst-case values. From Condition 1 we see that Sinkhorn DRO is finite if and only if under a light-tail condition on f*, while Wasserstein DRO is finite if and only if the loss function satisfies a growth condition [125, Theorem 1 and Proposition 2]:* $f(z) \leq L_f c(z, z_0) + M, \forall z \in \mathcal{Z}$ *for some constants* $L_f, M > 0$ *and* $z_0 \in \mathcal{Z}$*.* ♣

Remark 14 (Worst-case Distribution)**.** *Assume* $\overline{\rho} > 0$ *and Condition 1 holds, and there exists (actually, Lemma 5 ensures its uniqueness) an optimal Lagrangian multiplier* $\lambda^* > 0$ *in* (`Dual`)*. As we will demonstrate in the proof of Theorem 20, the worst-case distribution maps every* $x \in \operatorname{supp}\widehat{\mathbb{P}}$ *to a (conditional) distribution* γ_x^* *that solves a strictly convex program (i.e., Problem* (5.8)*), whose density function (with respect to* ν*) is*

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

where $\alpha_x := \left(\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)}\right]\right)^{-1}$ *is a normalizing constant to ensure the conditional distribution well-defined. The uniqueness of* $\gamma_x^*, x \in \operatorname{supp}\widehat{\mathbb{P}}$ *ensures the uniqueness of the worst-case distribution* $\mathbb{P}_*$*, whose density becomes*

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

As such, the worst-case distribution shares the same support as the measure ν*.*

Particularly, when $\widehat{\mathbb{P}}$ *is the empirical distribution* $\frac{1}{n}\sum_{i=1}^n \delta_{\hat{x}_i}$ *and* ν *is any continuous distribution on* $\mathbb{R}^d$*, the worst-case distribution* $\mathbb{P}_*$ *is supported on the entire* $\mathbb{R}^d$*. In contrast, the worst-case distribution for Wasserstein DRO is supported on at most* $n+1$ *points [125]. In Fig. 5.1 we visualize the worst-case distributions from Wasserstein/Sinkhorn DRO models. The loss function and transport cost used in this plot follow the setup described in Example 4. The Wasserstein ball radius, Sinkhorn ball radius, and entropic regularization value are fine-tuned to ensure that the optimal dual multipliers for all instances equal 5. Notably, the support points of the worst-case distributions from the Wasserstein DRO model correspond to the modes of the continuous worst-case distributions from the Sinkhorn DRO model.*

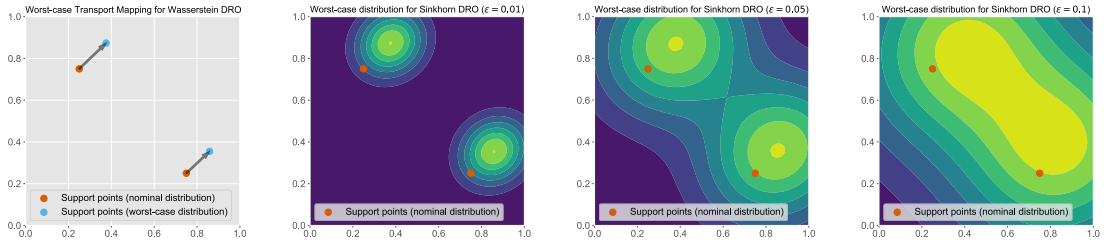

Figure 5.1: Visualization of worst-case distributions from Wasserstein DRO (left plot) and Sinkhorn DRO models (right three plots) with varying choices of ϵ.

The above demonstrates another difference (perhaps advantage) of Sinkhorn DRO compared with Wasserstein DRO. Indeed, for many practical problems, the underlying distribution is modeled as a continuous distribution. The worst-case distribution for Wasserstein DRO is often finitely supported, raising the concern of whether it hedges against the wrong

family of distributions and thus results in suboptimal solutions. The numerical results in Section 5.5 demonstrate some empirical advantages of Sinkhorn DRO. ♣

Remark 15 (Connection with KL-divergence DRO)**.** *Using Jensen's inequality, we can see that the dual objective function of the Sinkhorn DRO model is upper-bounded as*

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

which corresponds to the dual objective for the following KL-divergence DRO [24]

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

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

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

where $K_\epsilon(x) \propto \exp(-\|x\|_2^2/\epsilon)$ *represents the Gaussian kernel. By Lemma 4 and divergence inequality [89, Theorem 2.6.3], we can see the Sinkhorn DRO with* $\overline{\rho} = 0$ *is reduced to the following SAA model based on the distribution* $\mathbb{P}^0$*:*

$$V = \mathbb{E}_{z\sim\mathbb{P}^0}[f(z)] = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[f(z)]. \tag{5.4}$$

In non-parameteric statistics, the optimal bandwidth to minimize the mean-squared-error between the estimated distribution $\mathbb{P}_0$ *and the underlying true one is at rate* $\epsilon = O(n^{-1/(d+4)})$ *[143, Theorem 4.2.1]. However, such an optimal choice for the kernel density estimator may not be the optimal choice for optimizing the out-of-sample performance of the Sinkhorn DRO. In our numerical experiments in Section 5.5, we select* ϵ *based on cross-validation unless otherwise stated.*

We also note that data-driven KL-divergence DRO is typically more efficient to solve than Sinkhorn DRO. This is because KL-divergence DRO yields a worst-case distribution supported on the same set as the empirical observations, leading to a finite-dimensional convex reformulation that can be efficiently solved using off-the-shelf solvers (see the numerical study of CPU times in Appendix D.2.3). ♣

Remark 16 (Connection with Bayesian DRO)**.** *Bayesian DRO [281] proposed to solve*

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

where $\widehat{\mathbb{P}}$ is a special posterior distribution constructed from collected observations, and the ambiguity set $\mathcal{P}_x$ is typically constructed as a KL-divergence ball, i.e., $\mathcal{P}_x := \{\mathbb{P}:\ D_{KL}(\mathbb{P}\|\mathbb{Q}_x)\leq \eta\}$, with $\mathbb{Q}_x$ being the parametric distribution conditioned on x. According to [281, Section 2.1.3], a relaxation of the Bayesian DRO dual formulation is given by

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

When specifying the parametric distribution $\mathbb{Q}_x$ as the kernel probability distribution in (5.3) *and applying the change-of-variable technique such that λ is replaced with $\lambda\epsilon$, this relaxed formulation becomes*

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

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

5.3.3 Examples

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

Example 3 (Linear loss)**.** *Suppose that the loss function $f(z) = a^{\mathrm{T}} z$, support $\mathcal{Z} = \mathbb{R}^d$, ν is the corresponding Lebesgue measure, and the transport cost is the Mahalanobis distance, i.e., $c(x, y) = \frac{1}{2}(x - y)^{\mathrm{T}}\Omega(x - y)$, where Ω is a positive definite matrix. In this case, the kernel probability distribution $\mathbb{Q}_{x,\epsilon} = \mathcal{N}(x, \epsilon\Omega^{-1})$., and the dual problem can be written as*

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

where

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

Therefore

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

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

Example 4 (Quadratic loss)**.** *Consider the example of linear regression with quadratic loss $f_\theta(z) = (a^{\mathrm{T}}\theta - b)^2$, where $z := (a, b)$ denotes the predictor-response pair, $\theta \in \mathbb{R}^d$ denotes the fixed parameter choice, and $\mathcal{Z} = \mathbb{R}^{d+1}$. Taking ν as the Lebesgue measure and the transport cost as $c((a, b), (a', b')) = \frac{1}{2}\|a - a'\|_2^2 + \infty|b - b'|$. In this case, the dual problem becomes*

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

In comparison with the corresponding Wasserstein DRO formulation with radius ρ (see, e.g.,

[45, Example 4])

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

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

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

Corollary 3 (Conic Reformulation for Finite Support)**.** *Suppose that the support contains* $L_{\max}$ *elements, i.e.,* $\mathcal{Z} = \{z_\ell\}_{\ell=1}^{L_{\max}}$*, and the nominal distribution* $\widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^n \delta_{\hat{x}_i}$*. If Condition 1 holds and* $\overline{\rho} \geq 0$*, the dual problem* (Dual) *can be formulated as the following conic optimization:*

$$\begin{array}{lll} V_{\mathrm{D}} = \min\limits_{\substack{\lambda\geq 0, s\in\mathbb{R}^n, \\ a\in\mathbb{R}^{n\times L}}} & \lambda\overline{\rho} + \dfrac{1}{n}\displaystyle\sum_{i=1}^n s_i & \\ \textit{s.t.} & \lambda\epsilon \geq \displaystyle\sum_{\ell=1}^{L_{\max}} q_{i,\ell} a_{i,\ell}, i \in [n], & (5.5) \\ & (\lambda\epsilon, a_{i,\ell}, f(z_\ell) - s_i) \in \mathcal{K}_{\exp}, i \in [n], \ell \in [L]. & \end{array}$$

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* (5.3)*, and* $\mathcal{K}_{\exp}$ *denotes the exponential cone* $\mathcal{K}_{\exp} := \{(\nu, \lambda, \delta) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R} : \exp(\delta/\nu) \leq \lambda/\nu\}$.

Problem (5.5) is a convex program that minimizes a linear function with respect to linear and conic constraints, which can be solved using interior point algorithms [234, 299]. We will develop an efficient first-order optimization algorithm in Section 5.4 that is able to solve a more general problem (without a finite support).

5.3.4 Proof of Theorem 20

In this subsection, we present a sketch of the proof for Theorem 20. We begin with the weak duality result in Lemma 3, which can be shown by applying the Lagrangian weak duality.

Lemma 3 (Weak Duality)**.** *Under Assumption 4, it holds that*

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

Proof. Proof of Lemma 3. Based on Definition 11 of Sinkhorn distance, we reformulate V as

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

By Assumption 4, the constraint is equivalent to

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

and the primal problem is equivalent to

$$V = \sup_{\{\gamma_x\}_{x\in\mathrm{supp}\,\widehat{\mathbb{P}}}\subset\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}[f(z)] : \ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}\left[c(x,z) + \epsilon \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)} \right)\right] \leq \rho \right\}. \tag{5.6}$$

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

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

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

$$V \leq \inf_{\lambda\geq 0} \left\{ \lambda\rho + \sup_{\{\gamma_x\}_{x\in\mathrm{supp}\,\widehat{\mathbb{P}}}\subset\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\mathbb{E}_{z\sim\gamma_x}\left[f(z) - \lambda c(x,z) - \lambda\epsilon \log\left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)} \right)\right]\right\}\right\}. \tag{5.7}$$

For $x \in \operatorname{supp} \widehat{\mathbb{P}}$ and $\lambda \geq 0$, define

$$v_x(\lambda) := \sup_{\gamma_x \in \mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{z \sim \gamma_x} \left[f(z) - \lambda c(x,z) - \lambda \epsilon \log \left(\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}\nu(z)} \right) \right] \right\}. \tag{5.8}$$

Note that this function is measurable for any choice of λ (we omit its proof in Lemma 24). One can swap the supremum and the expectation operator on the right-hand-side of (5.7) to further upper bound it as

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

By Lemma 23, when there exists $\lambda > 0$ such that Condition 1 is satisfied, it holds that

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

and the desired result holds. Otherwise, for any $\lambda > 0$,

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

then intermediately we obtain

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

and the weak duality still holds. □ □

Next, we show the feasibility result in Theorem 20(I). The key observation is that the primal problem (`Primal`) can be reformulated as a generalized KL-divergence DRO problem. Consequently, Theorem 20(I) holds because of the non-negativity of KL-divergence.

Lemma 4 (Reformulation of (Primal))**.** *Under Assumption 4, it holds that*

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

Remark 17 (Comparison with Infinite-dimensional Convex Analysis)**.** *If assuming that* $\widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ *is finitely supported, by Lemma 4,* (Primal) *can be reformulated as a conic linear program*

$$V = \sup_{\{\gamma_i\}_{i\in[n]}\subset\mathcal{P}(\mathcal{Z})} \left\{ \frac{1}{n}\sum_{i\in[n]}\mathbb{E}_{z\sim\gamma_i}[f(z)] :\ \frac{1}{n}\sum_{i\in[n]} D_{KL}(\gamma_i\|\mathbb{Q}_{x_i,\epsilon}) \le \frac{\overline{\rho}}{\epsilon} \right\}.$$

Thus, strong duality from infinite-dimensional convex analysis (e.g., [278]) can be applied to show Theorem 20. However, our strong duality proof, as described below, differs from this one in several key aspects. First, our approach imposes less restrictive assumptions, holding for any measurable sample space $\mathcal{Z}$*, measurable loss function* f*, and nominal distribution* $\widehat{\mathbb{P}}$*. The strong duality result in [278] requires* $\mathcal{Z}$ *to be convex,* f *to be upper semicontinuous, and* $\widehat{\mathbb{P}}$ *to be finitely supported. Second, our approach is constructive: we explicitly characterize the worst-case distribution for Sinkhorn DRO, whereas a nonconstructive method was employed in [278]. Third, our approach provides a byproduct — an explicit necessary and sufficient condition for when the Sinkhorn ambiguity constraint is binding (Theorem 20*(IV)*). This insight offers practical guidance on choosing the ambiguity set size to avoid over-conservativeness.* ♣

Finally, we develop the strong duality. The general proof idea involves deriving the optimality condition of the dual minimizer, which then guide the construction of the worst-case distribution of (Primal). In the following, we provide the proof of the first part of Theorem 20(III) for the most representative case where $\overline{\rho} > 0$, the dual minimizer λ^* exists with $\lambda^* > 0$, and Condition 1 holds. Proofs of other cases are moved in Appendix D.5.

We first develop the optimality condition when the dual minimizer $\lambda^* > 0$, by setting the derivative of the dual objective function to zero.

Lemma 5 (First-order Optimality Condition when $\lambda^* > 0$)**.** *Suppose $\overline{\rho} > 0$ and Condition 1 is satisfied, and assume further that there exists a dual minimizer $\lambda^* > 0$, then the dual minimizer is unique and λ^* satisfies*

$$\frac{1}{\lambda^*}\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\frac{\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)} f(z)\right]}{\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)}\right]}\right] - \epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log \mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)}\right]\right] = \rho. \tag{5.9}$$

Proof. Proof of Theorem 20(III) for the case where Condition 1 holds and $\overline{\rho} > 0, \lambda^* > 0$. We take the transport mapping γ_* such that

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

and $\alpha_x := \left(\mathbb{E}_{z\sim\nu}\left[e^{(f(z)-\lambda^* c(x,z))/(\lambda^*\epsilon)}\right]\right)^{-1}$ is a normalizing constant such that $\mathrm{Proj}_{1\#}\gamma_* = \widehat{\mathbb{P}}$. Also define the primal (approximate) optimal distribution $\mathbb{P}_* := \mathrm{Proj}_{2\#}\gamma_*$. Recall the expression of the Sinkhorn distance in Definition 11, one can verify that

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

where the inequality relation is because γ_* is a feasible solution in $\Gamma(\widehat{\mathbb{P}}, \mathbb{P}_*)$, and the last two relations are by substituting the expression of γ_*. Since $\overline{\rho} > 0$ and the dual minimizer $\lambda^* > 0$, the optimality condition in (5.9) holds, which implies that $\mathcal{W}_\epsilon(\widehat{\mathbb{P}}, \mathbb{P}_*) \le \rho$, i.e., the

distribution $\mathbb{P}_*$ is primal feasible for the problem (Primal). Moreover, we can see that the primal optimal value is lower bounded by the dual optimal value:

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

where the third equality is by substituting the expression of γ_*, and the last equality is based on the optimality condition in (5.9). This, together with the weak duality, completes the proof. □ □

5.4 Efficient First-order Algorithm for Sinkhorn Robust Optimization

Consider the Sinkhorn robust optimization problem

$$\inf_{\theta\in\Theta}\sup_{\mathbb{P}\in\mathbb{B}_{\rho,\epsilon}(\widehat{\mathbb{P}})}\mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)]. \tag{5.10}$$

Here the feasible set $\Theta\subseteq\mathbb{R}^{d_\theta}$ is *closed and convex* containing all possible candidates of decision vector θ, and the Sinkhorn uncertainty set is centered around a given nominal distribution $\widehat{\mathbb{P}}$. Based on our strong dual expression (Dual), we reformulate (5.10) as

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

where the constant $\overline{\rho}$ and the distribution $\mathbb{Q}_{x,\epsilon}$ are defined in (5.2) and (5.3), respectively. In Examples 3 and 4, we have seen special instances of (D) where we can get closed-form expressions for the above integration. In this section, we develop an efficient algorithm for solving (D) for general loss functions where a closed-form expression is not available.

A typical approach for solving a stochastic optimization is the stochastic (sub)gradient

method such as stochastic mirror descent (SMD) [39]. Unlike many other stochastic optimization problems, one salient feature of (D) is that its inner objective involves a nonlinear transformation of the expectation. Consequently, based on a batch of simulated samples from $\mathbb{Q}_{x,\epsilon}$, an unbiased subgradient estimate could be challenging to obtain. In Section 5.4.1, we will combine SMD with biased subgradient estimators and bisection search to solve (D). We will analyze its computational complexity in Section 5.4.2.

5.4.1 Algorithm Framework

We define

$$F(\theta;\lambda) := \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_\theta(z)/(\lambda\epsilon)}\right]\right], \tag{5.11}$$

and define the objective value of the outer minimization in (D) as

$$\Psi(\lambda) := \lambda\overline{\rho} + \inf_{\theta\in\Theta} F(\theta;\lambda). \tag{5.12}$$

Solving (D) involves determining the optimal Lagrange multiplier λ of the function Ψ, where evaluating Ψ requires solving the minimization problem in (5.12) to obtain the optimal decision θ. We first introduce a biased SMD (BSMD) algorithm for finding the optimal decision θ given a fixed Lagrange multiplier λ in Section 5.4.1. Then, in Section 5.4.1, we present a bisection search algorithm to find the optimal Lagrange multiplier. Throughout this section, we assume the loss function $f_\theta(z)$ is convex in θ but it can be a potentially nonsmooth function. For any function $r(\theta)$ that is subdifferentiable in θ, we use the notation $\nabla_\theta r(\theta)$ to denote an arbitrary subgradient from its subdifferential, unless otherwise specified.

BSMD.

In this part, we omit the dependence of λ when defining objective or subgradient terms, e.g., we write $F(\theta)$ for $F(\theta;\lambda)$. We first introduce several notations that are standard in the mirror

descent algorithm. Let $\omega:\ \Theta \to \mathbb{R}$ be a distance generating function that is continuously differentiable and κ-strongly convex on Θ with respect to norm $\|\cdot\|$, which induces the Bregman divergence $D_\omega(\theta,\theta'):\ \Theta\times\Theta \to \mathbb{R}_+$: $D_\omega(\theta,\theta') = \omega(\theta') - \omega(\theta) - \langle \nabla\omega(\theta), \theta' - \theta\rangle$. Define the *prox-mapping* $\mathrm{Prox}:\ \mathbb{R}^{d_\theta} \to \Theta$ as

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

With these notations in hand, we present our algorithm in Algorithm 7, which iteratively obtains a biased stochastic (sub)gradient estimator and performs a proximal update.

Algorithm 7 BSMD for finding the optimal solution of (5.12) while fixing λ

Require: Maximal iteration T, constant step size h, initial guess θ_0, fixed multiplier λ.

1: **for** $t = 0, 1, \ldots, T-1$ **do**
2: Construct a (biased) subgradient estimator $v(\theta_t)$ of $F(\theta_t)$ using (5.15) or (5.17).
3: Update $\theta_{t+1} = \mathrm{Prox}_{\theta_t}\big(hv(\theta_t)\big)$.
4: **end for**

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

At the core of Algorithm 7 lies the crucial task of efficiently simulating the subgradient estimator in Step 2. It is noteworthy that the minimization in (5.12) is a special conditional stochastic optimization (CSO), as studied in [154, 156, 160]. CSO typically has the formulation

$$\min_{\theta\in\Theta}\ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}[H^1(\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[H^2(\theta;z)])], \tag{5.13}$$

and we specify $H^1(\cdot) = \lambda\epsilon\log(\cdot)$ and $H^2(\cdot;z) = \exp(f_\cdot(z)/(\lambda\epsilon))$ to convert (5.12) into (5.13). This structure suggests that ideas from CSO-related literature, particularly multi-level Monte Carlo (MLMC) estimators, can be applied to generate biased subgradient estimators with controlled bias, enhancing computational efficiency in Step 2. Our framework and analysis differs from the aforementioned references in several aspects, and see the discussion in Remark 23.

To generate the subgradient estimator, we first construct a function $F^\ell(\theta), \ell \in \mathbb{N}$ that

approximate the original objective function $F(\theta)$ with $\mathcal{O}(2^{-\ell})$-gap:

$$F^{\ell}(\theta) = \mathbb{E}_{x^{\ell}\sim\widehat{\mathbb{P}}}\, \mathbb{E}_{\{z_j^{\ell}\}_{j\in[2^{\ell}]}\sim\mathbb{Q}_{x^{\ell},\epsilon}}\left[\lambda\epsilon \log\Big(\frac{1}{2^{\ell}}\sum_{j\in[2^{\ell}]} e^{f_{\theta}(z_j^{\ell})/(\lambda\epsilon)}\Big)\right], \tag{5.14}$$

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

$$U_{n_1:n_2}(\theta,\zeta^{\ell}) := \lambda\epsilon\log\Big(\frac{1}{n_2-n_1+1}\sum_{j\in[n_1:n_2]} e^{f_{\theta}(z_j^{\ell})/(\lambda\epsilon)}\Big),$$
$$A^{\ell}(\theta,\zeta^{\ell}) := U_{1:2^{\ell}}(\theta,\zeta^{\ell}) - \frac{1}{2}U_{1:2^{\ell-1}}(\theta,\zeta^{\ell}) - \frac{1}{2}U_{2^{\ell-1}+1:2^{\ell}}(\theta,\zeta^{\ell}).$$

We take $U_{n_1:n_2}(\theta,\zeta^{\ell}) = 0$ if $[n_1:n_2] = \emptyset$. For fixed θ and $\ell\in\mathbb{N}$, we define

$$g^{\ell}(\theta,\zeta^{\ell}) := \nabla_{\theta}U_{1:2^{\ell}}(\theta,\zeta^{\ell}), \qquad G^{\ell}(\theta,\zeta^{\ell}) := \nabla_{\theta}A^{\ell}(\theta,\zeta^{\ell}).$$

The random vector $g^{\ell}(\theta,\zeta^{\ell})$ is an unbiased estimator of $\nabla_{\theta}F^{\ell}(\theta)$, while the random vector $G^{\ell}(\theta,\zeta^{\ell})$ is an unbiased estimator of $\nabla_{\theta}F^{\ell}(\theta) - \nabla_{\theta}F^{\ell-1}(\theta)$. We note that computing $G^{\ell}(\theta,\zeta^{\ell})$ involves computing subgradient vectors $\nabla_{\theta}f_{\theta}(z_j^{\ell}), j\in[2^{\ell}]$, and we use the same subgradient computation across $U_{1:2^{\ell}}, U_{1:2^{\ell-1}}, U_{2^{\ell-1}+1:2^{\ell}}$ to facilitate the reduction of the second-order moment of $G^{\ell}(\theta,\zeta^{\ell})$. Using these components, we define two types of subgradient estimators below.

– *Stochastic (sub)Gradient (SG) Estimator:* Fix the maximum level $L\in\mathbb{N}_{+}$. We first generate the sample set ζ^{L} and next construct the SG estimator

$$v^{\mathrm{SG}}(\theta) = g^{L}(\theta,\zeta^{L}). \tag{5.15}$$

– *Randomized Truncation MLMC (RT-MLMC) Estimator [49]:* Fix the maximum level

$L \in \mathbb{N}_+$. We first sample a random level $\widehat{\ell}$ following a truncated geometric distribution

$$p_\ell := \Pr(\widehat{\ell} = \ell) = \frac{2^{-\ell}}{2 - 2^{-L}}, \ell = 0, 1, \ldots, L. \tag{5.16}$$

Next, we construct the RT-MLMC estimator

$$v^{\text{RT-MLMC}}(\theta) = p_{\widehat{\ell}}^{-1} \cdot G^{\widehat{\ell}}(\theta, \zeta^{\widehat{\ell}}). \tag{5.17}$$

Remark 18 (Sampling from $\mathbb{Q}_{x,\epsilon}$)**.** *Sampling from $\mathbb{Q}_{x,\epsilon}$ is crucial for the construction of subgradient estimators. In many cases, it is an easy task: When the transport cost $c(\cdot, \cdot) = \frac{1}{2}\| \cdot - \cdot \|_2^2$ and $\mathcal{Z} = \mathbb{R}^d$, the distribution $\mathbb{Q}_{x,\epsilon}$ becomes a Gaussian distribution $\mathcal{N}(x, \epsilon I_d)$. When the transport cost $c(\cdot, \cdot)$ is decomposable in each coordinate, we can apply the acceptance-rejection method [9] to generate samples in each coordinate independently, the complexity of which only increases linearly in the data dimension.* ♣

Bisection Search

In this part, we introduce a bisection search algorithm to solve the one-dimensional convex minimization problem in (D). The algorithm relies on an efficient oracle to estimate the objective value of (D). We first define this oracle in Algorithm 8: Given a fixed multiplier λ, it solves problem (5.12) using Algorithm 7 and then estimates the corresponding objective value. It has m independent repetitions, whose value will be determined later in Section 5.4.2 to achieve the optimal complexity.

To implement Step 3 of Algorithm 8, we again leverage RT-MLMC to efficiently estimate the objective in (5.11). For given (θ, λ), let m' denote the mini-batch size. For $i \in [m']$, we sample $\widehat{\ell}_i$ following the distribution defined in (5.16) and sample an i.i.d. copy of $\zeta^{\widehat{\ell}_i}$ that is

Algorithm 8 Evaluating the objective value of (D)

Require: Fixed multiplier λ, error tolerance δ, batch size m.

1: **for** $j = 1, 2, \ldots, m$ **do**
2: Obtain a δ-optimal solution $\widehat{\theta}_j$ of problem (5.12) using Algorithm 7.
3: Estimate the objective in (5.11) with $\theta \equiv \widehat{\theta}_j$ using RT-MLMC estimator (5.18), denoted as $\widehat{F}(\widehat{\theta}_j; \lambda)$.
4: **end for**

Output $\widehat{\Psi}(\lambda) := \lambda\overline{\rho} + \min_{\theta \in \{\widehat{\theta}_1, \ldots, \widehat{\theta}_m\}} \widehat{F}(\theta; \lambda)$.

denoted as $\zeta_i^{\widehat{\ell}_i}$. Next, we construct the objective estimator at the i-th trial as

$$\widehat{F}_i(\theta; \lambda) = p_{\widehat{\ell}_i}^{-1} \cdot A^{\widehat{\ell}_i}(\theta, \zeta_i^{\widehat{\ell}_i}; \lambda).$$

To reduce the variance of objective estimator, the final estimator of $F(\theta; \lambda)$ is constructed by averaging the outcomes over all trials, denoted as

$$\widehat{F}(\theta; \lambda) = \frac{1}{m'} \sum_{i=1}^{m'} \widehat{F}_i(\theta; \lambda). \tag{5.18}$$

Given an inexact objective oracle of (D) (e.g., by querying Algorithm 8), we use bisection search to find a near-optimal multiplier in (D); see Algorithm 9 for details. Unlike conventional bisection that relies on gradient information, this algorithm leverages an inexact objective oracle $\widehat{\Psi}$ to iteratively shrink the search interval. It begins by dividing the interval into five evenly spaced points and selecting the minimum among the three central points. In each iteration, it updates the left, middle, and right points based on the the current minimum (from among the three middle points) and its two nearest neighbors. Then, it adjusts the middle-left and middle-right points to maintain evenly spacing. The oracle is queried twice per iteration, reusing one value from the previous iteration to identify the new minimum efficiently. This algorithm is adopted from [83, Algorithm 8], but is more efficient, as the original algorithm only shrinks the interval by $1/3$ with each iteration, whereas one can improve it to factor $1/2$. See its performance guarantee in Theorem 22.

Algorithm 9 Bisection search for finding the optimal multiplier of (D)

Require: Interval $[\lambda_l, \lambda_u]$ such that $\lambda_l < \lambda^* < \lambda_u$, maximum iterations T'
Require: Inexact objective oracle $\widehat{\Psi}(\cdot) : \mathbb{R}_+ \to \mathbb{R}$
1: Update $\beta_i^{(0)} = \frac{5-i}{4}\lambda_l + \frac{i-1}{4}\lambda_u$ for $i = 1, \ldots, 5$ {*Divide interval using 5 grid points*}
2: Query oracle to obtain $\widehat{\Psi}(\beta_j^{(0)}), j = 2, 3, 4$
3: Specify $i^{(1)} = \arg\min_{j=2,3,4} \widehat{\Psi}(\beta_j^{(0)})$
4: **for** $t = 1, \ldots, T'$ **do**
5: Update $(\beta_1^{(t)}, \beta_3^{(t)}, \beta_5^{(t)}) = (\beta_{i^{(t)}-1}^{(t-1)}, \beta_{i^{(t)}}^{(t-1)}, \beta_{i^{(t)}+1}^{(t-1)})$ {*Move left, middle, right points*}
6: Update $(\beta_2^{(t)}, \beta_4^{(t)}) = \left(\frac{\beta_1^{(t)}+\beta_3^{(t)}}{2}, \frac{\beta_3^{(t)}+\beta_5^{(t)}}{2}\right)$ {*Move middle-left, middle-right points*}
7: Query oracle to obtain $\widehat{\Psi}(\beta_j^{(t)}), j = 2, 4$, and update $\widehat{\Psi}(\beta_3^{(t)}) = \widehat{\Psi}(\beta_{i^{(t)}}^{(t-1)})$.
8: Specify $i^{(t+1)} = \arg\min_{j=2,3,4} \widehat{\Psi}(\beta_j^{(t)})$
9: **end for**
Output approximate optimal multiplier $\beta_{i^{(T'+1)}}^{(T')}$.

An alternative approach to solving (D) is to jointly optimize (λ, θ) using BSMD. Although this approach is theoretically sound, updating λ could lead to oscillations or divergence if the associated stepsize is not carefully tuned, due to the high variance of gradient estimators when λ is small. In our algorithm, we have developed a bisection method to update λ, which requires only specifying the maximum iterations, initial interval, and an inexact objective oracle but does not require tuning the stepsize for updating λ. We also remark that, as a practical alternative, one can solve (5.12) using Algorithm 7 alone and tune the hyperparameter λ, as tuning the radius $\overline{\rho}$ is equivalent to tuning the Lagrangian multiplier λ in (5.12). This corresponds to the Sinkhorn robust learning problem with a soft Sinkhorn constraint.

5.4.2 Convergence Analysis

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

Assumption 5. (I) *(Convexity): The loss function $f_\theta(z)$ is convex in θ.*

(II) *(Lipschitz Continuity): For any 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) *(Boundedness): The loss function $f_\theta(z)$ satisfies $0 \leq f_\theta(z) \leq B$ for any $\theta \in \Theta$ and $z \in \mathcal{Z}$.*

Assumption 5(I) ensures the convexity of the objective in Sinkhorn robust optimization, and enables us to develop globally convergent optimization algorithms. Assumption 5(II) is crucial for establishing the bounded subgradient norm condition required by the BSMD algorithm and deriving its global convergence rate. Assumption 5(III) guarantees the Lipschitz continuity of the nonlinear operator $H^1(\cdot) = \lambda\epsilon \log(\cdot)$ in (5.11), ensuring that the objective in (5.14) approximates (5.11) with $\mathcal{O}(2^{-\ell})$ gap. This assumption is restrictive in practice, whereas one can replace it by the following conditions (see the argument in [191, footnote 2]): (III-1): Θ is bounded; (III-2): Assumption 5(II); and (III-3): $\inf_{\theta\in\Theta} f_\theta(z_1) - \inf_{\theta\in\Theta} f_\theta(z_2) \leq B_0, \forall z_1, z_2 \in \mathcal{Z}$.

We note that our algorithm is assumed to have access to two sampling oracles: (i) Oracle $\mathbf{O}(\widehat{\mathbb{P}})$ that generates a sample from $\widehat{\mathbb{P}}$; (ii) Oracle $\mathbf{O}(\mathbb{Q}_{x,\epsilon})$ that, based on the input $x \in \operatorname{supp}\widehat{\mathbb{P}}$, generates a sample from $\mathbb{Q}_{x,\epsilon}$. In practical implementations, the cost of generating samples from these two distributions can differ. In data-driven applications, sampling from $\widehat{\mathbb{P}}$ often reduces to randomly selecting observed data points, whereas sampling from $\mathbb{Q}_{x,\epsilon}$ (described in Remark 18) usually involves stochastic noises such as Gaussian. In the subsequent analysis, we report the sample complexities from $\widehat{\mathbb{P}}$ or $\mathbb{Q}_{x,\epsilon}$ individually. Based on our algorithm design (see, e.g., objective and subgradient estimators in (5.15), (5.17), (5.18)), it is also easy to check that the total computational time is roughly proportional to the sum of these two sample complexities.

Complexity of BSMD.

In this part, we discuss the complexity of Algorithm 7. We say θ is a δ-optimal solution if $\mathbb{E}[F(\theta;\lambda)] - F(\theta^*;\lambda) \leq \delta$, where θ^* is the optimal solution of (5.12). By properly tuning

hyper-parameters to balance the trade-off between bias and second-order moment of the subgradient estimate, we establish its performance guarantees in Theorem 21. The explicit constants and proof can be found in Appendix D.6. Let us define the constant $K_{\lambda,\epsilon,B} = \frac{B}{\lambda\epsilon}$ that depends on λ, ϵ, B.

Theorem 21. *Under Assumption 5, when using BSMD (Algorithm 7) to find a δ-optimal solution of* (5.12)*, the following results hold:*

(I) *If using SG subgradient estimator, the sample complexity from $\widehat{\mathbb{P}}$ is $\mathcal{O}(\delta^{-2})$, and that from $\mathbb{Q}_{x,\epsilon}$ is $\mathcal{O}(\lambda\epsilon \exp(2K_{\lambda,\epsilon,B}) \cdot \delta^{-3})$, with $\mathcal{O}(\cdot)$ hiding constants depending on L_f, θ_0, κ.*

(II) *If using RT-MLMC subgradient estimator, the sample complexity from $\widehat{\mathbb{P}}$ is*

$$\widetilde{\mathcal{O}}(K_{\lambda,\epsilon,B} \exp(4K_{\lambda,\epsilon,B}) \cdot \delta^{-2}),$$

and that from $\mathbb{Q}_{x,\epsilon}$ is

$$\widetilde{\mathcal{O}}(K^2_{\lambda,\epsilon,B} \exp(4K_{\lambda,\epsilon,B}) \cdot \delta^{-2}),$$

with $\widetilde{\mathcal{O}}(\cdot)$ hiding constants depending on L_f, θ_0, κ and linearly depending on $(\log \frac{\lambda\epsilon}{\delta})^2$.

Theorem 21 shows that the sample complexity from $\widehat{\mathbb{P}}$ of BSMD, whether using the SG or RT-MLMC subgradient estimator, is of the same order with respect to the error tolerance δ. This rate matches the known lower bound for general convex stochastic programming problems [39]. However, this complexity associated with the RT-MLMC estimator has a worse constant dependence on the parameters λ, ϵ, and B. Despite this, the RT-MLMC estimator has a lower-order sample complexity from $\mathbb{Q}_{x,\epsilon}$ compared to the SG estimator. Our numerical experiments in Appendices D.2.1 and D.2.2 further demonstrate that the RT-MLMC estimator exhibits a significantly faster empirical convergence rate than the SG estimator.

Remark 19 (Comparison with Biased Sample Average Approximation)**.** *An alternative way to solve* (5.12) *is to approximate the objective using finite samples for both expectations. This leads to a biased sample estimate, called Biased Sample Average Approximation (BSAA). Under Assumption 5 and apply [154, Corollary 4.2], it can be shown that for BSAA, the sample complexity from* $\widehat{\mathbb{P}}$ *is* $n_1 = \widetilde{\mathcal{O}}\left(d_\theta B^2 \exp(2K_{\lambda,\epsilon,B}) \cdot \delta^{-2}\right)$ *and that from* $\mathbb{Q}_{x,\epsilon}$ *is* $\mathcal{O}(\lambda\epsilon \exp(2K_{\lambda,\epsilon,B}) \cdot n_1 \cdot \delta^{-1})$*. Our proposed BSMD with the RT-MLMC-based subgradient estimator has smaller order of the sample complexity from* $\mathbb{Q}_{x,\epsilon}$ *(in terms of error tolerance* δ*). Also, the BSAA method still requires computing the optimal solution of the approximated optimization problem as the output. Hence, it typically takes considerably less time and memory to run the BSMD step rather than solving the BSAA formulation.* ♣

Complexity of Bisection Search

We first provide the complexity analysis for Algorithm 8, which produces an estimator of the objective value of the outer minimization in (D). Define the constant $H_{\lambda,\epsilon,B} = \max(\exp(2K_{\lambda,\epsilon,B}), \lambda^2\epsilon^2)$.

Proposition 5. *Let* $\eta \in (0,1)$ *and set the batch size* $m = \lceil \log_2 \frac{2}{\eta} \rceil$*. Assume Assumption 5 holds, and we choose hyper-parameters in Step 3 of Algorithm 8 as*

$$L = \left\lceil \log_2 \frac{2\lambda\epsilon \exp(2K_{\lambda,\epsilon,B})}{\delta} \right\rceil, \quad m' = \mathcal{O}(1) \frac{\lambda^2\epsilon^2 \exp(2K_{\lambda,\epsilon,B})(L+1)}{\delta^2} \cdot \log \frac{m}{\eta}.$$

With probability at least $1-\eta$*, the output in Algorithm 8 satisfies* $|\widehat{\Psi}(\lambda) - \Psi(\lambda)| \leq \delta$*. When using RT-MLMC subgradient estimator* (5.17) *in the BSMD step and RT-MLMC objective estimator* (5.18)*, the sample complexity from* $\widehat{\mathbb{P}}$ *is* $\widetilde{\mathcal{O}}\left(H_{\lambda,\epsilon,B} K_{\lambda,\epsilon,B} \exp(2K_{\lambda,\epsilon,B}) \cdot \delta^{-2}\right)$ *and that from* $\mathbb{Q}_{x,\epsilon}$ *is* $\widetilde{\mathcal{O}}\left(H_{\lambda,\epsilon,B} K^2_{\lambda,\epsilon,B} \exp(2K_{\lambda,\epsilon,B}) \cdot \delta^{-2}\right)$*, with* $\widetilde{\mathcal{O}}(\cdot)$ *hiding constants depending on* L_f, θ_0, κ *and linearly depending on* $(\log \frac{\lambda\epsilon}{\delta})^2, (\log \frac{1}{\eta})^2$*.*

Next, we provide the convergence analysis for Algorithm 9.

Theorem 22. *Let $\eta \in (0,1)$. Assume Assumption 5 holds and $0 < \lambda_l \le \lambda^* \le \lambda_u < \infty$. Specify hyper-parameters in Algorithm 9 as*

$$T' = \left\lceil \log_2 \left(\frac{4L_\Psi(\lambda_u - \lambda_l)}{\delta} \right) \right\rceil, \quad \eta' = \frac{\eta}{3 + 2T'}, \quad L_\Psi = \overline{\rho} + \frac{B}{\lambda_l} \left[1 + \exp(K_{\lambda_l, \epsilon, B}) \right].$$

Suppose there exists an oracle $\widehat{\Psi}$ such that for any $\lambda > 0$, it estimates Ψ defined in (5.12) *with accuracy level $\delta/4$ with probability at least $1 - \eta'$, then with probability at least $1 - \eta$, Algorithm 9 finds the optimal multiplier with accuracy level δ (i.e., it finds λ such that $\Psi(\lambda) - \min_{\lambda_l \le \lambda \le \lambda_u} \Psi(\lambda) \le \delta$) by calling the inexact oracle $\widehat{\Psi}$ for $\widetilde{\mathcal{O}}(K_{\lambda_l, \epsilon, B})$ times, where $\widetilde{\mathcal{O}}(\cdot)$ hides constants depending on $\overline{\rho}$ and linearly depending on $\log \frac{\lambda_u - \lambda_l}{\delta}$ and $\log \frac{B}{\lambda_l}$.*

Remark 20 (Selection of λ_u and λ_l)**.** *Algorithm 9 requires the upper and lower bounds λ_u, λ_l on the optimal Lagrange multiplier λ^* as inputs. Under Assumption 5, we have a theoretical upper bound $\lambda_u := \overline{\rho}^{-1} B$ (see the proof in Lemma 32 in Appendix D.7). For the lower bound λ_l, it can be shown that as long as the condition in Theorem 20*(IV) *does not hold for $f_\theta(\cdot)$ for any θ, a valid lower bound $\lambda_l > 0$ exists. Unfortunately, deriving the closed-form expression of λ_l is infeasible. In practice, choosing an excessively small multiplier values may lead to solutions that are too conservative (where the Sinkhorn distance constraint becomes nearly unbinding). To mitigate this, we recommend empirical tuning of λ_l to ensure it remains sufficiently bounded away from zero, thereby avoiding degenerate cases. For examples in Section 5.5.1 and 5.5.2, we set $\lambda_l = 0.01$ and $\lambda_u = 500$.* ♣

Combining Proposition 5 and Theorem 22, the sample complexity from $\widehat{\mathbb{P}}$ for obtaining a δ-optimal solution of (D) with high probability is $\widetilde{\mathcal{O}}\left(H_{\lambda,\epsilon,B} K_{\lambda,\epsilon,B}^2 \exp(2K_{\lambda,\epsilon,B}) \cdot \delta^{-2} \right)$, and sample complexity from $\mathbb{Q}_{x,\epsilon}$ is $\widetilde{\mathcal{O}}\left(H_{\lambda,\epsilon,B} K_{\lambda,\epsilon,B}^3 \exp(2K_{\lambda,\epsilon,B}) \cdot \delta^{-2} \right)$.

Remark 21 (Comparison with Empirical Risk Minimization)**.** *The minimax lower bound of sample complexity from $\widehat{\mathbb{P}}$ for obtaining a δ-optimal solution from the empirical risk*

minimization (ERM) $\inf_{\theta\in\Theta}\ \mathbb{E}_{x\sim\widehat{\mathbb{P}}}[f_\theta(x)]$ *with a convex loss function* $f_\theta(z)$ *(regardless of the smoothness assumption) is* $\mathcal{O}(\delta^{-2})$ *[39]. The sample complexity from* $\widehat{\mathbb{P}}$ *of solving the Sinkhorn DRO model matches with its ERM counterpart, differing only by a (near-)constant factor in terms of error tolerance* δ*. However, we highlight that the constant factor has non-negligible dependence on related parameters* λ_l, ϵ, B*.* ♣

Remark 22 (Comparison with Wasserestein DRO)**.** *Recall from Table 5.1 that Wasserstein DRO is computationally efficient to solve for a restricted family of loss functions (Table 5.1). Specially, Wasserstein DRO with* $\widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^n \delta_{\hat{x}_i}$ *can be formulated as a minimax problem*

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

When $f_\theta(z)$ *is not piecewise concave in* z*, the above problem generally reduces to the convex-non-concave saddle point problem, whose global optimality is difficult to obtain. In comparison, we provided complexity guarantees for solving Sinkhorn DRO model for a broader class of loss functions.*

We also remark that the regularization parameter ϵ *is treated as a fixed intentional design choice when solving Sinkhorn DRO formulation. Our goal is not to use it as a computational approximation of Wasserstein DRO, even though it converges to Wasserstein DRO as* $\epsilon \to 0$*. Otherwise the constant part in our complexity bounds will explode to infinity.* ♣

Finally, we compare our algorithm design and analysis with existing references on CSO below.

Remark 23 (Comparison with [154, 156, 160])**.** *Our algorithm and analysis differ from existing references on CSO [154, 156, 160] in several aspects.*

Recall that we apply BSMD with the RT-MLMC subgradient estimator to solve (5.12)*. While the BSAA approach proposed in [154] is applicable, it is less efficient (see Remark 19). The BSMD method using the SG subgradient estimator from [160] can also be applied, but it results in worse sample complexity from* $\mathbb{Q}_{x,\epsilon}$*.*

Although one can apply Hu et al. [156] to consider BSMD with RT-MLMC or vanilla MLMC (V-MLMC) gradient estimators, their analysis focuses on unconstrained optimization with S_f-Lipschitz smooth loss functions, i.e., functions that are continuously differentiable with Lipschitz continuous gradients: Their analysis requires carefully tuning the parameters of the gradient estimators to balance the trade-off between bias and gradient variance, with stepsize that depends on the smoothness constant S_f (see [156, Theorem C.1]). This greatly limits the applicability of their theoretical guarantees, such as the newsvendor problem and portfolio optimization examples in our numerical study. In contrast, we demonstrate that the RT-MLMC estimator results in the same complexity bounds even for nonsmooth loss functions by appropriately balancing the trade-off between bias and subgradient second-order moment (Lemma 28). Notably, the V-MLMC estimator in [156] no longer has performance guarantees in nonsmooth case.

Another component of our algorithm requires estimating the objective in (5.12)*. Reference [160] used the SG objective estimator that leads to worse sample complexity from $\mathbb{Q}_{x,\epsilon}$; References [156, 160] did not investigate this problem. Our work provided the RT-MLMC estimator to estimate the Sinkhorn robust optimization objective, demonstrating its superior efficiency (Lemma 30).* ♣

5.5 Applications

In this section, we apply our methodology to three applications: the newsvendor problem, mean-risk portfolio optimization, and adversarial classification. We compare our model with three benchmarks: (i) the classical sample average approximation (SAA) model; (ii) the Wasserstein DRO model; and (iii) the KL-divergence DRO model. We choose the transport cost $c(\cdot,\cdot) = \|\cdot - \cdot\|_1$ for 1-Wasserstein or 1-Sinkhorn DRO model, and $c(\cdot,\cdot) = \frac{1}{2}\|\cdot - \cdot\|_2^2$ for 2-Wasserstein or 2-Sinkhorn DRO model. Throughout this section, we take the reference measure ν in the Sinkhorn distance to be the Lebesgue measure. The hyper-parameters are selected using the holdout method following from [226]. All experiments were conducted

on a Mac mini computer with 24GB of memory and M4 Pro GPU with 20 cores running Python 3.9. Further implementation details and experiments are included in Appendices D.1 and D.2, respectively.

5.5.1 Newsvendor Problem

Consider the following distributionally robust newsvendor problem:

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

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

In this experiment, we examine the performance of DRO models for various sample sizes $n \in \{10, 30, 100\}$ and under three different types of data distribution: (i) the exponential distribution with rate parameter 1, (ii) the gamma distribution with shape parameter 2 and scale parameter 1.5, (iii) the equiprobable mixture of two truncated normal distributions $\mathcal{N}(\mu = 1, \sigma = 1, a = 0, b = 10)$ and $\mathcal{N}(\mu = 6, \sigma = 1, a = 0, b = 10)$. We do not report the performance for 1-Wasserstein DRO model in this example, because it is identical to the SAA approach [226, Remark 6.7]. As 2-Wasserstein DRO is computationally intractable for this example, we solve the corresponding formulation by discretizing the support of the distributions.

We measure the out-of-sample performance of a solution θ based on training dataset $\mathcal{D}$ using the percentage of improvement (a.k.a., coefficient of prescriptiveness) in [31]:

$$\text{Prescriptiveness}(\theta) = \max\left(1 - \frac{J(\theta) - J^*}{J(\theta^{\text{SAA}}_{\mathcal{D}}) - J^*}, -1\right) \times 100\%, \tag{5.19}$$

where J^* denotes the true optimal value when the true distribution is known, $\theta^{\text{SAA}}_{\mathcal{D}}$ denotes

the decision from the SAA approach with dataset $\mathcal{D}$, and $J(\theta)$ denotes the expected loss of the solution θ under the true distribution, estimated through an SAA objective value with 10^5 testing samples. This coefficient is always bounded between -100% and 100%, and the higher this coefficient is, the better the solution's out-of-sample performance.

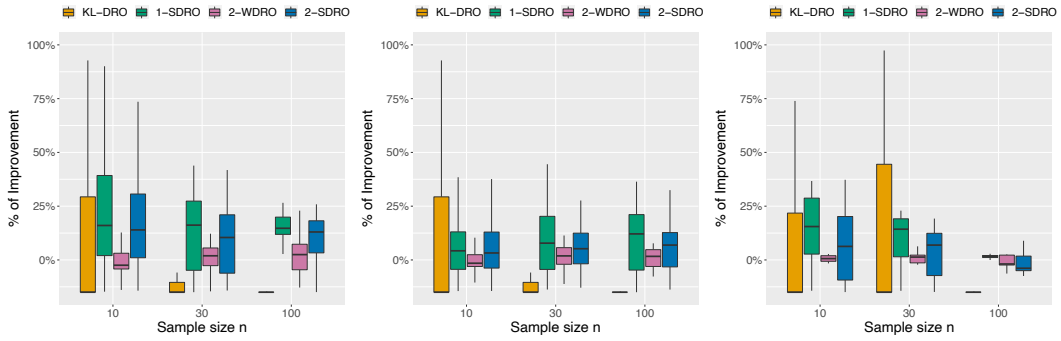

Figure 5.2: Experiment results of the newsvendor problem for different sample sizes and different data distributions in box plots. (a) Exponential distribution; (b) Gamma distribution; (c) Mixture of truncated normal distributions.

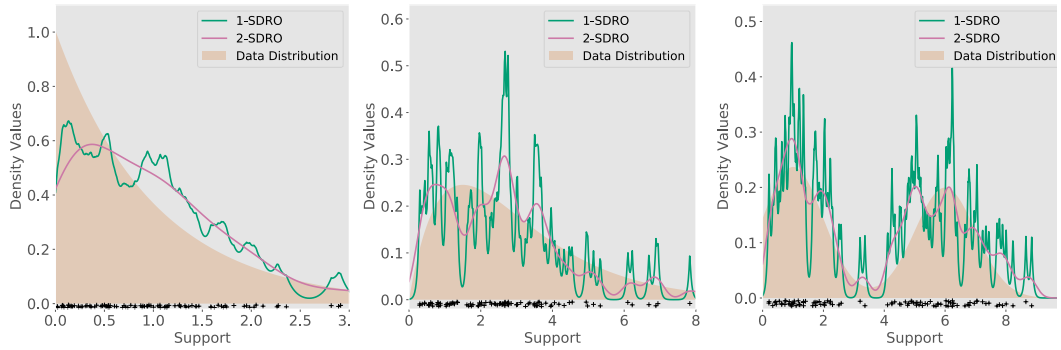

Figure 5.3: Plots for the density of worst-case distributions generated by the 1-SDRO or 2-SDRO model for newsvendor problem with different data distributions. (a) Exponential distribution; (b) Gamma distribution; (c) Mixture of truncated normal distributions.

We report the box-plots of the coefficients of prescriptiveness in Fig. 5.2 using 500 independent trials. We find that either 1-SDRO or 2-SDRO model achieve the best out-of-sample performance over all sample sizes and data distributions listed, as it consistently scores higher than other benchmarks in the box plots. In contrast, the KL-DRO model does not achieve satisfactory performance, and sometimes even underperforms the SAA model. While the 2-WDRO model demonstrates some improvement over the SAA model, the 2-

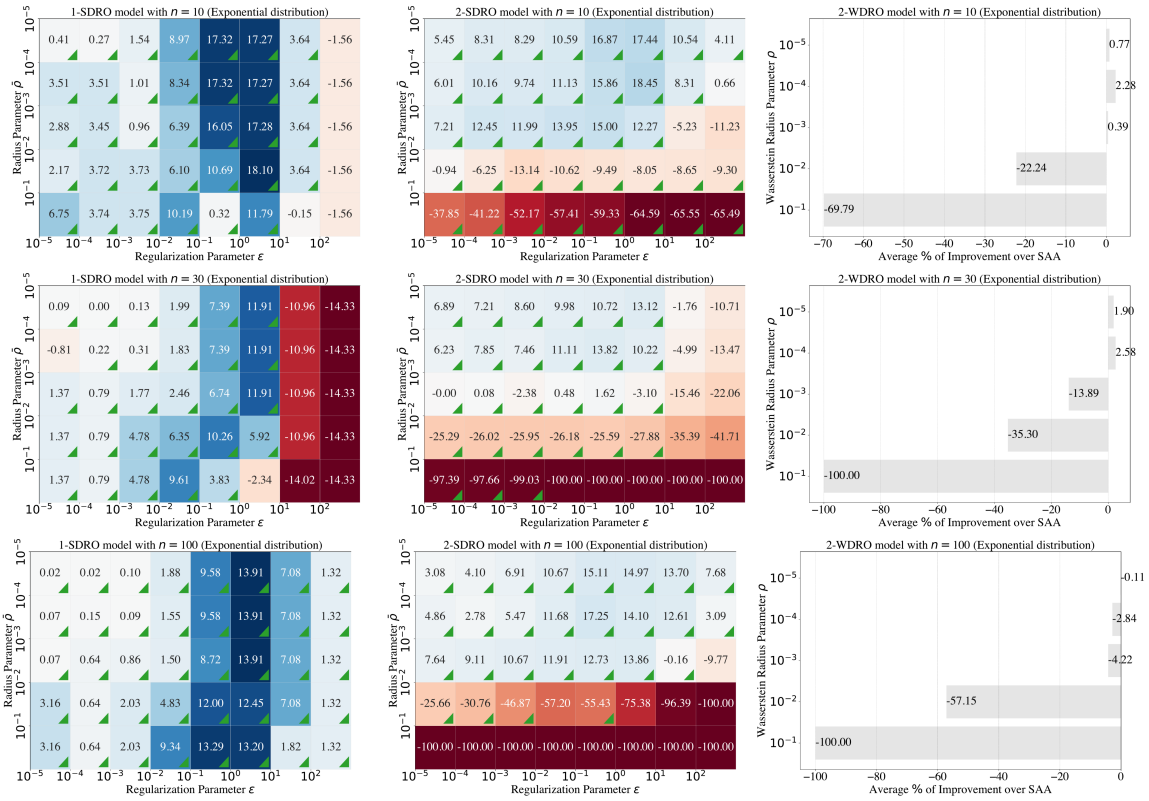

Figure 5.4: Experiment results of the newsvendor problem for exponential data distribution. Subplots from different rows correspond to different training sample sizes $n \in \{10, 30, 100\}$. Subplots from the first and second columns correspond to the heatmap plot of the coefficient of prescriptiveness for 1-SDRO and 2-SDRO models with different radius and regularization parameters, and the subplots from the last column correspond to the histogram plot of the coefficient of prescriptiveness for 2-WDRO model with different radius parameters. Each instance is taken the average of the simulation results over 50 independent trials. For SDRO models, we add a green triangle for each radius-regularization combination that outperforms the corresponding WDRO models with the same radius choice.

SDRO model shows more clear improvement. We plot the density of worst-case distributions for 1-SDRO or 2-SDRO model in Fig. 5.3. When specifying the data distribution as exponential, gamma, or Gaussian mixture, the corresponding worst-case distributions capture the shape of the ground truth distribution reasonably well, which partly explains why the Sinkhorn DRO model achieves superior performance when the data distribution is absolutely continuous. We report the coefficient of prescriptiveness of SDRO and WDRO models with different parameters when the data distribution is exponential in Fig. 5.4 (plots for other distributions are presented in Appendix D.2.4). We observe that there exists a large range of

parameter choices of SDRO models that lead to the superior performance over the WDRO models, since almost each row of the heatmap includes many green triangles. This fact justifies the benefits of adding entropic regularization.

5.5.2 Mean-risk Portfolio Optimization

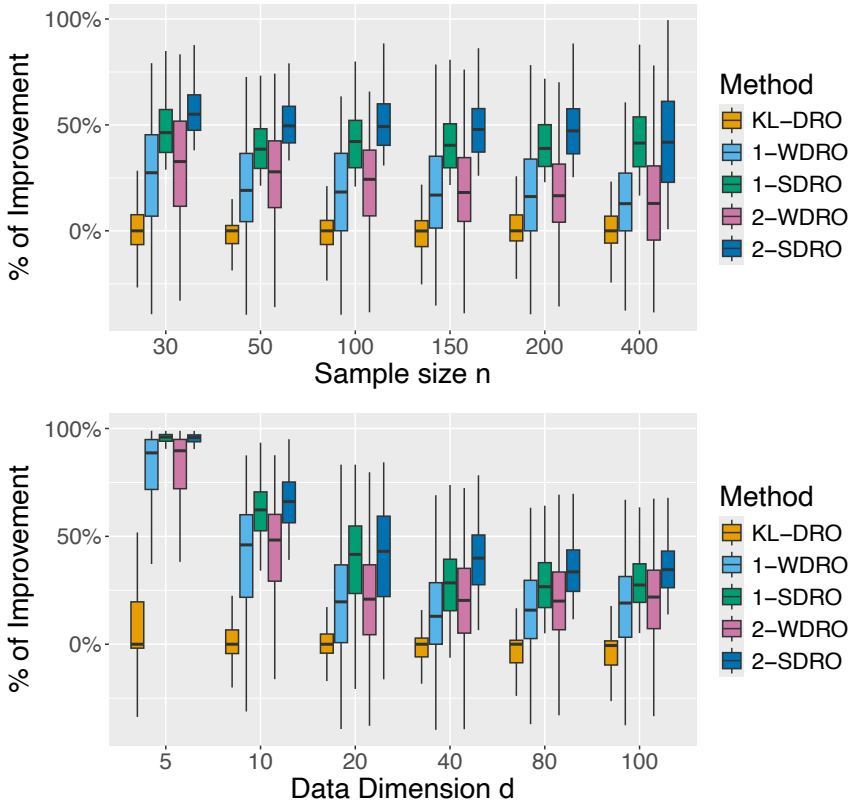

Figure 5.5: Experiment results of the portfolio optimization problem for different sample sizes and dimensions in box plots. (a) fixing data dimension $d = 30$ and varying sample size $n \in \{30, 50, 100, 150, 200, 400\}$; (b) fixing sample size $n = 100$ and varying data dimension $d \in \{5, 10, 20, 40, 80, 100\}$.

Consider the following distributionally robust mean-risk portfolio optimization problem

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

where the random vector $z \in \mathbb{R}^d$ stands for the returns of assets; the decision variable $\theta \in \Theta$

represents the portfolio strategy that invests a certain percentage θ_i of the available capital in the i-th asset; and the term $\mathbb{P}$-CVaR$_\alpha(-\theta^{\mathrm{T}} z)$ quantifies conditional value-at-risk [259], i.e., the average of the $\alpha \times 100\%$ worst portfolio losses under the distribution $\mathbb{P}$. We follow a similar setup as in Mohajerin Esfahani and Kuhn [226]. Specifically, we set $\alpha = 0.2, \varrho = 10$. The underlying true random return can be decomposed into a systematic risk factor $\psi \in \mathbb{R}$ and idiosyncratic risk factors $\epsilon \in \mathbb{R}^d$:

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

where $\psi \sim \mathcal{N}(0, 0.02)$ and $\epsilon_i \sim \mathcal{N}(i \times 0.03, i \times 0.025)$. When solving the Sinkhorn DRO formulation, we take the Bregman divergence D_ω as the KL-divergence when performing BSMD algorithm in Algorithm 7, allowing for efficient implementation [39].

We quantify the performance of a given solution using the same criterion defined in Section 5.5.1 and generate box plots using 500 independent trials. Fig. 5.5a) reports the scenario where the data dimension $d = 30$ is fixed and sample size $n \in \{30, 50, 100, 150, 200, 400\}$, and Fig. 5.5b) reports the scenario where the sample size $n = 100$ is fixed and the number of assets $d \in \{5, 10, 20, 40, 80, 100\}$. We find that the KL-DRO model does not have competitive performance compared to other DRO models, especially as the data dimension d increases. This is because the ambiguity set of KL-DRO model only takes into account those distributions sharing the same support as the nominal distribution, which seems to be restrictive, especially for high-dimensional settings. While 1-WDRO or 2-WDRO model has better out-of-sample performance than the SAA model, the corresponding 1-SDRO or 2-SDRO model has clearer improvements, as it consistently scores higher in the box plots. Finally, we show the coefficient of prescriptiveness of WDRO/SDRO models with different parameters for the instance $(n, d) = (30, 30)$ in Fig. 5.6 (other instances can be found in Appendix D.2.4). Similar as in the newsvendor problem, we observe that there exists a large number of parameter choices of SDRO models that outperform WDRO models, as indicated

by many green triangles in heatmap plots.

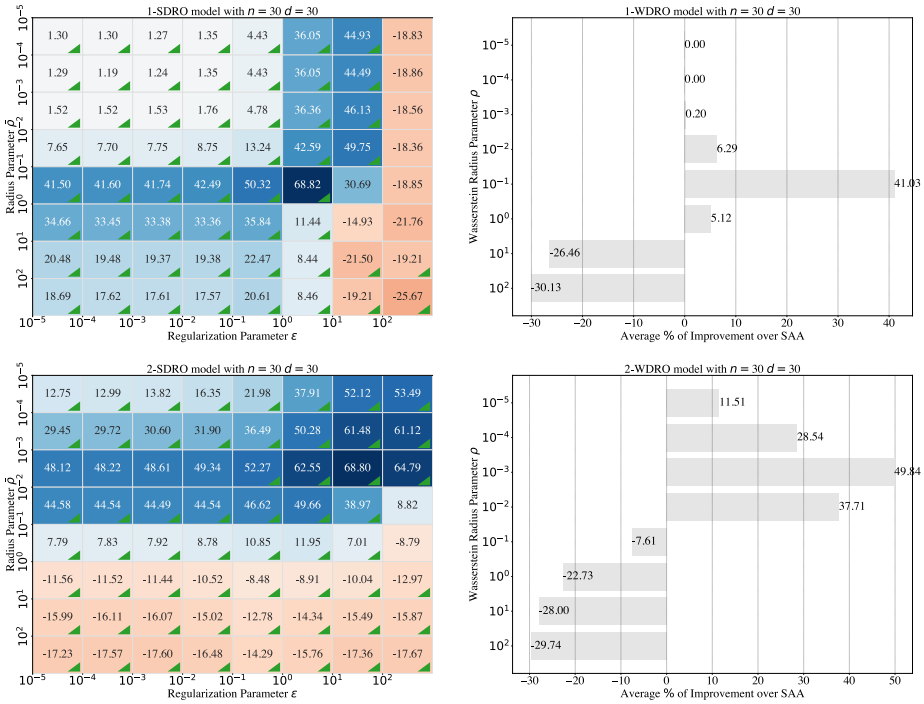

Figure 5.6: Experiment results of the portfolio optimization model with $(n, d) = (30, 30)$ in heatmaps. The four subplots correspond to the heatmap plot of the coefficient of prescriptiveness for 1-SDRO, 1-WDRO, 2-SDRO, and 2-WDRO models with varying parameters. For SDRO models, we add a green triangle for each radius-regularization combination that outperforms the corresponding WDRO models with the same radius choice.

5.5.3 Adversarial Multi-class Logistic Regression

Adversarial machine learning is an emerging topic in artificial intelligence, aiming to develop models that are robust against (potentially adversarial) data perturbations. It has been observed that small perturbations to the data can cause well-trained machine learning models to produce unexpectedly inaccurate predictions [137]. In real applications involving high-stake environments, such as self-driving and automated tumor detection, ensuring model robustness is essential for reliability and safety. Among existing approaches that produce robust machine learning models [146, 212, 240, 241, 242, 260, 286, 296], stands out as a particularly effective method and provides certifiable robustness [286]. In this subsection, we examine the performance of various DRO approaches for multi-class logistic

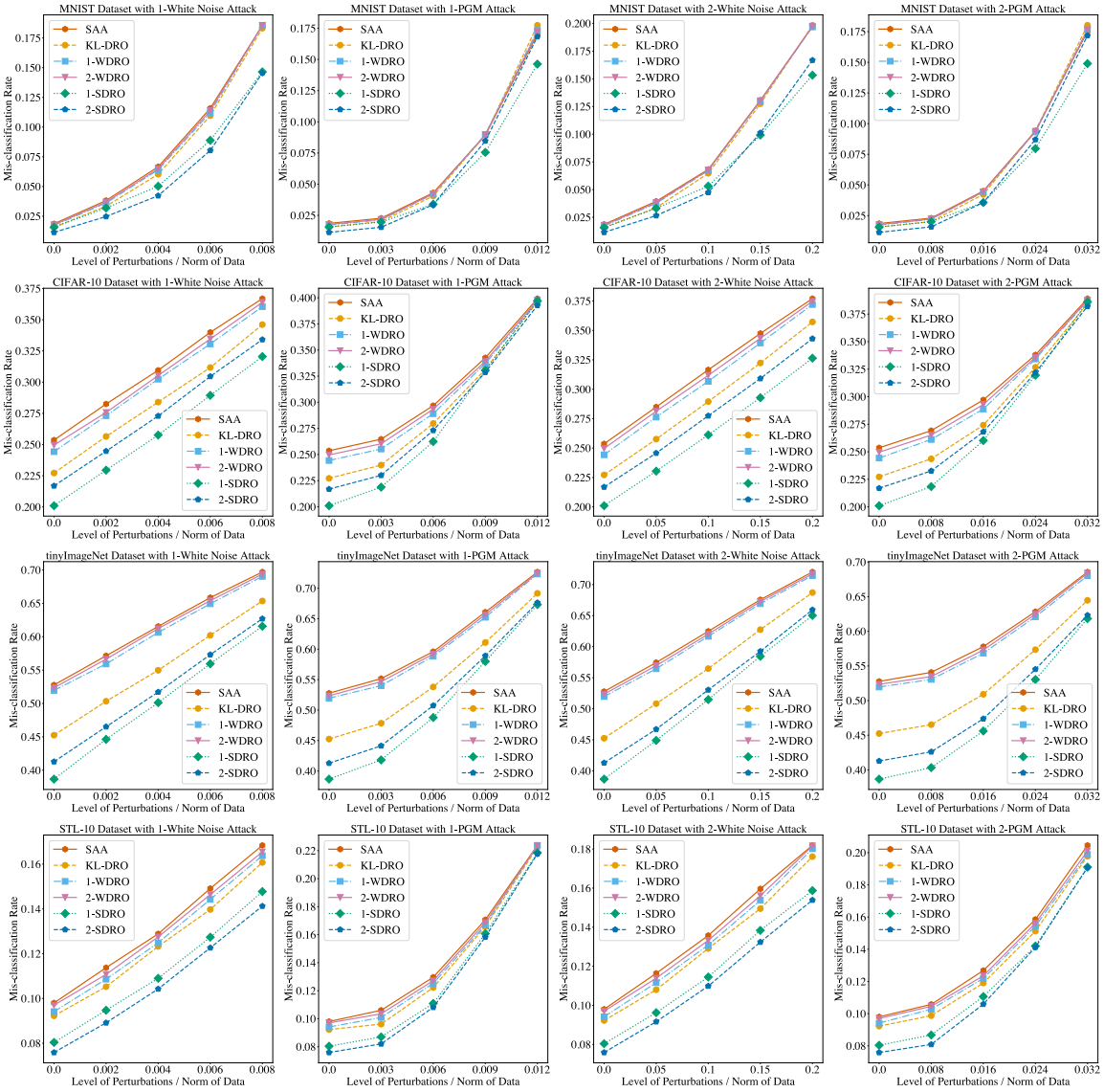

Figure 5.7: Results of adversarial training on various image datasets with different types of adversarial attack. From left to right, the figures correspond to (a) white Laplacian noise attack; (b) ℓ_1-norm PGD attack; (c) white Gaussian noise attack; and (d) ℓ_2-norm PGD attack. From top to bottom, the figures correspond to (a) MNIST dataset; (b) CIFAR-10 dataset; (c) tinyImageNet dataset; and (d) STL-10 dataset.

regression with data perturbations. Given a feature vector $x \in \mathbb{R}^d$ and its label $y \in [C]$, we denote $\boldsymbol{y} \in \{0,1\}^C$ as the corresponding one-hot label vector, and define the negative likelihood loss

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

where $B := [w_1, \ldots, w_K]$ denotes the parameters of the linear classifier. Let $\widehat{\mathbb{P}}$ be the

empirical distribution from training samples. Since the testing samples may have slightly different data distributions than the training samples, the DRO model aims to solve the following optimization problem to mitigate the impact of data perturbations:

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

It is assumed that the data perturbation only happens for the feature vector x but not the label $\boldsymbol{y}$.

We conduct experiments on four large-scale datasets: MNIST [188], CIFAR-10 [179], tinyImageNet [294], and STL-10 [82]. We pre-process these datasets using the ResNet-18 network [145] pre-trained on the ImageNet dataset to extract linear features. Since this network has learned a rich set of hierarchical features from the large and diverse ImageNet dataset, it typically extracts useful features for other image datasets. We then add different types of perturbations to the testing datasets, such as *white Laplace noise*, *white Gaussian noise*, and ℓ_p*-norm adversarial projected gradient descent (PGD) attacks [212] with* $p \in \{1, 2\}$. The level of perturbation is normalized by the averaged ℓ_2 norm of the feature vectors from testing dataset. See the detailed procedure for generating data perturbations and statistics on pre-processed datasets in Appendix D.1.3. We use the misclassification rate on testing dataset to measure the performance of the obtained classifers.

For baseline DRO models, we solve their Lagrangian relaxation, which adds the penalty of the statistical distance into the objective to ensure efficient implementation. To make fair comparisons, we tune the penalty parameter for each method such that the 2-Wasserstein distance between the nominal distribution and its perturbed one is controlled within $\varrho := 0.05 \cdot \mathbb{E}[\|\boldsymbol{x}\|_2^2]$ (expectation taken with respect to the training dataset). For SDRO methods, we fix $\epsilon = 0.1$ unless otherwise stated. It is noteworthy that solving the Lagrangian relaxation of 2-WDRO has global convergence guarantees only when the penalty parameter is sufficiently large [286], which is not the case for this example. In general, solving 1-

or 2-WDRO model reduces to solving a convex-non-concave minimax game, and we try gradient descent ascent [286, Algorithm 1] as a heuristic.

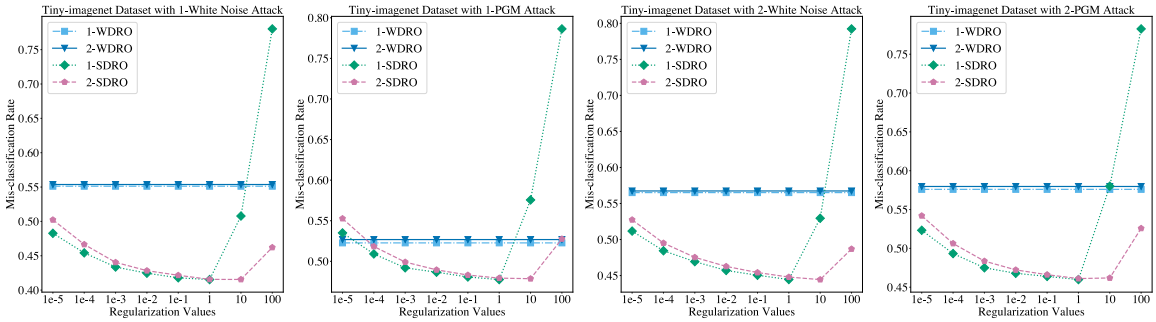

Figure 5.8: Experiment results of the adversarial classification problem with tinyImagenet dataset. The subplots from left to right correspond to the misclassification errors of SDRO and WDRO models with different types of adversarial attack. For SDRO models we vary the regularization parameter ϵ.

Fig. 5.7 presents the classification results for different types of adversarial attacks with varying levels of perturbations on the datasets. We observe that as the level of perturbations on the testing samples increases, all methods tend to perform worse. However, both the 1-SDRO and 2-SDRO models show a slower trend of increasing error rates than other benchmarks across all types of adversarial attacks and all datasets. This suggests that SDRO models can be a suitable choice for adversarial robust training. Fig.5.8 presents the misclassification rates of SDRO models under different types of adversarial attacks on the TinyImageNet dataset (additional results are provided in Appendix D.2.4). For baseline comparison, we also report the performance of WDRO models. In all subplots, the perturbation levels (normalized by the data norm) are fixed at 0.004, 0.006, 0.01, and 0.016 for the four adversarial attacks, while the penalty parameter is set to $\lambda = 10$ for both WDRO and SDRO models. The figure demonstrates that SDRO outperforms WDRO across a wide range of regularization parameters.

5.6 Concluding Remarks

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

In the meantime, several topics worth investigating are left for future work. For example, it is desirable to study the statistical performance guarantees under suitable choices of hyper-parameters. It is also of research interest to develop and analyze optimization algorithms with less restrictive assumptions and sharper complexity bounds. Exploring and discovering the benefits of Sinkhorn DRO in other applications may also be of future interest.

CHAPTER 6

REGULARIZATION FOR ADVERSARIAL ROBUST LEARNING

Despite the growing prevalence of artificial neural networks in real-world applications, their vulnerability to adversarial attacks remains a significant concern, which motivates us to investigate the robustness of machine learning models. While various heuristics aim to optimize the distributionally robust risk using the ∞-Wasserstein metric, such a notion of robustness frequently encounters computation intractability. To tackle the computational challenge, we develop a novel approach to adversarial training that integrates ϕ-divergence regularization into the distributionally robust risk function. This regularization brings a notable improvement in computation compared with the original formulation. We develop stochastic gradient methods with biased oracles to solve this problem efficiently, achieving the near-optimal sample complexity. Moreover, we establish its regularization effects and demonstrate it is asymptotic equivalence to a regularized empirical risk minimization framework, by considering various scaling regimes of the regularization parameter and robustness level. These regimes yield gradient norm regularization, variance regularization, or a smoothed gradient norm regularization that interpolates between these extremes. We numerically validate our proposed method in supervised learning, reinforcement learning, and contextual learning and showcase its state-of-the-art performance against various adversarial attacks. This work is mainly summarized in [310].

6.1 Introduction

Machine learning models are highly vulnerable to potential *adversarial attack* on their input data, which intends to cause wrong outputs. Even if the adversarial input is slightly different from the clean input drawn from the data distribution, these machine learning models can make a wrong decision. Goodfellow et al. [137] provided an example that, after adding a

tiny adversarial noise to an image, a well-trained classification model may make a wrong prediction, even when such data perturbations are imperceptible to visual eyes.

Given that modern machine learning models have been applied in many safety-critical tasks, such as autonomous driving, medical diagnosis, security systems, *etc*, improving the resilience of these models against adversarial attacks in such contexts is of great importance. Neglecting to do so could be risky or unethical and may result in severe consequences. For example, if we use machine learning models in self-driving cars, adversarial examples could allow an attacker to cause the car to take unwanted actions.

Adversarial training is a process of training machine learning model to make it more robust to potential adversarial attacks. To be precise, it aims to optimize the following robust optimization (RO) formulation, called *adversarial risk minimization*:

$$\min_{\theta\in\Theta} \Big\{\mathbb{E}_{z\sim\widehat{P}}\big[R_\rho(\theta;z)\big]\Big\}, \quad \text{where } R_\rho(\theta;z) \triangleq \sup_{z'\in\mathbb{B}_\rho(z)} f_\theta(z'). \tag{6.1}$$

Here $\widehat{P}$ represents the observed distribution on data, θ represents the machine learning model, $f_\theta(z)$ is a loss function, and the uncertainty set is defined as $\mathbb{B}_\rho(x) \triangleq \{z \in \mathcal{Z} : \|z-x\| \le \rho\}$ for some norm function $\|\cdot\|$ and some radius $\rho > 0$. In other words, this formulation seeks to train a machine learning model based on adversarial perturbations of data, where the adversarial perturbations can be found by considering all possible inputs around the data with radius ρ and picking the one that yields the worst-case loss. Unfortunately, problem (6.1) is typically intractable to solve because the inner supremum objective function is in general nonconcave in z. As pointed out by [287], solving the inner supremum problem in (6.1) with deep neural network loss functions is NP-hard. Several heuristic algorithms [63, 137, 185, 212, 241, 296] have been proposed to approximately find the optimal solution of (6.1), but they lack of global convergence guarantees and it remains an open question whether they can accurately and efficiently find the adversarial perturbations of data.

In this paper, we propose a new approach for adversarial risk minimization by adding a ϕ-divergence regularization. Here is a brief overview. By [123, Lemma EC.2], Problem (6.1) can be viewed as the dual reformulation of the following DRO problem:

$$\min_{\theta\in\Theta}\left\{\sup_{\mathbb{P}}\left\{\mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)]:\ \mathcal{W}_\infty(\mathbb{P},\widehat{P})\le\rho\right\}\right\}, \tag{∞-WDRO}$$

where $\mathcal{W}_\infty(\cdot,\cdot)$ is the ∞-Wasserstein metric defined as

$$\mathcal{W}_\infty(\mathbb{P},\mathbb{Q})=\inf_{\gamma}\left\{\text{ess.sup}_\gamma\,\|\zeta_1-\zeta_2\|:\ \begin{array}{l}\gamma\text{ is a joint distribution of }\zeta_1\text{ and }\zeta_2\\ \text{with marginals }\mathbb{P}\text{ and }\mathbb{Q}\text{, respectively}\end{array}\right\}.$$

Therefore, it is convenient to introduce the optimal transport mapping γ to re-write problem (∞-WDRO) as

$$\min_{\theta\in\Theta}\left\{\sup_{\mathbb{P},\gamma}\left\{\mathbb{E}_{\mathbb{P}}[f_\theta(z)]:\ \begin{array}{l}\text{Proj}_{1\#}\gamma=\widehat{P},\text{Proj}_{2\#}\gamma=\mathbb{P}\\ \text{ess.sup}_\gamma\|\zeta_1-\zeta_2\|\le\rho\end{array}\right\}\right\}. \tag{6.2}$$

As long as the loss $f_\theta(z)$ is nonconcave in z, such as neural networks and other complex machine learning models, problem (6.2) is intractable for arbitrary radius $\rho>0$. Instead, we add *ϕ-divergence regularization* to the objective in (6.2), and focus on solving the following formulation:

$$\min_{\theta\in\Theta}\left\{\sup_{\mathbb{P},\gamma}\left\{\mathbb{E}_{z\sim\mathbb{P}}\left[f_\theta(z)\right]-\eta\mathbb{D}_\phi(\gamma,\gamma_0):\ \begin{array}{l}\text{Proj}_{1\#}\gamma=\widehat{P},\text{Proj}_{2\#}\gamma=\mathbb{P}\\ \text{ess.sup}_\gamma\|\zeta_1-\zeta_2\|\le\rho\end{array}\right\}\right\}, \tag{Reg-∞-WDRO}$$

where γ_0 is the reference measure satisfying $\mathrm{d}\gamma_0(x,z)=\mathrm{d}\widehat{P}(x)\,\mathrm{d}\nu_x(z)$, with ν_x being the uniform probability measure on $\mathbb{B}_\rho(x)$, and $\mathbb{D}_\phi(\gamma,\gamma_0)$ is the ϕ-divergence [89] between γ and γ_0. In the following, we summarize several notable features of our proposed formulation.

Strong Dual Reformulation.

By the duality result in Theorem 23, (Reg-∞-WDRO) admits the strong dual reformulation:

$$\min_{\theta\in\Theta} \quad \mathbb{E}_{z\sim\widehat{P}}[\psi_\eta(z)], \tag{6.3a}$$

$$\text{where} \quad \psi_\eta(z) = \inf_{\mu\in\mathbb{R}} \left\{ \mu + \mathbb{E}_{z'\sim\nu_x} \left[(\eta\phi)^* \big(f_\theta(z') - \mu \big) \right] \right\}. \tag{6.3b}$$

Compared with the original formulation (6.1), we replace the worst-case loss $R_\rho(\theta; z)$ defined in (6.1) by $\phi_\eta(z)$, which is a variant of *optimized certainty equivalent* (OCE) risk measure studied in [26]. Subsequently, it can be shown that $\phi_\eta(x)$ is a smooth approximation of the optimal value $R_\rho(\theta; x)$.

Worst-case Distribution Characterization.

We characterize the worst-case distribution for problem (Reg-∞-WDRO) in Remark 6.2.1 and display its simplified expressions in Examples 5-9. In contrast to the conventional formulation (∞-WDRO) that *deterministically* transports each data from $\widehat{P}$ to its extreme perturbation, the worst-case distribution of our formulation transports each data x towards the entire domain set $\mathbb{B}_\rho(x)$ through specific absolutely continuous distributions. This observation indicates that our formulation (Reg-∞-WDRO) is well-suited for adversarial defense where the data distribution after adversarial attack manifests as absolutely continuous, such as through the addition of white noise to the data.

Efficient Stochastic Optimization Algorithm.

We adopt the idea of stochastic approximation to solve our reformulation (6.3) by iteratively obtaining a stochastic gradient estimator and next performing projected gradient update. To tackle the difficulty that one cannot obtain the unbiased gradient estimator, we introduce and analyze stochastic gradient methods with biased oracles inspired from [155]. Our proposed

algorithm achieves $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity for finding ϵ-optimal solution for convex $f_\theta(z)$ and general choices of ϕ-divergence, and $\widetilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity for finding ϵ-stationary point for nonconvex $f_\theta(z)$ and KL-divergence. These sample complexity results are near-optimal up to a near-constant factor.

Regularization Effects.

We develop regularization effects for problem (Reg-∞-WDRO) in Section 6.4. Specifically, we show that it is asymptotically equivalent to regularized ERM formulations under three different scalings of the regularization value ϵ and radius ρ: Let β be an uniform distribution supported on the unit norm ball $\mathbb{B}_1(0)$, and $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. When $\rho, \eta \to 0$, it holds that (Reg-∞-WDRO) $= \min_{\theta\in\Theta} \mathbb{E}_{z\sim\widehat{P}}[f_\theta(z)] + \mathcal{E}(f_\theta; \rho, \eta)$, where

$$\mathcal{E}(f;\rho,\eta) \simeq \begin{cases} \rho \cdot \mathbb{E}_{z\sim\widehat{P}}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu + \frac{1}{C}\mathbb{E}_{b\sim\beta}\left[\phi^*\Big(C\cdot(\nabla f(z)^{\mathrm{T}}b - \mu)\Big)\right]\right\}\right], & \text{if } \frac{\rho}{\eta} \to C, \\ \rho \cdot \mathbb{E}_{z\sim\widehat{P}}[\|\nabla f(z)\|_*], & \text{if } \frac{\rho}{\eta} \to \infty, \\ \frac{\rho^2}{2\eta\cdot\phi''(1)} \cdot \mathbb{E}_{z\sim\widehat{P}}\Big[\mathbb{V}\mathrm{ar}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}}b]\Big], & \text{if } \frac{\rho}{\eta} \to 0. \end{cases}$$

In other words, when $\rho/\eta \to \infty$, it corresponds to the gradient norm regularized ERM formulation; when $\rho/\eta \to 0$, it corresponds to a special gradient variance regularized ERM formulation; when $\rho/\eta \to C$, it corresponds to a regularized formulation that interpolates between these extreme cases.

Generalization Error Analysis.

We investigate the generalization properties of our proposed adversarial training framework. In particular, the optimal value in (Reg-∞-WDRO) is the confidence upper bound of its population version up to a negligible residual error. Next, we present the specific generalization error bound for linear and neural network function classes.

Numerical Applications.

Finally, we provide numerical experiments in Section 6.6 on supervised learning, reinforcement learning, and contextual learning. Numerical results demonstrate the state-of-the-art performance attained by our regularized adversarial learning framework against various adversarial attacks.

Related Work

Adversarial Learning.

Ever since the seminal work [137] illustrated the vulnerability of neural networks to adversarial perturbations, the research on adversarial attack and defense has progressively gained much attention in the literature. The NP-hardness of solving the adversarial training problem (6.1) with ReLU neural network structure has been proved in Sinha et al. [287], indicating that one should resort to efficient approximation algorithms with satisfactory solution quality. Numerous approaches for adversarial defense have been put forth [63, 137, 185, 212, 241, 296], aiming to develop heuristic algorithms to optimize the formulation (6.1) relying on the local linearization (i.e., first-order Taylor expansion) of the loss f_θ. Unfortunately, the Taylor expansion may not guarantee an accurate estimate of the original objective in (6.1), especially when the robustness level ρ is moderate or large. Henceforth, these algorithms often fail to find the worst-case perturbations of the adversarial training.

Distributionally Robust Optimization.

Our study is substantially related to the DRO framework. In literature, the modeling of distributional uncertainty sets (also called ambiguity sets) for DRO can be categorized into two approaches. The first considers finite-dimensional parameterizations of the ambiguity sets by taking into account the support, shape, and moment information [32, 74, 94,

135, 253, 265, 298, 318, 338]. The second approach, which has received great attention recently, constructs ambiguity sets using non-parametric statistical discrepancy, including f-divergence [22, 24, 108, 161, 315], Wasserstein distance and its entropic-regularized variant [13, 44, 73, 124, 226, 249, 305, 309, 319, 324, 336], and maximum mean discrepancy [290, 337].

There are many results on the computational traceability of DRO. Sinha et al. [287] showed that replacing ∞-Wasserstein distance with 2-Wasserstein distance in (∞-WDRO) yields more tractable formulations. Unfortunately, their proposed algorithm necessitates a sufficiently small robustness level such that the involved subproblem becomes strongly convex, which is not well-suited for adversarial training in scenarios with large perturbations. Wang et al. [305] added entropic regularization regarding the p-WDRO formulation to develop more efficient algorithms. We highlight that their result cannot be applied to the entropic regularization for ∞-WDRO setup because the associated transport cost function is not finite-valued.

Stochastic Gradient Methods with Biased Gradient Oracles.

Stochastic biased gradient methods have received great attention in both theory and applications. References [4, 68, 151, 153] construct gradient estimators with small biases at each iteration and analyze the iteration complexity of their proposed algorithms, ignoring the cost of querying biased gradient oracles. Hu et al. [155, 157] proposed efficient gradient estimators using multi-level Monte-Carlo (MLMC) simulation and provided a comprehensive analysis of the total complexity of their algorithms by considering both iteration and per-iteration costs. This kind of algorithm is especially useful when constructing unbiased estimators can be prohibitively expansive or even infeasible for many emerging machine learning and data science applications, such as ϕ-divergence/Sinkhorn DRO [191, 305, 335], meta learning [158, 167], and contextual learning [104]. We show that our formulation (Reg-∞-WDRO) can also be solved using this type of approach.

Notations. Denote by $\mathrm{Proj}_{1\#}\gamma, \mathrm{Proj}_{2\#}\gamma$ the first and the second marginal distributions of γ, respectively. For a measurable set $\mathcal{Z}$, denote by $\mathcal{P}(\mathcal{Z})$ the set of probability measures on $\mathcal{Z}$. Denote by $\operatorname{supp}\mathbb{P}$ the support of probability distribution $\mathbb{P}$. Given a measure μ and a measurable variable $f:\ \mathcal{Z}\to\mathbb{R}$, we write $\mathbb{E}_{z\sim\mu}[f]$ for $\int f(z)\,\mathrm{d}\mu(z)$. Given a subset E in Euclidean space, let $\mathrm{vol}(E)$ denote its volume. Let $\theta^*\in\arg\min_{\theta\in\Theta}\ F(\theta)$. We say a given random vector θ is a δ-optimal solution if $\mathbb{E}[F(\theta)-F(\theta^*)]\le\delta$. In addition, we say θ is a δ-stationary point if for some step size $\gamma>0$, it holds that $\mathbb{E}\left\|\frac{1}{\gamma}\left[\theta-\mathrm{Proj}_\Theta\big(\theta-\delta\nabla F(\theta)\big)\right]\right\|_2^2\le\delta^2$. For a given probability measure μ in $\mathbb{R}^d$, denote by $f_{\#}\mu$ the pushforward measure of μ by $f:\ \mathbb{R}^d\to\mathbb{R}$.

6.2 Phi-Divergence Regularized Adversarial Robust Training

In this section, we discuss the regularized formulation of the adversarial robust training problem (6.2). Define the reference measure ν_z as the uniform probability measure supported on $\mathbb{B}_\rho(z)\subseteq\mathcal{Z}$, i.e.,

$$\frac{\mathrm{d}\nu_z(\omega)}{\mathrm{d}\omega}=\frac{\mathbf{1}\{\omega\in\mathbb{B}_\rho(z)\}}{\mathrm{vol}(\mathbb{B}_\rho(z))}\triangleq\mathsf{V}_\rho^{-1}\mathbf{1}\{\omega\in\mathbb{B}_\rho(z)\},\tag{6.4}$$

where we denote $\mathsf{V}_\rho=\mathrm{vol}(\mathbb{B}_\rho(z))$, since the volume of $\mathbb{B}_\rho(z)$ is independent of the choice of z. Next, we take the reference measure γ_0, a transport mapping from $\mathcal{Z}$ to $\mathcal{Z}$, as

$$\mathrm{d}\gamma_0(z,z')=\ \mathrm{d}\widehat{P}(z)\,\mathrm{d}\nu_z(z'),\quad\ \forall z,z'\in\mathcal{Z}.$$

Such a reference measure transports the probability mass of $\widehat{P}$ at z to its norm ball $\mathbb{B}_\rho(z)$ uniformly. With such a choice of γ_0, each probability mass of $\widehat{P}$ is allowed to move around its neighborhood (the norm ball of radius ρ) according to certain continuous probability density values, which takes account into a flexible type of adversarial attack. With these notations, we add the following ϕ-divergence regularization on the formulation (6.2). Notably, it ensures the worst-case distribution is absolutely continuous.

Definition 12 (ϕ-divergence Regularization)**.** *Let $\phi:\ \mathbb{R}\to\mathbb{R}_+\cup\{\infty\}$ be a convex lower semi-continuous function such that $\phi(1)=0,\phi(x)=\infty$ if $x<0$. Given an optimal transport mapping $\gamma\in\mathcal{P}(\mathcal{Z}^2)$, define the ϕ-divergence regularization*

$$\mathbb{D}_\phi(\gamma,\gamma_0)=\mathbb{E}_{(z,z')\sim\gamma_0}\left[\phi\left(\frac{\mathrm{d}\gamma(z,z')}{\mathrm{d}\gamma_0(z,z')}\right)\right]. \qquad \clubsuit$$

For simplicity, we focus solely on the inner maximization and omit the dependence of the parameter θ on the loss $f_\theta(z)$. Now, the regularized formulation of (6.2) becomes

$$\sup_{\mathbb{P},\gamma}\left\{\mathbb{E}_{z\sim\mathbb{P}}\left[f(z)\right]-\eta\mathbb{D}_\phi(\gamma,\gamma_0):\quad\begin{array}{l}\mathrm{Proj}_{1\#}\gamma=\widehat{P},\mathrm{Proj}_{2\#}\gamma=\mathbb{P}\\ \text{ess.sup}_\gamma\|\zeta_1-\zeta_2\|\le\rho\end{array}\right\}. \quad \text{(Primal-}\phi\text{-Reg)}$$

By convention, we say for an optimal solution $(\mathbb{P}_*,\gamma^*)$ to (Primal-ϕ-Reg), if exists, the distribution $\mathbb{P}_*$ is its *worst-case* distribution, and γ^* is its *worst-case* transport mapping. Define the dual formulation of (Primal-ϕ-Reg) as

$$\mathbb{E}_{z\sim\widehat{P}}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\mathbb{E}_{z'\sim\nu_z}\left[(\eta\phi)^*\big(f(z')-\mu\big)\right]\right\}\right]. \qquad \text{(Dual-}\phi\text{-Reg)}$$

The following summarizes the main result in this section, which shows the strong duality result, and reveals how to compute the worst-case distribution of (Primal-ϕ-Reg) from its dual. The proof of Theorem 23 is provided in Appendix E.2.

Theorem 23 (Strong Duality)**.** *Assume that $\mathcal{Z}$ is a measurable space, $f:\ \mathcal{Z}\to\mathbb{R}\cup\{\infty\}$ is a measurable function, and for every joint distribution $\gamma\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z})$ with $\mathrm{Proj}_{1\#}\gamma=\widehat{P}$, it has a regular conditional distribution γ_z given the value of the first marginal equals z. Then for any $\eta>0$, it holds that*

(I) (Primal-ϕ-Reg) = (Dual-ϕ-Reg)*;*

(II) *Additionally assume that for $\widehat{P}$-almost surely every z, there exists a primal-dual pair*

(μ_z^*, ζ_z^*) *such that*

$$\zeta_z^* \in \mathcal{Z}_+^*, \quad \mathbb{E}_{\nu_z}[\zeta_z^*] = 1, \quad \zeta_z^*(\omega) = (\eta\phi)^{*\prime}\big[f(\omega) - \mu_z^*\big], \tag{6.5}$$

then there exists a worst-case distribution $\mathbb{P}_*$ *having the density*

$$\frac{\mathrm{d}\mathbb{P}_*(\omega)}{\mathrm{d}\omega} = \mathsf{V}_\rho^{-1} \cdot \mathbb{E}_{z\sim\widehat{P}}\Big[\mathbf{1}\{\omega \in \mathbb{B}_\rho(z)\} \cdot \zeta_z^*(\omega)\Big].$$

Theorem 23 requires γ having a regular conditional distribution γ_z given the value of its first marginal equals z. It simply means that for any given $z \in \operatorname{supp} \widehat{P}$, γ_z is a well-defined transition probability kernel, which always holds for Polish probability space. We refer to [173, Chapter 5] for a detailed discussion of the regular conditional distribution.

6.2.1 Discussions

In the following examples, we show that for some common choices of the function ϕ, Condition (6.5) can be further simplified such that one can obtain more analytical expressions of the worst-case distribution $\mathbb{P}_*$.

Example 5 (Indicator Regularization)**.** *For* $\alpha \in (0, 1]$*, consider the indicator function* ϕ *such that* $\phi(x) = 0$ *for* $x \in [0, \alpha^{-1}]$ *and otherwise* $\phi(x) = \infty$*. Let* μ_z^* *be the left-side* $(1-\alpha)$*-quantile of* $f_\#\nu_z$*, which is also called the value-at-risk and denoted as* $\mathrm{V@R}_{\alpha,\nu_z}(f)$*, and define* $\zeta_z^*(\omega) = \alpha^{-1} \cdot \mathbf{1}\{f(\omega) \geq \mu_z^*\}$*. One can verify that* (μ_z^*, ζ_z^*) *is a primal-dual optimal solution to* (6.5)*, and therefore the worst-case distribution* $\mathbb{P}_*$ *has the density*

$$\frac{\mathrm{d}\mathbb{P}_*(\omega)}{\mathrm{d}\omega} = (\alpha\mathsf{V}_\rho)^{-1} \cdot \mathbb{E}_{z\sim\widehat{P}}\Big[\mathbf{1}\Big\{\big(\omega \in \mathbb{B}_\rho(z)\big) \bigwedge \big(f(\omega) \geq \mathrm{V@R}_{\alpha,\nu_z}(f)\big)\Big\}\Big].$$

Define the average risk-at-risk (AVaR) functional $\mathrm{AV@R}_{\alpha,\mathbb{P}}(f) = \inf_\mu \big\{\mu + \alpha^{-1}\mathbb{E}_{z\sim\mathbb{P}}[f(z) - \mu]_+\big\}$*, then*

$$\text{(Dual-}\phi\text{-Reg)} = \mathbb{E}_{z\sim\widehat{P}}\Big[\mathrm{AV@R}_{\alpha,\nu_z}(f)\Big]. \qquad \clubsuit$$

Example 6 (Entropic Regularization)**.** *Consider* $\phi(x) = x\log x - x + 1, x \geq 0$*. In this case, it can be verified that the primal-dual pair to Condition* (6.5) *is unique and has closed-form expression:*

$$\mu_z^* = \eta \log \mathbb{E}_{\omega\sim\nu_z}\Big[\exp\Big(\frac{f(\omega)}{\eta}\Big)\Big], \quad \zeta_z^*(\omega) = \alpha_z \cdot \exp\Big(\frac{f(\omega)}{\eta}\Big).$$

where $\alpha_z := \big(\mathbb{E}_{\omega\sim\nu_z}[e^{f(\omega)/\eta}]\big)^{-1}$ *is a normalizing constant. Consequently, the worst-case distribution* $\mathbb{P}_*$ *satisfies*

$$\frac{\mathrm{d}\mathbb{P}_*(\omega)}{\mathrm{d}\omega} = \mathsf{V}_\rho^{-1} \cdot \mathbb{E}_{z\sim\widehat{P}}\Big[\alpha_z \cdot \exp\Big(\frac{f(\omega)}{\eta}\Big) \cdot \boldsymbol{1}\{\omega \in \mathbb{B}_\rho(z)\}\Big],$$

and Problem (Dual-ϕ-Reg) *simplifies into an expectation of logarithm of another conditional expectation, which corresponds to the objective of conditional stochastic optimization (CSO) [154, 159]:*

$$\text{(Dual-}\phi\text{-Reg)} = \mathbb{E}_{z\sim\widehat{P}}\left[\psi_{\mathit{Entr}}(z;\eta)\right], \quad \textit{where} \quad \psi_{\mathit{Entr}}(z;\eta) = \eta\log\mathbb{E}_{z'\sim\nu_z}\Big[\exp\Big(\frac{f(z')}{\eta}\Big)\Big].$$

Compared with the original formulation (6.1)*, the entropic regularization framework replaces the worst-case loss* $\sup_{z'\in\mathbb{B}_\rho(z)} f(z')$ *by* $\psi_{\mathit{Entr}}(z;\eta)$*. Based on the well-known Laplace's method (also called the log-sum-exp approximation) [61], this framework provides a smooth approximation of the optimal value in* (6.1)*.* ♣

Example 7 (Quadratic Regularization)**.** *Consider* $\phi(x) = \frac{1}{2}(x^2 - 1), x \geq 0$*. By Condition* (6.5)*, one can verify that* μ_z^* *is a solution to the scalar equation* $\mathbb{E}_{\omega\sim\nu_z}[f(\omega) - \mu_z]_+ = \eta$ *and* $\zeta_z^*(\omega) = \eta^{-1}\big(f(\omega) - \mu_z^*\big)_+$*. Hence, the worst-case distribution* $\mathbb{P}_*$ *has the density*

$$\frac{\mathrm{d}\mathbb{P}_*(\omega)}{\mathrm{d}\omega} = (\eta\mathsf{V}_\rho)^{-1} \cdot \mathbb{E}_{z\sim\widehat{P}}\Big[\boldsymbol{1}\big\{\omega \in \mathbb{B}_\rho(z)\big\} \cdot \big(f(\omega) - \mu_z^*\big)_+\Big].$$

Additionally,

$$\text{(Dual-}\phi\text{-Reg)} = \mathbb{E}_{z\sim\widehat{P}}\left[\inf_{\mu\in\mathbb{R}}\left\{\frac{1}{2\eta}\mathbb{E}_{z'\sim\nu_x}[f(z')-\mu]_+^2+\frac{\eta}{2}+\mu\right\}\right]. \qquad \clubsuit$$

Compared to entropic regularization, this method requires solving a one-dimensional minimization problem before determining the worst-case distribution density or evaluating the dual reformulation, which can be accomplished using a bisection search algorithm. Consequently, the computational cost may be higher. However, this regularization is more stable, especially for small values of η, whereas using small η in entropic regularization can lead to numerical errors due to the log-sum-exp operator. Besides, quadratic regularization implicitly promotes sparsity on the support of the worst-case distribution, as the density corresponding to support point ω equals zero when $f(\omega) < \mu_z^*$ for $\widehat{P}$-almost sure z.

Example 8 (Absolute Value Regularization)**.** *Consider $\phi(x) = |x-1|, x \geq 0$. Assume*

$$\|f\|_{\nu_z,\infty} := \textit{ess-sup}_{\nu_z} f = \max_{z'\in\mathbb{B}_\rho(z)} f(z') < \infty \tag{6.6}$$

for $\widehat{P}$-almost surely z. One can verify that

$$\mu_z^* = -\eta + \|f\|_{\nu_z,\infty}, \quad \zeta_z^*(\omega) = \mathbf{1}\Big\{f(\omega)+2\eta-\|f\|_{\nu_z,\infty}\geq 0\Big\}.$$

Hence, the worst-case distribution $\mathbb{P}_*$ has the density

$$\frac{\mathrm{d}\mathbb{P}_*(\omega)}{\mathrm{d}\omega} = \mathsf{V}_\rho^{-1}\cdot\mathbb{E}_{z\sim\widehat{P}}\Big[\mathbf{1}\Big\{\omega\in\mathbb{B}_\rho(z)\bigwedge f(\omega)+2\eta-\|f\|_{\nu_z,\infty}\geq 0\Big\}\Big].$$

In this case,

$$\text{(Dual-}\phi\text{-Reg)} = \mathbb{E}_{z\sim\widehat{P}}\Big[\|f\|_{\nu_z,\infty}-2\eta+\mathbb{E}_{z'\sim\nu_z}\Big[f(z')-\|f\|_{\nu_z,\infty}+2\eta\Big]_+\Big]. \qquad \clubsuit$$

Example 9 (Hinge Loss Regularization)**.** *Consider $\phi(x) = (x-1)_+, x \geq 0$. Under the*

same assumption as in Example 8, one can verify

$$\mu_z^* = -\eta + \|f\|_{\nu_z,\infty}, \quad \zeta_z^*(\omega) = \mathbf{1}\{f(\omega) - \mu_z^* \geq 0\}.$$

Hence, the worst-case distribution $\mathbb{P}_*$ has the density

$$\frac{\mathrm{d}\mathbb{P}_*(\omega)}{\mathrm{d}\omega} = \mathsf{V}_\rho^{-1} \cdot \mathbb{E}_{z\sim\widehat{P}}\Big[\mathbf{1}\Big\{\omega \in \mathbb{B}_\rho(z) \bigwedge f(\omega) + \eta - \|f\|_{\nu_z,\infty} \geq 0\Big\}\Big].$$

In this case,

$$\text{(Dual-}\phi\text{-Reg)} = \mathbb{E}_{z\sim\widehat{P}}\Big[\|f\|_{\nu_z,\infty} - \eta + \mathbb{E}_{z'\sim\nu_z}\Big[f(z') - \|f\|_{\nu_z,\infty} + \eta\Big]_+\Big]. \qquad \clubsuit$$

Remark 24 (Connections with Bayesian DRO)**.** *Our formulation is closely related to the dual formulation of Bayesian DRO [282, Eq. (2.10)] with two major differences: (i) we treat the reference measure ν_z as an uniform distribution supported on $\mathbb{B}_\rho(z)$, while the authors therein consider a more general conditional distribution; (ii) we fix the regularization value η, while the authors therein treat it as a Lagrangian multiplier associated with the hard constraint $\mathbb{E}_{z'\sim\nu_z}\left[\phi\left(\frac{\mathrm{d}\gamma_z(z')}{\mathrm{d}\nu_z(z')}\right)\right] \leq \eta'$ for some constant $\eta' > 0$.* $\clubsuit$

Following the discussion in Example 6, we are curious to study under which condition will the regularized formulation serve as the smooth approximation of the classical adversarial training formulation as the regularization value vanishes, called the *consistency property*. Proposition 6 gives its sufficient condition. Its proof is provided in Appendix E.2.

Assumption 6. *Assume either one of the following conditions hold:*

(I) $\lim_{t\to\infty} \frac{\phi(t)}{t} < \infty$*;*

(II) $\lim_{t\to\infty} \frac{\phi(t)}{t} = \infty$*, dom*$(\phi) = \mathbb{R}_+$.

Proposition 6 (Consistency of Regularized Formulation)**.** *Suppose Assumption 6 holds, and for any $\eta > 0$ and $\widehat{P}$-almost sure z, the minimizer to the inner infimum problem in*

(Dual-ϕ-Reg) *exists and is finite. Then, as* $\eta \to 0$*, the optimal value of* (Primal-ϕ-Reg) *converges to* $\mathbb{E}_{z\sim\widehat{P}}\left[\max_{z'\in\mathbb{B}_\rho(z)} f(z')\right]$.

Assumption 6 is widely utilized in the ϕ-divergence DRO literature [207]. Within this context, Assumption 6(I) is referred as *popping property* and the first condition of 6(II) is referred as *non-popping property*. Under the non-popping property, we further assume that the domain of ϕ is $\mathbb{R}_+$. This assumption is crucial: the indicator function in Example 5 does not satisfy this assumption, and consequently, the consistency property in this case does not hold. However, as demonstrated in Proposition 6, the divergence choices in Examples 6-9 do satisfy this consistency property. As we will demonstrate in the next subsection, by taking the regularization value $\eta \to 0$, the worst-case distributions from Examples 6-9 indeed concentrate around the worse-case perturbation area.

6.2.2 Visualization of Worst-case Distribution

In this subsection, we display the worst-case distributions studied in Examples 5-9 using a toy example. We obtain the densities of these distributions by discretizing the continuous distribution ν_z using 10^4 grid points. The loss $f(\cdot)$ is constructed using a three-layer neural network, whose detailed configuration is provided in Appendix E.7 and landscape is displayed in Figure 6.1. We take $\widehat{P} = \delta_0$, and the domain of adversarial attack as $\mathbb{B}_\rho(z) = [-5, 5]$. The inner maximization problem corresponding to the un-regularized formulation (6.1) amounts to solve the optimization problem $\max_{z\in[-5,5]} f(z)$. From the plot of Figure 6.1, we can see that solving the un-regularized adversarial learning problem is highly non-trivial because the inner maximization problem contains many local maxima, and it is difficult to find the global maxima.

The worse-case distributions are provided in Figure 6.2, in which different columns correspond to different regularizations (i.e., indicator, entropic, quadratic, absolute value, or hinge loss), and different rows correspond to different choices of parameters: for indicator regularization, we tune the risk level $\alpha \in (0, 1]$, whereas for other regularizations, we tune

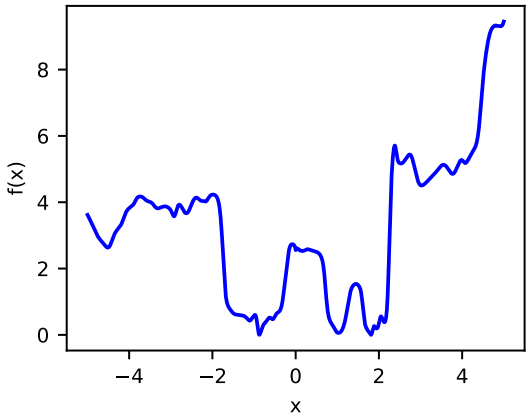

Figure 6.1: Landscape of the 1-dimensional objective $f(\cdot)$

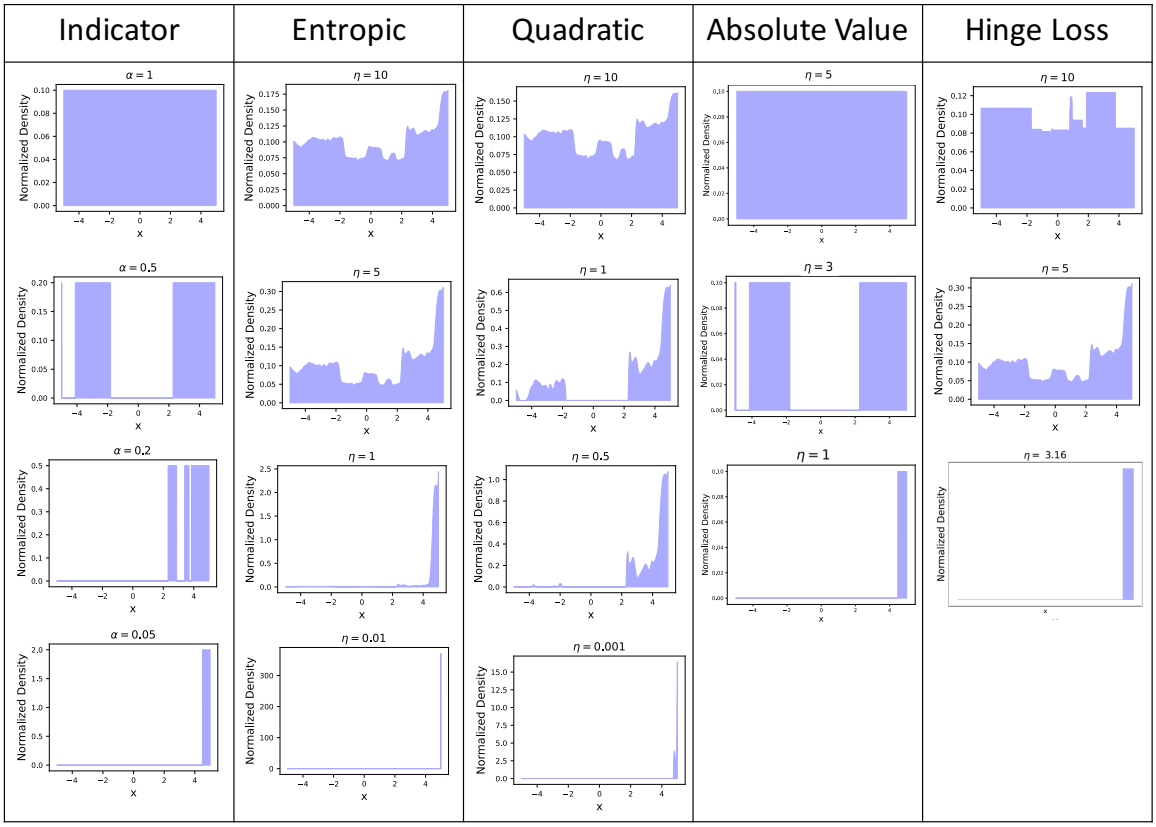

Figure 6.2: Worse-case distributions for different kinds of regularizations and different choices of parameters (including risk level α and regularization level η).

the regularization level η. Our findings are summarized as follows.

(I) For indicator regularization, the worst-case distribution does not vary w.r.t. the choice of regularization value η but the risk level α. From the plot, we can see that the

worst-case distribution for $\alpha = 1$ becomes the uniform distribution around the domain, whereas as α decreases, it tends to center around the worst-case perturbation area, i.e., the area close to $x = 5$.

(II) For entropic regularization, we find for large η, the worst-case distribution tends to be uniform over its support, whereas, for small η, the worst-case distribution tends to center around the worst-case perturbation area. Compared to the plot for indicator regularization, the worst-case distribution here demonstrates greater flexibility by allowing unequal weight values across different support points.

(III) For quadratic regularization, we obtain similar observations as in entropic regularization. Besides, the maximum density value does not increase to infinity at such a quick rate, which demonstrates that the quadratic regularized formulation is more numerically stable to solve. One should also notice that even for small η, the support of the worst-case distribution from entropic regularization is still the whole domain $[-5, 5]$, whereas most density values can be extremely small. In contrast, the support of that from quadratic regularization only takes a tiny proportion of the whole domain.

(IV) For absolute value and hinge loss regularizations, unlike entropic or quadratic regularization, the worst-case distributions here are constructed using histograms with equal weights assigned to different support points. This indicates that these two choices lack the flexibility needed to represent a meaningful worst-case distribution.

Based on the discussions above, we recommend using entropic or quadratic regularization for adversarial robust learning in practice. Since these two regularizations are special cases of the Cressie-Read family of ϕ-divergences [90], exploring other types of ϕ-divergences as regularizations opens an interesting avenue for further research.

6.3 Optimization Algorithm

We now develop stochastic gradient-type methods to solve the proposed formulation (Reg-∞-WDRO). We first re-write it as

$$\min_{\theta\in\Theta}\left\{F(\theta)=\mathbb{E}_{z\sim\widehat{P}}\ [R(\theta;z)]\right\} \tag{6.7a}$$

$$\text{where}\quad R(\theta;z)=\sup_{\gamma\in\mathcal{P}(\mathcal{Z})}\left\{\mathbb{E}_{z'\sim\gamma}[f_\theta(z)]-\eta\mathbb{E}_{z'\sim\nu_z}\left[\phi\left(\frac{\mathrm{d}\gamma(z')}{\mathrm{d}\nu_z(z')}\right)\right]\right\},\quad\forall z. \tag{6.7b}$$

The formulation above is difficult to solve because each $z\in\operatorname{supp}\widehat{P}$ corresponds to a lower-level subproblem (6.7b). Consequently, Problem (6.7) necessitates solving a large number of these subproblems, given that the size of $\operatorname{supp}\widehat{P}$ is typically large or even uncountably infinite. In contrast, we will provide an efficient optimization algorithm whose sample complexity is *near-optimal* and *independent* of the size of $\operatorname{supp}\widehat{P}$.

Throughout this section, we assume the divergence function ϕ is strongly convex with modulus κ, which is a standard condition studied in literature. Thus, for each z, the maximizer (denoted as $\bar{\gamma}_z$) to the lower level problem (6.7b) is unique and guaranteed to exist. Therefore, we update θ in the outer minimization problem according to the projected stochastic gradient descent (SGD) [1] outlined in Algorithm 10.

Algorithm 10 Projected SGD for solving (6.7)

Require: Maximum iteration T, initial guess θ_1, constant stepsize τ

1: **for** $t=1,\ldots,T-1$ **do**
2: Obtain a stochastic estimator of the (sub-)gradient $\nabla F(\theta_t)$.
3: Update $\theta_{t+1}=\operatorname{Proj}_\Theta\Big(\theta_t-\tau V(\theta_t)\Big)$.
4: **end for**

Output iteration points $\{\theta_t\}_{t=1}^T$.

The Step 2 of Algorithm 10 requires the construction of the gradient of the objective at

[1] The projection $\operatorname{Proj}_\Theta(\cdot)$ can be replaced by the generalized projection mapping defined by the proximal operator. This modified algorithm is called stochastic mirror descent, which incorporates the geometry of the constraint set Θ and results in the same order of complexity bound but with (potentially) lower constant.

the upper level. According to the Danskin's theorem, it holds that

$$\nabla F(\theta) = \mathbb{E}_{z\sim\widehat{P}}[\nabla R(\theta; z)] = \mathbb{E}_{z\sim\widehat{P}}\mathbb{E}_{z'\sim\bar{\gamma}_z}[\nabla f_\theta(z')].$$

In the following, we discuss how to construct the stochastic (sub-)gradient estimator $V(\cdot)$ in Step 2 of Algorithm 10. More specifically, how to construct the estimator of $\nabla R(\theta; z)$.

6.3.1 Gradient Estimators

Since $\nabla R(\theta; z)$ is challenging to estimate, we first construct an approximation objective of $R(\theta; z)$ whose gradient is easier to estimate. Denote the collection of random sampling parameters $\zeta^\ell := (z, \{z'_i\}_{i\in[2^\ell]})$, where $z \sim \widehat{P}$ and $\{z'_i\}_{i\in[2^\ell]}$ are 2^ℓ i.i.d. samples generated from distribution ν_z. Then, we define the approximation function

$$F^\ell(\theta) = \mathbb{E}_{\zeta^\ell}\left[\widehat{R}\big(\theta; \{z'_i\}_{i\in[2^\ell]}\big)\right], \tag{6.8}$$

where for fixed decision θ, sample z, and sample set $\{z'_i\}$ consisting of m samples, define

$$\widehat{R}\big(\theta; \{z'_i\}\big) = \max_{\gamma\in\Delta^m}\left\{\sum_{i\in[m]}\gamma_i f_\theta(z'_i) - \frac{\eta}{m}\sum_{i\in[m]}\phi(m\gamma_i)\right\}. \tag{6.9}$$

The function $\widehat{R}\big(\theta; \{z'_i\}\big)$ can be viewed as the optimal value of the ϕ-divergence DRO with discrete empirical distribution supported on $\{z'_i\}$. The high-level idea of the approximation function $F^\ell(\theta)$ is to replace the lower-level problem (6.7b), the ϕ-divergence DRO with continuous reference distribution, using another ϕ-divergence DRO with its empirical reference distribution. As the number of samples of the empirical reference distribution goes to infinity, one can expect that $F^\ell(\theta)$ approximates the original objective $F(\theta)$ with negligible error.

It is easy to generate gradient estimators for the approximation function $F^\ell(\theta)$. For fixed random sampling parameter ζ^ℓ, assume there exists an oracle that returns $\widetilde{\gamma}_{n_1:n_2}$ as

near-optimal probability mass values for $\widehat{R}\big(\theta; \{z_i'\}_{i\in[n_1:n_2]}\big)$, and define the gradient

$$\nabla\widetilde{R}\big(\theta; \{z_i'\}_{i\in[n_1:n_2]}\big) = \sum_{i\in[n_1:n_2]} (\widetilde{\gamma}_{n_1:n_2})_i \nabla_\theta f_\theta(z_i'). \tag{6.10}$$

Due to the near-optimality of $\widetilde{\gamma}_{n_1:n_2}$, it holds that

$$\nabla\widetilde{R}\big(\theta; \{z_i'\}_{i\in[n_1:n_2]}\big) \approx \nabla\widehat{R}\big(\theta; \{z_i'\}_{i\in[n_1:n_2]}\big).$$

Next, we define

$$g^\ell(\theta, \zeta^\ell) = \nabla\widetilde{R}\big(\theta; \{z_i'\}_{i\in[1:2^\ell]}\big) \tag{6.11}$$

$$G^\ell(\theta, \zeta^\ell) = \nabla\widetilde{R}\big(\theta; \{z_i'\}_{i\in[1:2^\ell]}\big) - \frac{1}{2}\Big[\nabla\widetilde{R}\big(\theta; \{z_i'\}_{i\in[1:2^{\ell-1}]}\big) + \nabla\widetilde{R}\big(\theta; \{z_i'\}_{i\in[2^{\ell-1}+1:2^\ell]}\big)\Big]. \tag{6.12}$$

Now, we list two choices of gradient estimators at a point θ:

Stochastic Gradient (SG) Estimator. For fixed level L, generate n_L^{o} i.i.d. copies of ζ^L, denoted as $\{\zeta_i^L\}$. Then construct

$$V^{\mathrm{SG}}(\theta) = \frac{1}{n_L^{\mathrm{o}}} \sum_{i=1}^{n_L^{\mathrm{o}}} g^L(\theta, \zeta_i^L). \tag{6.13a}$$

Randomized Truncation MLMC (RT-MLMC) Estimator. For fixed level L, generate n_L^{o} i.i.d. random levels following the *truncated geometric distribution* $\mathbb{P}(\widehat{L} = \ell) = \frac{2^{-\ell}}{2-2^{-L}}, \ell = 0, \ldots, L$, denoted as $\widehat{L}_1, \ldots, \widehat{L}_{n_L^{\mathrm{o}}}$. Then construct

$$V^{\text{RT-MLMC}}(\theta) = \frac{1}{n_L^{\mathrm{o}}} \sum_{i=1}^{n_L^{\mathrm{o}}} \mathbb{P}(\widehat{L} = \widehat{L}_i)^{-1} \cdot G^{\widehat{L}_i}(\theta, \zeta_i^{\widehat{L}_i}). \tag{6.13b}$$

The SG estimator is a conventional approach to estimate $\nabla F^\ell(\theta)$. Instead, the RT-MLMC estimator has the following attractive features:

(I) It constitutes a gradient estimator of the approximation function $F^L(\theta)$ with a small bias. More specifically, RT-MLMC and SG estimators have the same bias:

$$\begin{aligned}\mathbb{E}[V^{\text{RT-MLMC}}(\theta)] &= \mathbb{E}_{\widehat{L}_1}\left[\frac{1}{\mathbb{P}(\widehat{L}=\widehat{L}_1)}\mathbb{E}_{\zeta^{\widehat{L}_1}}[G^{\widehat{L}_1}(\theta,\zeta^{\widehat{L}_1})]\right] \\ &= \sum_{\ell=0}^{L}\mathbb{P}(\widehat{L}=\ell)\cdot\left[\frac{1}{\mathbb{P}(\widehat{L}=\ell)}\mathbb{E}_{\zeta^\ell}[G^\ell(\theta,\zeta^\ell)]\right] \\ &= \sum_{\ell=0}^{L}\mathbb{E}_{\zeta^\ell}[G^\ell(\theta,\zeta^\ell)] = \mathbb{E}_{z\sim\widehat{P}}\mathbb{E}_{\{z_i'\}_{i\in[2^\ell]}\sim\nu_z}\left[\nabla\widetilde{R}\big(\theta;\{z_i'\}_{i\in[2^\ell]}\big)\right] \\ &= \mathbb{E}[V^{\text{SG}}(\theta)],\end{aligned}$$

and the bias vanishes quickly as $L\to\infty$.

(II) Since $\nabla\widetilde{R}\big(\theta;\{z_i'\}_{i\in[1:2^\ell]}\big)$, $\nabla\widetilde{R}\big(\theta;\{z_i'\}_{i\in[1:2^{\ell-1}]}\big)$, and $\nabla\widetilde{R}\big(\theta;\{z_i'\}_{i\in[2^{\ell-1}+1:2^\ell]}\big)$ are generated using the same random sampling parameters ζ^ℓ, they are highly correlated, which indicates the stochastic estimator $G^\ell(\theta,\zeta^\ell)$ defined in (6.12) has small second-order moment and variance thanks to the control variate effect [261], making it a suitable recipe for gradient simulation.

(III) The construction of SG estimator requires generating $n_L^{\text{o}}\cdot 2^L=\mathcal{O}(2^L)$ samples, while the (expected) number of samples for RT-MLMC is $n_L^{\text{o}}\cdot\frac{L}{2-2^{-L}}=\mathcal{O}(L)$. As a result, the computation of RT-MLMC estimator is remarkably smaller than that of SG estimator.

6.3.2 Solving penalized ϕ-divergence DRO with finite support

In the last subsection, it is assumed that one has the oracle for solving the generic penalized ϕ-divergence DRO with m support points $\{f_1,\ldots,f_m\}$:

$$\mathcal{R}=\max_{\gamma\in\Delta^m}\left\{\sum_{i\in[m]}\gamma_i f_i-\frac{\eta}{m}\sum_{i\in[m]}\phi(m\gamma_i)\right\}. \tag{6.14}$$

The formulation (6.9) is a special case of this problem by taking $f_i = f_\theta(z_i'), \forall i$. In the following, we provide an algorithm that returns returns the optimal solution to (6.14) up to precision ϵ. We write its Lagrangian reformulation as

$$\min_{\mu} \max_{\gamma \in \mathbb{R}_+^m} \Big\{ \mathcal{L}(\lambda, \gamma) = \sum_{i \in [m]} \gamma_i f_i - \frac{\eta}{m} \sum_{i \in [m]} \phi(m\gamma_i) + \mu \Big(1 - \sum_{i \in [m]} \gamma_i \Big) \Big\}.$$

Based on this minimax formulation, we present an efficient algorithm that finds a near-optimal primal-dual solution to (6.14) in Algorithm 11, whose complexity analysis is presented in Proposition 7. The complexity is quantified as the number of times to query samples $f_1, \ldots, f_m$. Its proof is provided in Appendix E.3.

Algorithm 11 Bisection search for solving (6.14)

Require: Interval $[\underline{\mu}, \overline{\mu}]$, maximum iteration T, constant $K = \lim_{s \to 0+} \phi'(s)$.

1: **for** $t = 1, \ldots, T$ **do**
2: Update $\mu = \frac{1}{2}(\underline{\mu} + \overline{\mu})$
3: Obtain index set $\mathcal{N} = \Big\{ i \in [m] : \ f_i \leq \mu + \eta K \Big\}$.
4: Compute $h(\mu) := \frac{1}{m} \sum_{i \in [m] \setminus \mathcal{N}} (\phi')^{-1} \Big(\frac{f_i - \mu}{\eta} \Big) - 1$.
5: Update $\overline{\mu} = \mu$ if $h(\mu) \leq 0$ and otherwise $\underline{\mu} = \mu$.
6: **end for**
7: Obtain γ^* such that $\gamma_i = 0$ if $i \in \mathcal{N}$ and otherwise $\gamma_i = \frac{1}{m} (\phi')^{-1} (\frac{f_i - \mu}{\eta})$.
Output the estimated primal-dual optimal solution (γ, μ) and estimated optimal value $h(\mu)$.

Proposition 7 (Performance Guarantees of Algorithm 11)**.** *Fix the precision $\epsilon > 0$. Suppose we choose hyper-parameters in Algorithm 11 as*

$$T = \frac{1}{2} \log_2 \left(\frac{\varrho^2}{2\kappa\eta} \cdot \frac{1}{\epsilon} \right) = \mathcal{O}(\log \frac{1}{\epsilon}),$$

$$\underline{\mu} = \underline{f},$$

$$\overline{\mu} = \begin{cases} \overline{f} - \eta(\underline{f} - \overline{f}), & \text{if } \phi'(s) \to -\infty \text{ as } s \to 0+, \\ \overline{f} - \eta K, & \text{if } \phi'(s) \to K > -\infty \text{ as } s \to 0+, \end{cases}$$

where $\varrho = \overline{\mu} - \underline{\mu}$, $\underline{f} = \min_{i\in[m]} f_i$, *and* $\overline{f} = \max_{i\in[m]} f_i$. *As a consequence, Algorithm 11 finds a primal-dual solution to* (6.14) *such that*

(I) *the estimated objective value* $\widetilde{\mathcal{R}}$ *satisfies* $|\widetilde{\mathcal{R}} - \mathcal{R}| \leq \epsilon$;

(II) *the difference between the estimated primal solution* γ *and the optimal primal solution* γ^* *is bounded:* $\|\gamma - \gamma^*\|_\infty \leq \frac{1}{m}\sqrt{\frac{2\epsilon}{\kappa\eta}}$;

(III) *the distance between the estimated dual solution* μ *and the set of optimal dual solutions* $\mathcal{S}^*$ *is bounded:* $\mathsf{D}(\mu, \mathcal{S}^*) \leq \varrho \cdot 2^{-T} = (2\eta\kappa\epsilon)^{1/2}$;

(IV) *its worst-case computational cost is* $\mathcal{O}(mT) = \mathcal{O}(m \log \frac{1}{\epsilon})$.

Most literature (such as [133, 155, 231]) provided algorithms for solving hard-constrained ϕ-divergence DRO problem, but they did not study how to extend their framework for penalized ϕ-divergence DRO in (6.14). One exception is that Levy et al. [191] mentioned this problem can be solved using bisection search as in Algorithm 11 but did not provide detailed parameter configurations.

Remark 25 (Near-Optimality of Algorithm 11)**.** *For some special choices of the divergence function* ϕ, *the optimal solution of problem* (6.9) *can be obtained with closed-form solution, such as the entropy function* $\phi(s) = s\log s - s + 1$. *However, Algorithm 11 is applicable to solving problem* (6.9) *for general choices of the divergence function. When considering* $\phi(s) = s\log s - s + 1$, *the optimal solution to* (6.14) *becomes*

$$\gamma_i^* = \frac{e^{f_i/\eta}}{\sum_{i\in[m]} e^{f_i/\eta}}, \quad \forall i \in [m].$$

Computing this optimal solution requires computational cost at least $\Omega(m)$. *Compared with the complexity in Proposition 7, Algorithm 11 is a near-optimal choice because it matches the lower bound up to negligible constant* $\mathcal{O}(\log \frac{1}{\epsilon})$. ♣

6.3.3 Complexity Analysis

In this subsection, we provide the convergence analysis of our projected SGD algorithm using SG and RT-MLMC gradient estimators. Throughout this subsection, the computational cost is quantified as the number of times to generate samples $z \sim \widehat{P}$ or samples $z' \sim \nu_z$ for any $z \in \operatorname{supp} \widehat{P}$. We consider the following assumptions regarding the loss function.

Assumption 7 (Loss Assumptions)**.** (I) *(Convexity): The loss $f_\theta(z)$ is convex in θ.*

(II) *(Lipschitz Continuity): 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) *(Boundedness): for any z and θ, it holds that $0 \leq f_\theta(z) \leq B$.*

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

Nonsmooth Convex Loss

To analyze the convergence rate, we rely on the following technical assumptions.

Assumption 8. (I) *For any data points $\{f_1, \ldots, f_m\}$, the optimal probability vector γ^* in the data-driven penalized ϕ-divergence DRO problem* (6.14) *satisfies $\mathcal{D}_{\mathcal{X}^2}(\gamma^*, \frac{1}{m}\mathbf{1}_m) \leq C$.*

(II) *For each $\theta \in \Theta$ and $z \in \operatorname{supp} \widehat{P}$, the inverse cdf of the random variable $(f_\theta)_{\#}\nu_z$ is G_{idf}-Lipschitz.*

Assumption 8 is relatively mild and has originally been proposed in [191, Assumption A1] to investigate the complexity of solving standard ϕ-divergence DRO. Assumption 8(I) holds by selecting proper divergence function ϕ, such as quadratic or entropy function in Examples 7 and 6. As long as the probability density of $f_\theta(Z)$ is lower bounded by Υ within its support, Assumption 8(II) holds with $G_{\mathrm{idf}} = \Upsilon^{-1}$. Now, we derive statistics of SG and RT-MLMC estimators.

Proposition 8 (Bias/Second-order-Moment/Cost of SG and RT-MLMC Estimators)**.** *Fix the precision $\epsilon > 0$. Suppose Assumption 8*(I) *holds, and during the construction of SG/RT-MLMC estimators, one query Algorithm 11 with optimality gap controlled by ϵ. Then it holds that*

(I) *(Bias): Suppose, in addition, Assumption 8*(II) *holds, then*

$$\mathbb{E}[V^{\mathrm{SG}}(\theta)] = \mathbb{E}[V^{\text{RT-MLMC}}(\theta)] = \nabla \widetilde{F}(\theta),$$

where $|\widetilde{F}(\theta) - F(\theta)| \leq \epsilon + G_{\mathrm{idf}} \cdot 2^{-L}$.

(II) *(Second-order Moment):*

$$\begin{aligned}
\mathbb{E}\|V^{\mathrm{SG}}(\theta)\|_2^2 &\leq 2L_f^2\left[1 + (2\epsilon)/(\kappa\eta)\right], \\
\mathbb{E}\|V^{\text{RT-MLMC}}(\theta)\|_2^2 &\leq \frac{96L_f^2}{\kappa\eta} \cdot (2^L\epsilon) + 6(L+1)L_f^2 C.
\end{aligned}$$

(III) *(Variance):*

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[V^{\mathrm{SG}}(\theta)] &\leq \frac{2L_f^2\left[1 + (2\epsilon)/(\kappa\eta)\right]}{n_L^{\mathrm{o}}}, \\
\mathbb{V}\mathrm{ar}[V^{\text{RT-MLMC}}(\theta)] &\leq \frac{1}{n_L^{\mathrm{o}}}\left[\frac{96L_f^2}{\kappa\eta} \cdot (2^L\epsilon) + 6(L+1)L_f^2 C\right].
\end{aligned}$$

(IV) *(Cost): Generating a single SG estimator requires cost $\mathcal{O}(n_L^{\mathrm{o}} \cdot 2^L \log\frac{1}{\epsilon})$, whereas generating a single RT-MLMC estimator requires expected cost $\mathcal{O}(n_L^{\mathrm{o}} \cdot L\log\frac{1}{\epsilon})$.*

Let the estimated solution returned by the projected SGD algorithm be $\widetilde{\theta}_{1:T} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$. Based on Proposition 8, we derive complexity bounds for solving (6.7) when the loss function is convex and Lipschitz continuous. We formalize our results in the following theorem. Its proof is provided in Appendix E.3.

Theorem 24 (Complexity for Nonsmooth Convex Loss)**.** *Suppose Assumptions 7*(I)*, 7*(II)*, 8 hold, and $\delta > 0$ is a sufficiently small precision level. During the construction of SG/RT-MLMC estimators, let the optimality gap of querying Algorithm 11 controlled by $\epsilon = \frac{\delta}{8}$, and specify the hyper-parameters of SGD algorithm with SG or RT-MLMC estimators as in Table 6.1. As a result,*

(I) *(SG Estimator) The SGD algorithm with SG estimator finds a δ-optimal solution to* (6.7) *with computational cost $\mathcal{O}(T \cdot n_L^{\circ} 2^L \log \frac{1}{\epsilon}) = \mathcal{O}(\delta^{-3} \log \frac{1}{\delta})$.*

(II) *(RT-MLMC Estimator) The SGD algorithm with RT-MLMC estimator finds a δ-optimal solution to* (6.7) *with computational cost $\mathcal{O}(T \cdot n_L^{\circ} L \log \frac{1}{\epsilon}) = \mathcal{O}(\delta^{-2} (\log \frac{1}{\delta})^4)$.*

Table 6.1: Hyper-parameters used in the projected SGD algorithm with SG/RT-MLMC gradient estimators for nonsmooth convex loss.

Method	Batch Size n_L°	Max Level L	Max Iteration T	Step Size γ
SG	1	$\log \frac{8G_{\text{idf}}}{\delta}$	$\mathcal{O}(1/\delta^2)$	$\mathcal{O}(\delta)$
RT-MLMC	1	$\log \frac{8G_{\text{idf}}}{\delta}$	$\mathcal{O}((\log 1/\delta)^2/\delta^2)$	$\mathcal{O}((\log 1/\delta)^{-2}\delta)$

Smooth Nonconvex Loss

When the loss $f_\theta(z)$ is nonconvex in θ, we focus on finding the near-stationary point of (6.7) instead. The key in this part is to build the bias between our gradient estimator in (6.13a) or (6.13b) and the true gradient of the objective. Unfortunately, such a result for general choice of ϕ-divergence regularization is hard to show. In this part, we only investigate the convergence behavior of entropic regularization (see Example 6). In such a case, we have the closed-form expression regarding the optimal solution of the lower-level problem (6.7b), and therefore, Problem (6.7) can be reformulated as

$$\min_{\theta \in \Theta} \left\{ F(\theta) = \mathbb{E}_{z \sim \widehat{P}} \left[\eta \log \mathbb{E}_{z' \sim \nu_z} \left[\exp\left(\frac{f_\theta(z')}{\eta} \right) \right] \right] \right\}. \tag{6.15}$$

Similarly, the approximation function F^ℓ defined in (6.8) becomes

$$F^\ell(\theta) = \mathbb{E}_{\zeta^\ell}\left[\eta \log\left(\frac{1}{2^\ell}\sum_{i\in[2^\ell]} \exp\left(\frac{f_\theta(z_i')}{\eta}\right)\right)\right]. \tag{6.16}$$

In this case, we do not use (6.11) or (6.12) but adopt the following way to construct the random vectors $g^\ell(\theta, \zeta^\ell)$ and $G^\ell(\theta, \zeta^\ell)$: define

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

and construct

$$g^\ell(\theta, \zeta^\ell) = \nabla_\theta U_{1:2^\ell}(\theta, \zeta^\ell), \tag{6.17}$$

$$G^\ell(\theta, \zeta^\ell) = \nabla_\theta\left[U_{1:2^\ell}(\theta, \zeta^\ell) - \frac{1}{2}U_{1:2^{\ell-1}}(\theta, \zeta^\ell) - \frac{1}{2}U_{2^{\ell-1}+1:2^\ell}(\theta, \zeta^\ell)\right]. \tag{6.18}$$

The following theorem presents the complexity of obtaining a δ-stationary point for our projected SGD algorithm using either SG or RT-MLMC estimator. Its proof is provided in Appendix E.3.

Theorem 25 (Complexity for Smooth Nonconvex Loss)**.** *Under Assumptions 7*(II)*, 7*(III)*, and 7*(IV)*, with properly chosen hyper-parameters of the RT-MLMC estimator as in Table 6.2, the following results hold:*

(I) *(Smooth Nonconvex Optimization) The computation cost of RT-MLMC scheme for finding ϵ-stationary point is of $\widetilde{\mathcal{O}}(\epsilon^{-4})$ with memory cost $\widetilde{\mathcal{O}}(\epsilon^{-2})$.*

(II) *(Unconstrainted Smooth Nonconvex Optimization) Additionally assume the constraint set $\Theta = \mathbb{R}^{d_\theta}$, then the memory cost of RT-MLMC improves to $\widetilde{\mathcal{O}}(1)$.*

Remark 26 (Comparision with Sinkhorn DRO)**.** *Sinkhorn DRO [305] introduces entropic regularization to the ambiguity set constructed using the p-Wasserstein distance, resulting*

Table 6.2: Hyper-parameters used in the projected SGD algorithm with SG/RT-MLMC gradient estimators for smooth nonconvex loss.

Scenarios	Hyper-parameters	Comp./Memo.
Smooth Nonconvex Optimization	$L = \mathcal{O}(\log \frac{1}{\epsilon^2}), \quad T = \widetilde{\mathcal{O}}(\epsilon^{-2})$ $n_L^{\circ} = \widetilde{\mathcal{O}}(\epsilon^{-2}), \quad \gamma = O(1)$	Comp. $= \mathcal{O}(T(n_L^{\circ} L)) = \widetilde{\mathcal{O}}(\epsilon^{-4})$ Memo. $= \mathcal{O}(n_L^{\circ} L) = \widetilde{\mathcal{O}}(\epsilon^{-2})$
Unconstrainted Smooth Nonconvex Optimization	$L = \mathcal{O}(\log \frac{1}{\epsilon^2}), \quad T = \widetilde{\mathcal{O}}(\epsilon^{-4})$ $n_L^{\circ} = \mathcal{O}(1), \quad \gamma = \widetilde{\mathcal{O}}(\epsilon^{2})$	Comp. $= \mathcal{O}(T(n_L^{\circ} L)) = \widetilde{\mathcal{O}}(\epsilon^{-4})$ Memo. $= \mathcal{O}(n_L^{\circ} L) = \widetilde{\mathcal{O}}(1)$

in a dual reformulation that closely resembles (6.15)*, with the key modification of replacing the uniform distribution ν_z with certain kernel probability distributions. The primary distinction lies in the fact that the authors of the original work provide only the SG or RT-MLMC estimator for nonsmooth convex loss, whereas we extend their analysis to the smooth nonconvex loss setting. A crucial aspect of the convergence analysis is that for the constrained smooth nonconvex case, as highlighted in [132], a large mini-batch size n_L° is required at each iteration to estimate the gradient with sufficiently small variance to ensure convergence. In contrast, for the unconstrained case, a mini-batch size $n_L^{\circ} = \mathcal{O}(1)$ is sufficient.* ♣

Remark 27 (Comparison with ∞-WDRO)**.** *When solving* (∞-WDRO)*, the involved sub-problems are finding the global optimal value of the supremum $\sup_{z \in \mathbb{B}_\rho(x)} \ f(z)$ for $x \in \operatorname{supp} \widehat{P}$, which are computationally challenging in general. Various heuristics [137, 185, 212] have been proposed to approximately solve it by replacing $f(z)$ with its linear approximation $f(x) + \nabla f(x)^{\mathrm{T}} z$. It is worth noting that such an approximation is not accurate, especially when the radius ρ of domain set $\mathbb{B}_\rho(x)$ is moderately large, which corresponds to large adversarial perturbation scenarios. For example, for the loss $f(z)$ depicted in Figure 6.1, its linear approximation around $x = 0$ will yield a wrong global maximum estimate. In contrast, we proposed stochastic gradient methods to solve the regularized formulation* (Reg-∞-WDRO) *with provable convergence guarantees, which avoids solving*

such a hard maximization subproblem. Numerical comparisons in Section 6.6.1 also suggest that our method outperforms those heuristics when adversarial perturbations are moderately large. ♣

6.4 Regularization Effects of Regularized Adversarial Robust Learning

In this section, we provide an interpretation on how our proposed formulation (Reg-∞-WDRO) works by showing its close connection to the regularized ERM problem:

$$\min_{\theta\in\Theta} \mathbb{E}_{z\sim\widehat{P}}[f_\theta(z)] + \mathcal{R}(f_\theta;\rho,\eta)$$

for certain regularization $\mathcal{R}(f_\theta;\rho,\eta)$. As we focus on small-perturbation attacks, it is assumed that $\rho,\eta\to 0$. Also, we omit the dependence of f_θ on θ for simplicity. Subsequently, we derive the regularization effects of (Reg-∞-WDRO) by considering different scaling of ρ and η. To begin with, we define regularizer $\mathcal{E}$ as the difference between regularized robust loss in (Primal-ϕ-Reg) and non-robust loss:

$$\mathcal{E}_{\widehat{P}}(f;\rho,\eta) = \text{Optval(Primal-}\phi\text{-Reg)} - \mathbb{E}_{\widehat{P}}[f]. \tag{6.19}$$

Besides, we define the following regularizations. Let β be the uniform probability distribution supported on $\mathbb{B}_1(0)$, and define

$$\mathcal{R}_1(f;\rho,\eta) = \rho\cdot\mathbb{E}_{z\sim\widehat{P}}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\frac{1}{C}\mathbb{E}_{b\sim\beta}\left[\phi^*\Big(C\cdot(\nabla f(z)^{\mathrm{T}}b-\mu)\Big)\right]\right\}\right], \tag{6.20a}$$

$$\mathcal{R}_2(f;\rho,\eta) = \rho\cdot\mathbb{E}_{z\sim\widehat{P}}\Big[\|\nabla f(z)\|_*\Big], \tag{6.20b}$$

$$\mathcal{R}_3(f;\rho,\eta) = \frac{\rho^2}{2\eta\cdot\phi''(1)}\cdot\mathbb{E}_{z\sim\widehat{P}}\Big[\mathbb{V}\mathrm{ar}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}}b]\Big], \tag{6.20c}$$

where $C>0$ is some constant to be specified. These three regularizations correspond to the asymptotic approximations of the regularizer under three different scaling regions of ρ and

η.

We impose the following smoothness assumption on the loss f, which is a standard technique assumption when investigating the regularization effects of Wasserstein DRO [123].

Assumption 9 (Smooth Loss)**.** *The loss $f(\cdot)$ is smooth with respect to the norm $\|\cdot\|$ such that $\|\nabla f(x) - \nabla f(x')\|_* \leq S(x) \cdot \|x - x'\|, \quad \forall x, x'$.*

We now present our main result in this section in Theorem 26, whose proof is provided in Appendix E.4.

Theorem 26 (Regularization Effects)**.** *Suppose Assumption 9 holds, and $\rho \to 0, \eta \to 0$, we have the following results.*

(I) *(OCE Regularization) When $\rho/\eta \to C \in (0, \infty)$, it holds that*

$$\left|\mathcal{E}_{\widehat{P}}(f; \rho, \eta) - \mathcal{R}_1(f; \rho, \eta)\right| = o(\rho).$$

(II) *(Variation Regularization) When $\rho/\eta \to \infty$ and suppose additionally Assumption 6 holds, then*

$$\left|\mathcal{E}_{\widehat{P}}(f; \rho, \eta) - \mathcal{R}_2(f; \rho, \eta)\right| = o(\rho).$$

(III) *(Variance Regularization) When $\rho/\eta \to 0$ and suppose additionally that $\phi(t)$ is two times continuously differentiable in a neighborhood of $t = 1$ with $\phi''(1) > 0$, then*

$$\left|\mathcal{E}_{\widehat{P}}(f; \rho, \eta) - \mathcal{R}_3(f; \rho, \eta)\right| = o(\rho).$$

The proof idea is to consider the surrogate of $\mathcal{E}_{\widehat{P}}(f; \rho, \eta)$, by replacing f with its first-order Taylor expansion, which leads to

$$\widetilde{\mathcal{E}}_{\widehat{P}}(f; \rho, \eta) = \rho \cdot \mathbb{E}_{z \sim \widehat{P}}\left[\inf_{\mu \in \mathbb{R}}\left\{\mu + \frac{1}{\rho/\eta}\mathbb{E}_{b \sim \beta}\left[\phi^*\Big(\frac{\rho}{\eta} \cdot (\nabla f(z)^{\mathrm{T}} b - \mu)\Big)\right]\right\}\right]. \tag{6.21}$$

Based on Assumption 9, it can be shown that $\mathcal{E}_{\widehat{P}}(f;\rho,\eta) = \widetilde{\mathcal{E}}_{\widehat{P}}(f;\rho,\eta) + O(\rho^2)$. Thus, it suffices to derive approximations of $\widetilde{\mathcal{E}}$ under different scaling regimes of ρ/η. Below, we provide interpretations on this main result for each scaling regime.

Case 1: $\rho/\eta \to C \in (0,\infty)$.

In this case, the perturbation budget ρ and regularization level η decay in the same order. It is noteworthy that the drived regularization $\mathcal{R}_1(f;\rho,\eta)$ in (6.20a) has close connection to the optimized certainty equivalent risk (OCE) measure studied in [26]: Define the OCE of random variable X with parameter η as

$$\mathfrak{S}_\eta(X) = \eta \cdot \inf_{\mu\in\mathbb{R}} \Big\{\mu + \mathbb{E}\Big[\phi^*\Big(\frac{X}{\eta} - \mu\Big)\Big]\Big\} = \inf_{\mu\in\mathbb{R}} \Big\{\mu + \mathbb{E}\Big[\eta\phi^*\Big(\frac{X-\mu}{\eta}\Big)\Big]\Big\},$$

then $\mathcal{R}_1(f;\rho,\eta) = \rho \cdot \mathbb{E}_{z\sim\widehat{P}}\Big[\mathfrak{S}_{1/C}\Big((\nabla f(z))_{\#}\beta\Big)\Big]$, where $(\nabla f(z))_{\#}\beta$ is the pushforward probability measure of β by the linear projection map $\langle\cdot, \nabla f(z)\rangle$. Namely, the regularization term $\mathcal{R}_1(f;\rho,\eta)$ represents the averaged value of OCE across the projection of the loss gradient $\nabla f(z)$ with random projection directions.

Interestingly, it can be shown that the regularization $\mathcal{R}_1(f;\rho,\eta)$ converges to $\mathcal{R}_2(f;\rho,\eta)$ as the constant $C \to \infty$, and converges to $\mathcal{R}_3(f;\rho,\eta)$ as $C \to 0$. When $\rho/\eta \to C$ for some $C > 0$, the corresponding regularized ERM is an interpolation between the regularized ERM formulations corresponding to other two extreme cases.

Case 2: $\rho/\eta \to \infty$.

In this case, the convergence rate of the regularization level η is faster than that of the perturbation budget ρ. We showed that (6.19) is asymptotically equivalent to the gradient norm regularization in (6.20b). Recall that Gao et al. [123] showed the standard ∞-Wasserstein DRO can be approximated using the same regularization term, and the authors therein call it the *variation regularization*. Therefore, our finding in this case matches our intuition since

the regularization level η has little impact on the regularization effect.

Case 3: $\rho/\eta \to 0$.

Finally, we consider the case where the convergence rate of ρ is faster than that of the regularization η. We showed that (6.19) is asymptotically equivalent to the variance regularization (6.20c) in terms of the projected gradient of the loss f. Note that the regularization effect for this case requires the assumption that the divergence function $\phi(t)$ should be two times continuously differentiable and locally strongly convex around $t = 1$, which is a common condition in the study of general ϕ-divergence DRO [48, 107, 107, 186].

Recall that [48, 107] showed the ϕ-divergence DRO with a sufficiently small size of ambiguity set can be well-approximated by the ERM with variance regularization in terms of the loss. By taking the first-order Taylor expansion regarding the loss, these regularizations relate to each other. An intuitive explanation is that the impact of regularization level dominates in this case, which corresponds to the regime where the worst-case transport mapping (Primal-ϕ-Reg) is sufficiently close to the reference mapping. This indeed corresponds to the case studied in the aforementioned reference.

6.5 Generalization Error Bound

In this section, we investigate the generalization properties of our proposed adversarial learning framework in (Reg-∞-WDRO). To simplify our analysis, we focus on the *multi-class* classification setup, i.e., the loss function $f_\theta(z)$ is defined as $f_\theta(z) = \ell\big(g_\theta(x), y\big)$, where the data point $z = (x, y)$ represents the feature-label pair, $g_\theta(x)$ is the predictor function parameterized by θ, and $\ell : \mathbb{R}^K \times \{1, \ldots, K\} \to [0, 1]$ denotes a K-class classification loss function such as the cross-entropy loss. Throughout this section, we take the norm function that appears in the ∞-Wasserstein metric as

$$\|z - z'\| = \|x - x'\|_\infty + \infty \cdot \mathbf{1}\{y \neq y'\},$$

where $z = (x, y)$ and $z' = (x', y')$ are two different data points.Thus, we take into account only the distribution shift of the feature vector and omit the label distribution shift. Let $S = \{z_i\}_{i=1}^n$ denote the set of n i.i.d. sample points generated from $\mathbb{P}_{\text{true}}$, and $\mathbb{P}_n$ be the empirical distribution supported on S. For fixed parameter θ, let $\widehat{R}_{\text{adv}}(\theta)$ be the objective value of (Reg-∞-WDRO) with $\widehat{P} = \mathbb{P}_n$, and $R_{\text{adv}}(\theta)$ be its population version, i.e.,

$$\begin{aligned} R_{\text{adv}}(\theta) &= \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{(x,y)\sim\mathbb{P}} \left[\ell\big(g_\theta(x), y\big)\right] - \eta \mathbb{D}_\phi(\gamma, \gamma_0) : \begin{array}{l} \text{Proj}_{1\#}\gamma = \mathbb{P}_{\text{true}}, \text{Proj}_{2\#}\gamma = \mathbb{P} \\ \text{ess.sup}_\gamma \|\zeta_1 - \zeta_2\| \le \rho \end{array} \right\} \\ &= \mathbb{E}_{(x,y)\sim\mathbb{P}_{\text{true}}} \left[\inf_{\mu\in\mathbb{R}} \left\{ \mu + \mathbb{E}_{b\sim\beta} \left[(\eta\phi)^* \left(\ell\big(g_\theta(x+b), y\big) - \mu \right) \right] \right\} \right], \end{aligned}$$

where the last equality is based on the strong duality result in Theorem 23, and $b \sim \beta$ is a random vector uniformly distributed on $\mathbb{B}_\rho(0)$, the $\|\cdot\|_\infty$-ball of radius ρ centered at the origin. One of the most important research questions in learning theory is to provide the gap between the *empirical regularized adversarial risk* $\widehat{R}_{\text{adv}}(\theta)$ and the *population regularized adversarial risk* $R_{\text{adv}}(\theta)$ (see, e.g., [10, 11, 330]). We answer this question leveraging the covering number argument.

Let us begin with some technical preparation. Let $\epsilon > 0$ and $(\mathcal{V}, \|\cdot\|)$ be a normed space. We say $\mathcal{C} \subseteq \mathcal{V}$ is an ϵ-cover of $\mathcal{V}$ if for any $V \in \mathcal{V}$, there exists $V' \in \mathcal{C}$ such that $\|V - V'\| \le \epsilon$. The least cardinality of $\mathcal{C}$ is called the ϵ-covering number, denoted as $\mathcal{N}(\mathcal{V}, \epsilon, \|\cdot\|)$. For any $x, x' \in \mathcal{V}^n$, we take the norm $\|x - x'\| = \max_{i\in[n]} \|x_i - x_i'\|$. Define the regularized adversarial function class

$$\mathcal{G}_{\text{adv}} = \left\{ (x,y) \mapsto \inf_{\mu\in\mathbb{R}} \left\{ \mu + \mathbb{E}_{b\sim\beta} \left[(\eta\phi)^* \left(\ell\big(g_\theta(x+b), y\big) - \mu \right) \right] \right\} : \quad \theta \in \Theta \right\}.$$

For dataset $S = \{z_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n$, define

$$\mathcal{G}_{\text{adv}|_S} = \Big\{ (g(x_1, y_1), \ldots, g(x_n, y_n)) : \ g \in \mathcal{G}_{\text{adv}} \Big\}.$$

An intermediate consequence of the covering number argument is the following.

Proposition 9 ([276, 330])**.** *Suppose the range of the loss function* $(x, y) \mapsto \ell(g_\theta(x), y)$ *is* $[0, 1]$*. With probability at least* $1 - \delta$ *with respect to* $\widehat{P} = \mathbb{P}_n$*, it holds that for all* $\theta \in \Theta$*,*

$$R_{adv}(\theta) \leq \widehat{R}_{adv}(\theta) + \inf_{\alpha > 0} \left(8\alpha + \frac{24}{\sqrt{n}} \int_\alpha^1 \sqrt{\log \mathcal{N}(\mathcal{G}_{adv|_S}, \epsilon, |\cdot|)}\, \mathrm{d}\epsilon \right) + 3\sqrt{\frac{\log(2/\delta)}{n}}. \tag{6.22}$$

The remaining challenge is to provide the upper bound on the covering number $\mathcal{N}(\mathcal{G}_{\text{adv}|_S}, \epsilon, |\cdot|)$. Define the new function class of interest as

$$\mathcal{G} = \Big\{ (x, y) \mapsto \ell(g_\theta(x+\cdot), y) \in \mathbb{R}^{\mathbb{B}_\rho(0)} : \theta \in \Theta \Big\}, \qquad \mathcal{G}_{|_S} = \Big\{ (g(x_1, y_1), \ldots, g(x_n, y_n)) : g \in \mathcal{G} \Big\}.$$

and the associated norm $\|g\|_\infty = \sup_{b \in \mathbb{B}_\rho(0)} |g(b)|, \forall g \in \mathcal{G}$. The following proposition controls the covering number of $\mathcal{G}_{\text{adv}|_S}$ using that of $\mathcal{G}_{|S}$.

Proposition 10. *Assume that* ϕ *is strictly convex. Then, it holds that* $\mathcal{N}(\mathcal{G}_{adv|_S}, \epsilon, |\cdot|) \leq \mathcal{N}(\mathcal{G}_{|_S}, \epsilon, \|\cdot\|_\infty)$.

Proposition 10 gives an estimate of $\mathcal{N}(\mathcal{G}_{\text{adv}|_S}, \epsilon, |\cdot|)$ by taking $\mathcal{G}_{|_S}$ that involves the perturbation set $\mathbb{B}_\rho(0)$ into account. The advantage is that the covering number $\mathcal{N}(\mathcal{G}_{|_S}, \epsilon, \|\cdot\|_\infty)$ can be easily computed. The following presents several applications of Propositions 9 and 10.

Example 10 (Linear Function Class)**.** *Let us model the predictor* $g_\theta(x)$ *using the linear function class*

$$g_\theta(x) = Wx, \quad \theta \in \Theta := \Big\{ W \in \mathbb{R}^{K \times d}, \quad \|W\|_{1,\infty} \leq \Lambda_1, \|W\|_{2,2} \leq \Lambda \Big\},$$

where the feature vector $x \in \mathbb{R}^d$ *is assumed to be bounded:* $\sup_x \|x\|_2 \leq \Psi$*. Next, we*

consider the ramp loss $\ell:\ \mathbb{R}^K \times \{1,\dots,K\} \to \mathbb{R}_+$

$$\ell(t,y) = \begin{cases} 1, & \text{if } M(t,y) \le 0, \\ 1 - \dfrac{1}{\varrho} M(t,y), & \text{if } 0 < M(t,y) < \varrho, \\ 0, & \text{if } M(t,y) \ge \varrho, \end{cases} \tag{6.23}$$

where $M(t,y) = t_y - \max_{y' \neq y} t_{y'}$. *By [229, Lemmas 4.4 and 5.2], it holds that* $\log \mathcal{N}(\mathcal{G}_{|_S}, \epsilon, \|\cdot\|_\infty) = \widetilde{O}(\frac{\Lambda^2 d(\Psi+\sqrt{d}\rho)^2}{\epsilon^2 \varrho^2})$, *where* $\widetilde{O}(\cdot)$ *hides constant logarithmically dependent on related parameters. As a consequence, the generalization bound* (6.22) *further simplifies to*

$$R_{adv}(\theta) \le \widehat{R}_{adv}(\theta) + 3\sqrt{\frac{\log(2/\delta)}{n}} + \frac{8}{n} + \widetilde{O}\left(\frac{\Lambda\sqrt{d}(\Psi+\sqrt{d}\rho)}{\varrho\sqrt{n}}\right).$$

Example 11 (Neural Network Function Class)**.** *We next consider the predictor* $g_\theta(x)$ *belongs to the* L*-layer and* m*-width neural network function class with* 1*-Lipschitz nonlinear activation* σ*:*

$$f(x) = W_L \cdot \sigma\Big(W_{L-1} \cdots \sigma(W_1 x)\Big),$$

where the feature vector $x \in \mathbb{R}^d$ *satisfies* $\sup_x \|x\|_2 \le \Psi$*, and the model parameter*

$$\theta \in \Theta := \Big\{(W_1,\dots,W_L):\ \|W^l\|_2 \le a_l, \|W^l\|_{sp} \le s_l, l = 1,\dots,L-1 \\ \|W^L\|_2 \le a_L, \|W^L\|_{2,\infty} \le s_L, \|W^1\|_{1,\infty} \le s_1'\Big\}.$$

When considering the ramp loss in (6.23)*, according to [229, Lemma 5.14],*

$$\log \mathcal{N}(\mathcal{G}_{|_S}, \epsilon, \|\cdot\|_\infty) = \widetilde{O}\left(\frac{L^2 d(\Psi+\sqrt{d}\rho)^2}{\varrho^2 \epsilon} \cdot \prod_{l\in[L]} s_l^2 \cdot \sum_{l\in[L]} \frac{a_l^2}{s_l^2}\right).$$

Then, the generalization bound (6.22) *becomes*

$$R_{adv}(\theta) \leq \widehat{R}_{adv}(\theta) + 3\sqrt{\frac{\log(2/\delta)}{n}} + \frac{8}{n} + \widetilde{O}\left(\frac{L\sqrt{d}(\Psi + \sqrt{d}\rho)}{\varrho\sqrt{n}} \cdot \prod_{l\in[L]} s_l \cdot \sqrt{\sum_{l\in[L]} \frac{a_l^2}{s_l^2}}\right).$$

Remark 28. *Compared to the generalization analysis of the unregularized adversarial robust learning formulation in* (6.1) *(see, e.g., [11, 174, 321, 330]), our error bound for the regularized case is similar. It aligns with the state-of-the-art error bound for the unregularized case. The novelty in our theoretical analysis lies in upper bounding the covering number of $\mathcal{G}_{adv|S}$ in Proposition 9 by that of $\mathcal{G}_{|S}$. Based on the lower bound of Rademacher complexity for neural network function classes [20, Theorem 3.4], our generalization bound for the linear function class matches the lower bound in terms of the parameters Ψ, Λ, and n, but introduces an additional $O(d)$ term. For neural network function classes, our bound introduces an additional $O(Ld \cdot \sqrt{\sum_{l\in[L]} \frac{a_l^2}{s_l^2}})$ term. While these additional terms are relatively small, developing new proof techniques to further tighten the generalization analysis would be desirable.*

6.6 Numerical Study

In this section, we examine the numerical performance of our proposed algorithm on three applications: supervised learning, reinforcement learning, and contextual learning. We compare our method with the following baselines: (i) empirical risk minimization (ERM), (ii) fast-gradient method (FGM) [137], (iii) and its iterated variant (IFGM) [185]. Those baseline methods are heuristic approaches to approximately solving the ∞-WDRO model.

6.6.1 Supervised Learning

We validate our method on three real-world datasets: MNIST [187], Fashion-MNIST [320], and Kuzushiji-MNIST [81]. The experiment setup largely follows from Sinha et al. [287]. Specifically, we build the classifer using a neural network with 8×8, 6×6, and 5×5

convolutional filter layers and ELU activations, and followed by a connected layer and softmax output. After the training process with those listed methods, we then add various perturbations to the testing datasets, such as the ℓ_2-norm and ℓ_∞-norm adversarial projected gradient method (PGM) attacks [212], and white noises uniformly distributed in a ℓ_2 or ℓ_∞ norm ball. We use the mis-classification rate on testing dataset to quantify the performance for the obtained classifers. For fair comparison, we take the same level of robustness parameter $\rho = 0.45$ for all approaches, and specify the number of epochs (i.e., the number of times the data points are passed through the model) as 30 and use Adam optimizer with stepsize $\gamma = 1e\text{-}3$. For PGM attack and FGM/IFGM defense, we specify the stepsize for the attack step as 0.1. The number of iterations for the attack step of PGM attack and IFGM defense is set to be 15. Since the scaling of η that satisfies $\rho/\eta \to C$ for some constant C corresponds to an interpolation of gradient norm and gradient variance regularized ERM training as suggested by Section 6.4, we specify the regularization value $\eta = 2 \cdot \rho$ in this experiment. Since the loss is highly nonconvex, we examine our regularized adversarial training using entropic regularization only, with RT-MLMC gradient estimator and the maximum level $L = 7$.

Figure 6.3 presents the mis-classification results of various methods. Specifically, three rows correspond to different kinds of datasets (MNIST, Fashion-MNIST, and Kuzushiji-MNIST), and four columns correspond to different types of perturbations (ℓ_2/ℓ_∞ adversarial attack, and ℓ_2/ℓ_∞ white noise attack). For every single plot, the x-axis corresponds to the magnitude of perturbation (ϵ_{white} or ϵ_{adv}) normalized by the average of the norm overall feature vectors (in terms of 2- or ∞-norm, denoted as C_2 or C_∞, respectively). From these plots, we find that all methods tend to have worse performance as the perturbation level increases, but the regularized adversarial risk model consistently outperforms all baselines. Especially, it performs well when the perturbation levels are large. This suggests that our model has superior performance for adversarial training in scenarios with large perturbations.

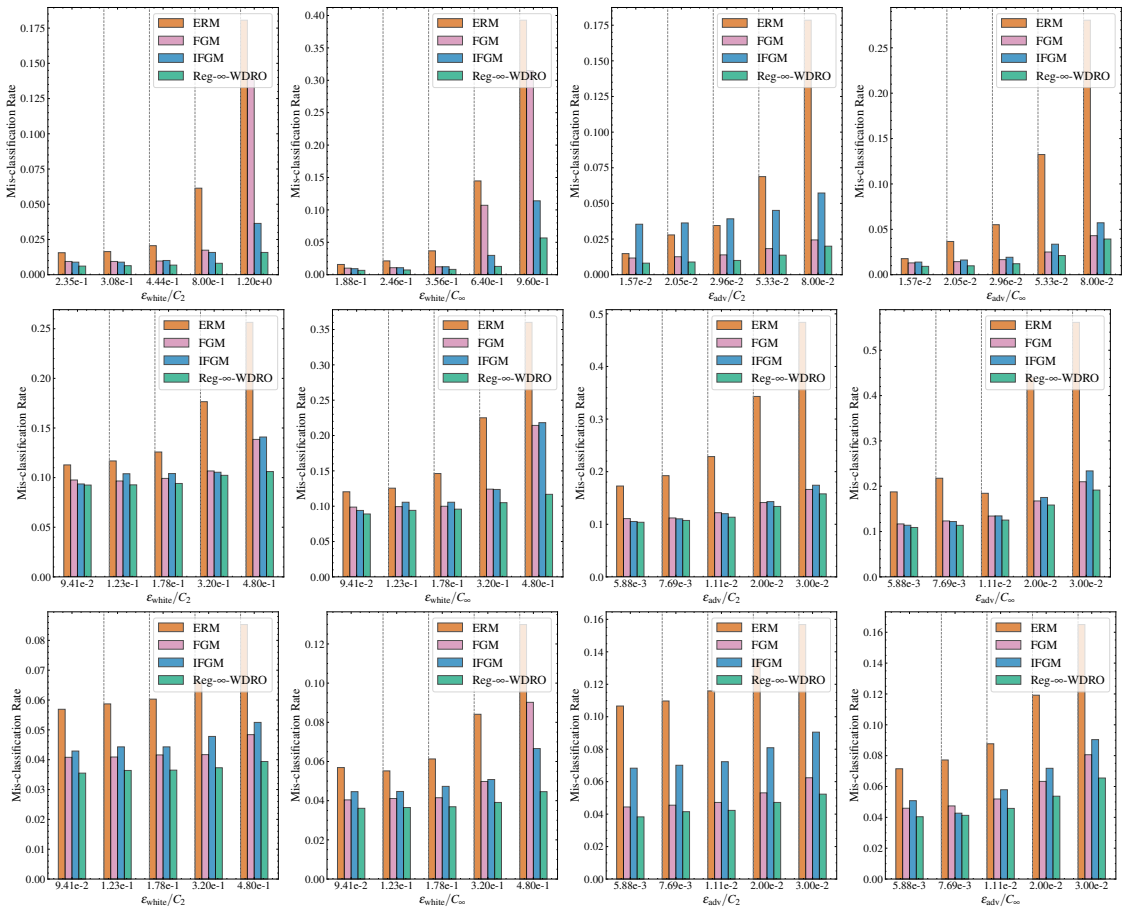

Figure 6.3: Results of adversarial training in terms of mis-classification rates. From top to bottom, the figures correspond to (a) MNIST; (b) Fashion-MNIST; (c) and Kuzushiji-MNIST datasets. From left to right, the figures correspond to (a) ℓ_2-norm white noise attack; (b) ℓ_∞-norm white noise attack; (c) ℓ_2-norm PGM attack; and (d) ℓ_∞-norm PGM attack.

6.6.2 Reinforcement Learning

Next, we provide a robust algorithm for reinforcement learning (RL). Consider an infinite-horizon discounted finite state MDP represented by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma \rangle$, where $\mathcal{S}, \mathcal{A}$ denotes the state and action space, respectively; $\mathbb{P} = \{\mathbb{P}(s' \mid s, a)\}_{s,s',a}$ is the set of transition probability metrics; $R = \{r(s,a)\}_{s,a}$ is the reward table with (s,a)-th entry being the reward for taking the action a at state s; and $\gamma \in (0,1)$ is the discounted factor. Similar to problem (∞-WDRO), robust reinforcement learning seeks to maximize the worst-case risk function $\sup_{\mathbb{P} \in \mathfrak{R}} \mathbb{E}[\sum_t \gamma^t r(s_t, a_t)]$, with $\mathfrak{R}$ represents the ambiguity set for state-action transitions. For simplicity, we consider a tabular Q-learning setup in this subsection. The

standard Q-learning algorithm in RL learns a Q-function $Q: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with iterations

$$Q(s^t, a^t) \leftarrow (1-\alpha_t)Q(s^t, a^t) + \alpha_t r(s^t, a^t) - \gamma\alpha_t \min_a(-Q(s^{t+1}, a)), \quad s^{t+1} \sim \mathbb{P}(\cdot \mid s^t, a^t). \tag{6.24}$$

We modify the last term of the update (6.24) with an adversarial state perturbation to take ∞-Wasserstein distributional robustness with entropic regularization into account, leading to the new update

$$Q(s^t, a^t) \leftarrow (1-\alpha_t)Q(s^t, a^t) + \alpha_t r(s^t, a^t) - \gamma\alpha_t \min_a \left\{\eta \log \mathbb{E}_{\widehat{s}^{t+1}\sim\beta(s^{t+1};\rho)} e^{-Q(\widehat{s}^{t+1}, a)/\eta}\right\},$$

Table 6.3: Performance of Q-learning algorithms in original MDP and shifted MDP environments. Error bars are produced using 10 independent trials.

Environment	Regular	Robust
Original MDP	469.42 ± 19.03	$\mathbf{487.11 \pm 9.09}$
Perturbed MDP (Heavy)	187.63 ± 29.40	$\mathbf{394.12 \pm 12.01}$
Perturbed MDP (Short)	355.54 ± 28.89	$\mathbf{443.17 \pm 9.98}$
Perturbed MDP (Strong g)	271.41 ± 20.7	$\mathbf{418.42 \pm 13.64}$

where $\beta(s^{t+1}; \rho)$ denotes an uniform distribution supported on a $\|\cdot\|_\infty$-norm ball of s^{t+1} with radius ρ. Standard fixed point iteration analysis [292, 312, 327] can be modified to show the convergence of the modified Q-learning iteration. Our proposed algorithm in Section 6.3 can be naturally applied to proceed the updated Q-learning iteration.

We test our algorithm in the cart-pole environment [59], where the objective is to balance a pole on a cart by moving the cart to left or right, with state space including the physical parameters such as chart position, chart velocity, angle of pole rotation, and angular of pole velocity. To generate perturbed MDP environments, we perturb the physical parameters of the system by magnifying the pole's mass by 2, or shrinking the pole length by 2, or magnifying the strength of gravity g by 5. We name those three perturbed environments as *Heavy*, *Short*, or *Strong* g MDP environments, respectively.

Figure 3 demonstrates the training process of regular and robust Q-learning algorithms

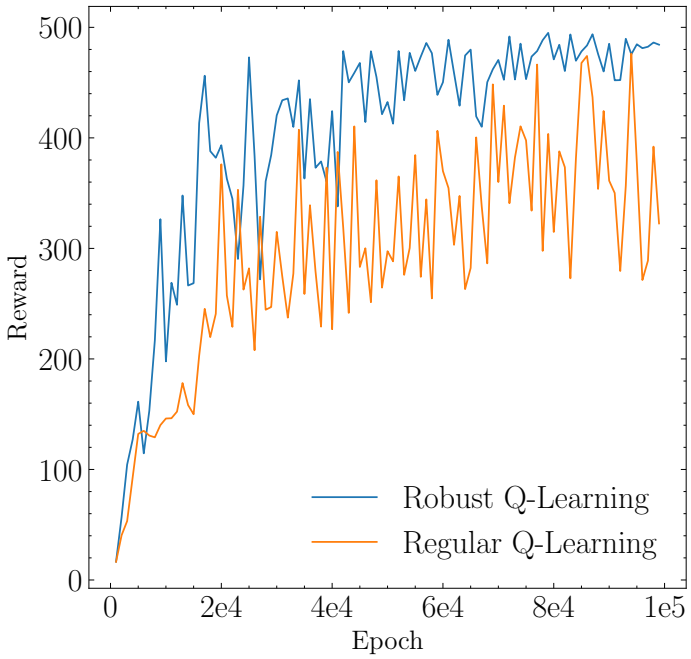

Figure 6.4: Episode lengths during training. The environment caps episodes to 400 steps.

on the original MDP environment. Interestingly, the robust Q-learning algorithm learns the optimal policy more efficiently than the regular MDP. One possible explanation is that taking account into adversarial perturbations increase the exploration ability of the learning algorithm. Next, we report the performance of trained policies in original and perturbed MDP environments in Table 2, from which we can see that our proposed robust Q-learning algorithm consistently outperforms the regular non-robust algorithm.

6.6.3 Contextual Learning

Contextual stochastic optimization (CSO) seeks the optimal decision to minimize the cost function Ψ involving random parameters Z, whose distribution is affected by a vector of relevant covariates denoted as X. Since one has access to covariates X before decision making, we parameterize the optimal decision using $f_\theta(\cdot)$ that maps from X to the final decision. This paradigm, inspired by the seminar work [31], has achieved phenomenal success in operations research applications. See the survey [262] that summarizes its recent developments.

Distributionally robust CSO with ∞-type casual optimal transport distance has gained great popularity in recent literature [121, 329]. It seeks the optimal decision parameter θ to minimize the worst-case risk, where the worst-case means we simultaneously find the casual optimal transport γ that maps $\widehat{P}$, the empirical distribution from available data $\{(x_i, z_i)\}_i$, to $\mathbb{P}$ up to certain transportation budget, Its strong dual reformulation can be reformulated as a special case of (∞-WDRO):

$$\min_{\theta} \left\{ \mathbb{E}_{\widehat{x}\sim\widehat{P}_{\widehat{X}}} \left[\sup_{x'\in\mathbb{B}_\rho(\widehat{x})} \mathbb{E}_{\widehat{z}\sim\widehat{P}_{\widehat{Z}|\widehat{X}=\widehat{x}}} [\Psi(f_\theta(x'), \widehat{z})] \right] \right\}. \tag{6.25}$$

Similar to adversarial robust learning, Problem (6.25) can be challenging to solve because computing the optimal value of the inner maximization problem is usually NP-hard. Instead, we replace the inner maximization with OCE risk, leading to the approximation problem

$$\min_{\theta} \left\{ \mathbb{E}_{\widehat{x}\sim\widehat{P}_{\widehat{X}}} \left[\inf_{\mu\in\mathbb{R}} \left\{ \mu + \mathbb{E}_{x'\sim\nu_{\widehat{x}}} \Big[(\eta\phi)^*(\mathbb{E}_{\widehat{z}\sim\widehat{P}_{\widehat{Z}|\widehat{X}=\widehat{x}}} [\Psi(f_\theta(x'), \widehat{z})] - \mu) \Big] \right\} \right] \right\}, \tag{6.26}$$

where $\nu_{\widehat{x}}$ denotes the uniform distribution supported on $\mathbb{B}_\rho(\widehat{x})$. Alternatively, we can express (6.26) as a special case of (Reg-∞-WDRO). See the detailed discussion in Appendix E.6.

In the following, we test our algorithm in the application of data-driven personalized pricing problem, using the similar setup in [329, Example 6]: Let $w \in \mathbb{R}$ denote price, $x \in \mathbb{R}^{10}$ denote side information, and $z \in \mathbb{R}^2$ denote price sensitivity coefficient that describes how price influences customer demand. The loss $\Psi(w, z) = -wz^{\mathrm{T}} \begin{pmatrix} w \\ 1 \end{pmatrix}$, denoting the negative revenue under price w and coefficient z. Assume z depends on x in a nonlinear way:

$$z = \begin{pmatrix} \tanh(3\beta_1^{\mathrm{T}} x) \\ \exp(-2\beta_2^{\mathrm{T}} x) \end{pmatrix} + \mathcal{N}(0, \mathbf{I}_2),$$

where $\beta_1, \beta_2 \sim \mathcal{U}([-0.1, 0.1]^{10})$ and $x \sim \mathcal{N}(0, \mathbf{I}_{10})$. We solve this problem using the linear decision rule approach, by taking $f_\theta(x) = \theta^{\mathrm{T}} g(x)$, where $g : \mathbb{R}^{10} \to \mathbb{R}^{100}$ is a random

feature model:

$$g(x) = \left(\cos(\omega_i^{\mathrm{T}} x + b_i)\right)_{i\in[100]}, \quad \omega_i \sim \mathcal{N}(0, \mathbf{I}_{10}), b \sim \mathcal{U}([0, 2\pi]).$$

Throughout the experiment, we take hyper-parameters $\rho = 0.45$ and $\eta = 0.9$. When creating training dataset, we generate $M \in \{25, 50, 100, 200\}$ samples of x, denoted as $\{x_i\}_{i\in[M]}$, and for each x_i, we generate $m \in \{10, 30, 50, 100, 200\}$ samples of z from the conditional distribution of z given $x = x_i$.

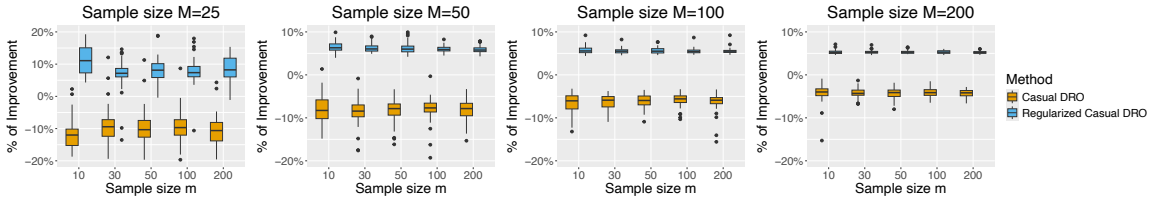

Figure 6.5: Results of ∞-type Casual DRO and its regularized version in terms of percentage of improvements. From left to right, the figures correspond to $M = 25, 50, 100, 200$, respectively.

We quantify the performance of a given decision θ using the percentage of improvements (compared to ERM) measure:

$$\mathcal{J}(\theta) = 1 - \frac{\mathcal{R}(\theta) - \mathcal{R}^*}{\mathcal{R}(\theta_{\mathrm{ERM}}) - \mathcal{R}^*},$$

where $\mathcal{R}^*$ denotes the ground truth optimal revenue provided the distribution of (x, z) is exactly known, θ_{ERM} denotes the decision obtained from the ERM, the non-robust training approach, and $\mathcal{R}(\theta)$ denotes the expected revenue of the decision θ under the ground truth distribution. The plots in Figure 6.5 report the percentage of improvements obtained either by solving the standard casual CSO problem (6.25) using the heuristic FGM method or its (KL-divergence-)regularized formulation (6.26). The error bars are reproduced using 50 independent trials. For all scenarios, we can see the regularized robust CSO model outperforms the un-regularized one. Besides, the un-regularized DRO model has negative improvements in general, mainly due to the computational intractability of Problem (6.25).

6.7 Conclusion

In this paper, we proposed a ϕ-divergence regularized framework for adversarial robust training. From the computational perspective, this new formulation is easier to solve compared with the original one. From the statistical perspective, this framework is asymptotically equivalent to certain regularized ERM under different scaling regimes of the regularization and robustness hyper-parameters. From the generalization perspective, we derived the population regularized adversarial risk is upper bounded by the empirical one up to small residual error. Numerical experiments indicate that our proposed framework achieves state-of-the-art performance, in the applications of supervised learning, reinforcement learning, and contextual learning.

CHAPTER 7
CONCLUDING REMARKS

In this thesis, we leverage advanced statistical and optimization techniques to tackle two decision-making under uncertainty problems: hypothesis testing and distributionally robust stochastic optimization (DRO). Our proposed framework not only provides decisions with satisfactory out-of-sample performance but also aligns with key ethical and societal principles, including robustness and interpretability. We develop efficient algorithms with solid computational guarantees and favorable statistical properties, and validate their superior performance in many real-world applications.

7.1 Future Directions

In addition to the future directions outlined in previous chapters, we discuss additional topics that merit further investigation.

Robust Hypothesis Testing. Existing work typically assumes the input data for hypothesis testing follows the independent and identically distributed probability distribution, which may not always hold in reality. There have also been several trials to develop robust methods to tackle this issue for hypothesis two-sample testing [126, 164, 214, 309, 314]. I aim to incorporate DRO to design more effective algorithms for large-scale, high-dimensional, and noisy data. I am also interested in providing robust approaches for other hypothesis testing problems such as independence testing and multi-class hypothesis testing.

DRO with Nonconvex Objectives. Traditional DRO framework typically assumes the training objective is **convex** to obtain global convergence guarantees. I plan to develop global optimization algorithms for solving DRO with nonconvex objectives. On the one hand, it is promising to leverage mixed-integer optimization techniques to solve those problems up to moderate problem size. On the other hand, the preliminary understanding

of the landscape of special nonconvex problems with hidden convexity suggests that I can develop customized first-order algorithms to achieve global convergence. I foresee these nonconvex problems finding important applications in inventory control [70], pricing, and reinforcement learning [71].

Deep Learning Applications. Deep learning has ushered in a new era of data science. Its applications in hypothesis testing and DRO areas have achieved some success [75, 303, 326], as neural networks are powerful in function representation and are adaptive to the structure of data. I aim to use the state-of-the-art tools of deep learning, including generative/large-language models, to enhance the performance of decision-making, problem-solving, and analytics.

Appendices

APPENDIX A

PROOFS AND ADDITIONAL DETAILS OF CHAPTER 2

A.1 Preliminary Technical Results

Theorem 27 (Pinsker's Inequality [89])**.** *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 11 (Lipschitz Properties of Retraction Operator [56])**.** *There exists constants* L_1, L_2 *such that the following inequalities hold:*

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

Inspired from Appendix A.3 in [169], we are able to compute the constants in Proposition 11 explicitly: $L_1 = 1$ and $L_2 = \frac{1}{2}$. The proof is provided below.

Proof. By definition, we have that

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

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

Secondly, we can see that

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

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 $\|\mathrm{Retr}_s(\zeta) - (s+\zeta)\|_2 \leq \frac{1}{2}\|\zeta\|^2$. □

Theorem 28 (McDiarmid's Inequality [217])**.** *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 6 (Equivalent Definition for Sub-Gaussian variables (Lemma 2.3.2 in [134]))**.**

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 29 (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* $X \sim \mu^n$*:*

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

A.2 Introduction to Manifold Optimization

A brief introduction to manifold optimization can be found in [152]. 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 $\text{Grad}f(x)$, 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:

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

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

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

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

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

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 [110] and [316] 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 [56] investigated the Riemannian gradient methods that are guaranteed to converge into stationary points globally, the key proof technique of which relies on Proposition 11. 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

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

where the last equality is because that $(K_\mu - K_\nu)a$ is a zero function for any a satisfying $\|a\|_2 \le 1$. Hence, $\left\|\mathbb{E}_\mu[f(x)] - \mathbb{E}_\nu[f(y)]\right\|_2 < 2\varepsilon$ for any $\varepsilon > 0$ and $f \in \mathcal{C}_b(\mathcal{X})$. Then we conclude that the distribution $\mu = \nu$. □

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:

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

The Lagrangian function becomes

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

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 [258]. Moreover, the dual optimal value equals the primal optimal value because the Slater's condition [58] 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 29. *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

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

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 30. *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 [163]. 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 7 (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 $\lambda \in [0, 1]$:*

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

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 7. 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

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

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

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

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 8 (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 8. Note that

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

where the second inequality is by applying Lemma 7, and the last inequality is by applying Proposition 11. Moreover, we have that

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

Combining those inequalities above implies that

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

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

Lemma 9 (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 $k \geq 1$:*

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

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 9. 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 27 implies that

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

□

Lemma 10 (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 $t \ge 1$:*

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

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 10. 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 11. *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 11. The condition $\|\xi^{t+1}\| \leq \epsilon_1$ directly implies that

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

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 12 (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 12. 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 8,

Lemma 9, and Lemma 10:

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

Then we have that

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

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 29 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

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

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

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

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

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

Applying the ϵ-net argument similar to the Dudley's entropy integral bound [300, Theorem 5.22] gives

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

Taking $\mathcal{N}(\mathcal{F}, \mathsf{d}, \epsilon) = \lceil \frac{1}{\epsilon} \rceil$ 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 29 implies that

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

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

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

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

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

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

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

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

The upper bound for $(\mathcal{KPW}(\hat{\nu}_m,\nu))^{1/p}$ 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

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

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 13. *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 13. 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 14. *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 14. Following the similar argument as in Lemma 1, we can see that

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

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 28 implies

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

Applying Lemma 6 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 [300, 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 $\mathcal{N}(\mathcal{F}, \mathsf{d}, \epsilon) = \lceil \frac{1}{\epsilon} \rceil$ 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 15. *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 15. 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 28 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 14 and Lemma 15, 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:

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

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

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

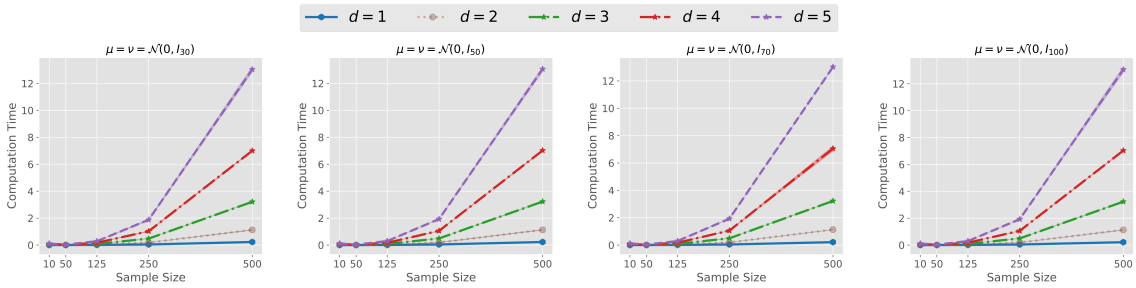

Figure A.1: Mean computation time for computing $\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_n)$ for varying n. Results are averaged over 10 independent trials.

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.

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

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

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

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

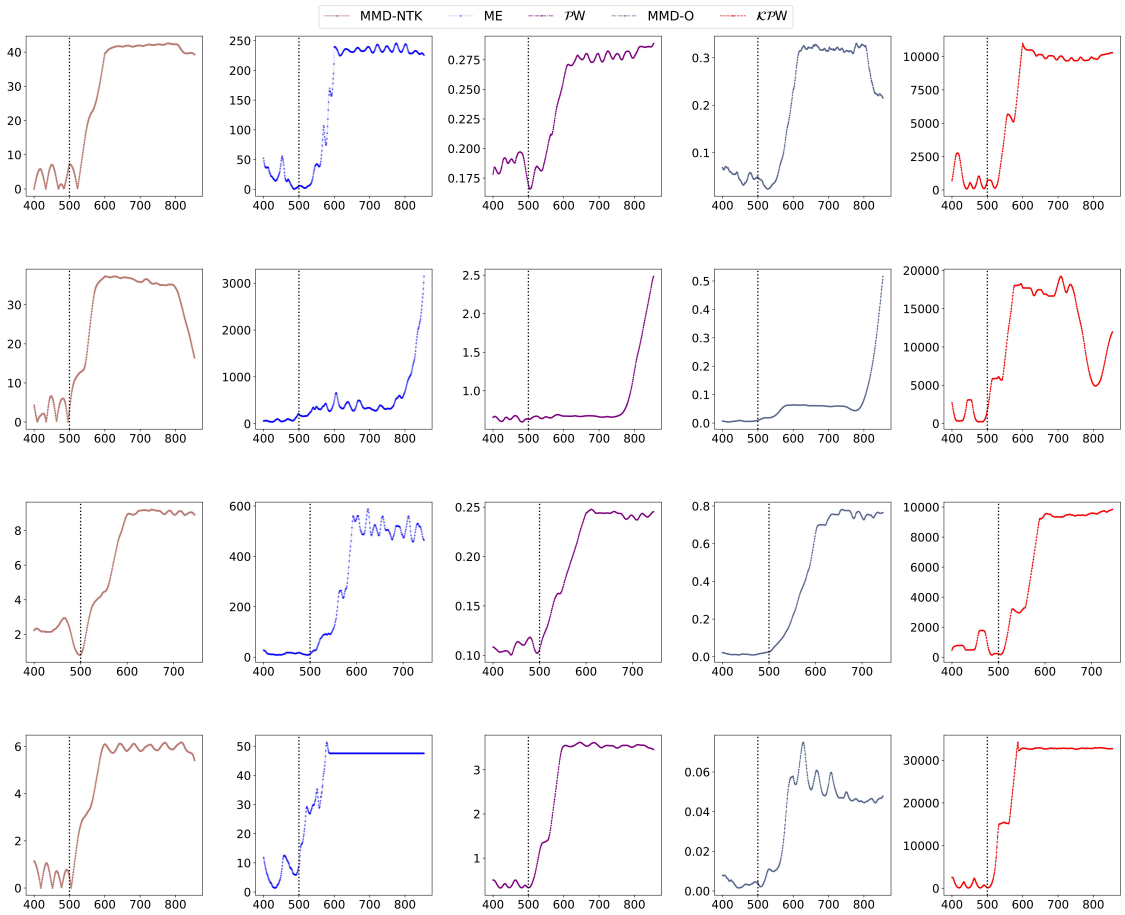

Figure A.2: Comparison of detection statistics from bending to throwing for various testing procedures. Black dash line indicates the true change-point. Each row corresponds to detection results for each user.

A.8.2 Impact of Entropic Regularization Parameter η

As pointed out in [128], 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 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

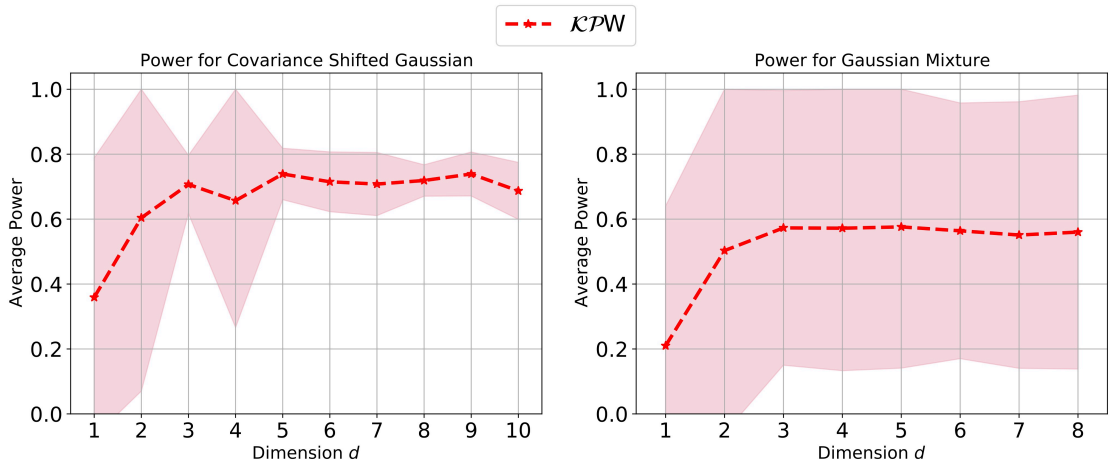

Figure A.3: Average power for KPW test across different choices of projected dimension d. Left: Gaussian distribution; Right: Gaussian mixture distribution. Results are averaged over 10 independent trials.

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.

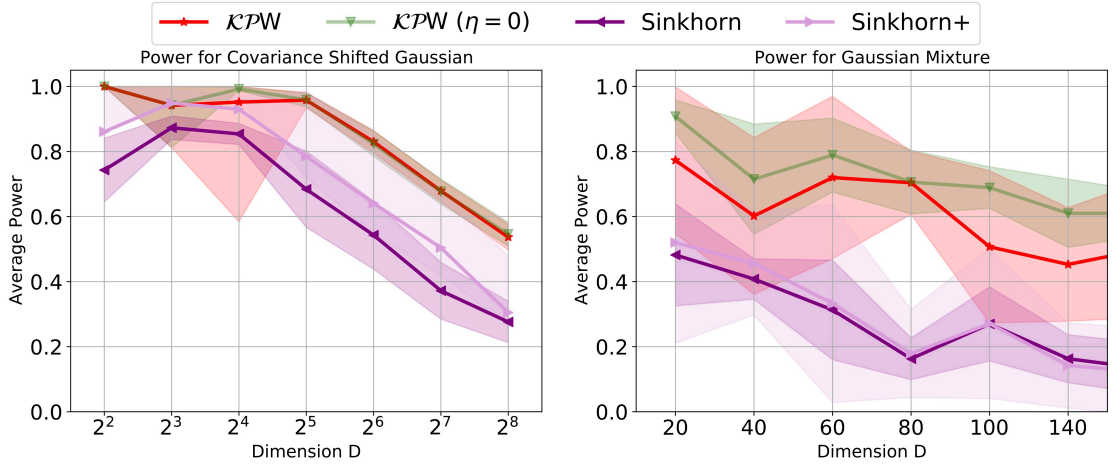

Figure A.4: Average power for KPW tests and Sinkhorn tests across different choices of data dimension D. Left: Gaussian distribution; Right: Gaussian mixture distribution. Results are averaged over 10 independent trials.

APPENDIX B

PROOFS AND ADDITIONAL DETAILS OF CHAPTER 3

B.1 Comparison with Optimal Transport Divergences

Table B.1: Time and sample complexity of empirical OT estimators in terms of the number of samples n. Here $\widehat{\mu}_n, \widehat{\nu}_n$ represent two empirical distributions based on i.i.d. n sample points in $\mathbb{R}^d$. p denotes the order of the metric defined in the cost function of standard Wasserstein distance.

Reference	Estimator	Name	Time Complexity	Sample Complexity
[116]	$W_p(\widehat{\mu}_n, \widehat{\nu}_n)$	Wasserstein distance	$\mathcal{O}(n^3 \log n)$	$\mathcal{O}(n^{-1/d})$
[128]	$\mathcal{S}_{p,\epsilon}(\widehat{\mu}_n, \widehat{\nu}_n)$	Sinkhorn Divergence	$\mathcal{O}(n^2)$ per iteration	$\mathcal{O}(n^{-1/2}(1+\epsilon^{d/2}))$
[239]	$\mathcal{MS}_k(\widehat{\mu}_n, \widehat{\nu}_n)$	Max Sliced Wasserstein distance with k-dimensional projector	$\tilde{\mathcal{O}}(n^2 d)$	$\mathcal{O}(n^{-1/(\max\{k,2\})})$
[230]	$\mathcal{SW}_p(\widehat{\mu}_n, \widehat{\nu}_n)$	Sliced Wasserstein distance	$\tilde{\mathcal{O}}(nd)$	$\mathcal{O}(n^{-1/2})$
[307] and this work	$\mathcal{KMS}_p(\widehat{\mu}_n, \widehat{\nu}_n)$	Kernel Max Sliced Wasserstein Distance	$\tilde{\mathcal{O}}(n^2 d^3)$	$\mathcal{O}(n^{-1/2})$

In Table B.1, we summarize the time and sample complexity of various OT divergences. Notably, the sample complexity of the standard Wasserstein distance suffers from the curse of dimensionality. While the Sinkhorn divergence mitigates this issue, its sample complexity depends on $\epsilon^{d/2}$, which can be prohibitively large when the regularization parameter ϵ is very small. For Max-Sliced (MS) and Sliced Wasserstein distances, their sample complexity is independent of the data dimension. However, they rely on linear projections to analyze data samples, which may not be optimal, particularly when the data exhibits a nonlinear low-dimensional structure. This limitation motivates the study of the KMS Wasserstein distance by Wang et al. [307] and in our work. In Example 1 and Section 6.6, we numerically validate this observation and demonstrate its practical advantages.

B.2 Proof of Theorem 6 and Corollary 1

The proof in this part relies on the following technical results.

Theorem 30. ***(Finite-Sample Guarantee on MS*** 1***-Wasserstein Distance on Hilbert Space, Adopted from [50, Corollary 2.8])*** *Let* $\delta \in (0,1]$*, and* μ *be a probability measure on a separable Hilbert space* $\mathcal{H}$ *with* $\int_{\mathcal{H}} \|x\| \,\mathrm{d}\mu(x) < \infty$*. Let* $X_1, \ldots, X_n$ *be i.i.d. random elements of* $\mathcal{H}$ *sampled according to* μ*, and* $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$*, then it holds that*

$$\mathbb{E}\mathcal{MS}_1\left(\mu, \widehat{\mu}_n\right) \leq C \cdot \left(\int_{\mathcal{H}} \|x\|^{2+2\delta} \,\mathrm{d}\mu(x)\right)^{1/(2+2\delta)} \cdot (\delta n)^{-1/2},$$

where $C \geq 1$ *is a universal constant.*

Theorem 31 (Functional Hoefffding Theorem [300, Theorem 3.26])**.** *Let* $\mathcal{F}$ *be a class of functions, each of the form* $h: \mathcal{B} \to \mathbb{R}$*, and* $X_1, \ldots, X_n$ *be samples i.i.d. drawn from* μ *on* $\mathcal{B}$*. For* $i \in [n]$*, assume there are real numbers* $a_{i,h} \leq b_{i,h}$ *such that* $h(x) \in [a_{i,h}, b_{i,h}]$ *for any* $x \in \mathcal{B}, h \in \mathcal{F} \cup \{-\mathcal{F}\}$*. Define*

$$L^2 = \sup_{h \in \mathcal{F} \cup \{-\mathcal{F}\}} \frac{1}{n}\sum_{i=1}^n (b_{i,h} - a_{i,h})^2.$$

For all $\delta \geq 0$*, it holds that*

$$\mathbb{P}\left\{\sup_{h\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n h(X_i)\right| \geq \mathbb{E}\left[\sup_{h\in\mathcal{F}} \left|\frac{1}{n}\sum_{1}^n h(X_i)\right|\right] + \delta\right\} \leq \exp\left(-\frac{n\delta^2}{4L^2}\right).$$

We first show the one-sample guarantees for KMS p-Wasserstein distance.

Proposition 12. *Fix* $p \in [1, \infty)$*, error probability* $\alpha \in (0,1)$*, and suppose Assumption 2 holds. Let* $C \geq 1$ *be a universal constant. Then, we have the following results:*

(I) $\mathbb{E}\mathcal{KMS}_p(\mu, \widehat{\mu}_n) \leq A(2C^{1/p}) \cdot n^{-1/(2p)}$

(II) *With probability at least* $1-\alpha$*, it holds that*

$$\mathcal{KMS}_p(\mu, \widehat{\mu}_n) \leq 2^{1-1/p} A \left(C + 4\sqrt{\log \frac{1}{\alpha}} \right)^{1/p} \cdot n^{-1/(2p)}.$$

Proof of Proposition 12. Recall from (3.3) that

$$\mathcal{KMS}_p(\mu, \nu) = \mathcal{MS}_p\Big(\Phi_{\#}\mu, \Phi_{\#}\nu\Big).$$

Therefore, it suffices to derive one-sample guarantees for $\mathcal{MS}_p\Big(\Phi_{\#}\mu, \Phi_{\#}\widehat{\mu}_n\Big)$.

(I) Observe that under Assumption 2, we have

$$A^2 \geq K(x,x) = \langle K_x, K_x \rangle = \|K_x\|_{\mathcal{H}}^2,$$

and therefore $\|\Phi(x)\|_{\mathcal{H}} = \|K_x\|_{\mathcal{H}} \leq A, \forall x \in \mathcal{B}$. In other words, for every probability measure μ on $\mathcal{B}$, the probability measure $\Phi_{\#}\mu$ is supported on the ball in $\mathcal{H}$ centered at the origin with radius A. By Theorem 30 with $\delta = 1$, we obtain

$$\mathbb{E}\mathcal{KMS}_1(\mu, \widehat{\mu}_n) = \mathbb{E}\mathcal{MS}_1\Big(\Phi_{\#}\mu, \Phi_{\#}\widehat{\mu}_n\Big) \leq \frac{AC}{\sqrt{n}}.$$

Since $\Phi_{\#}\mu$ and $\Phi_{\#}\widehat{\mu}_n$ are supported on the ball of $\mathcal{H}$ centered at the origin with radius A, it holds that

$$\mathcal{MS}_p\Big(\Phi_{\#}\mu, \Phi_{\#}\widehat{\mu}_n\Big) \leq \Big[\mathcal{MS}_1\Big(\Phi_{\#}\mu, \Phi_{\#}\widehat{\mu}_n\Big) \cdot (2A)^{p-1}\Big]^{1/p}.$$

In other words,

$$\mathcal{KMS}_p(\mu, \widehat{\mu}_n) \leq \Big[\mathcal{KMS}_1(\mu, \widehat{\mu}_n) \cdot (2A)^{p-1}\Big]^{1/p}. \tag{B.1}$$

It follows that

$$\begin{aligned}\mathbb{E}\mathcal{KMS}_p(\mu,\widehat{\mu}_n) &= \mathbb{E}\mathcal{MS}_p\Big(\Phi_{\#}\mu,\Phi_{\#}\widehat{\mu}_n\Big)\\ &\le \mathbb{E}\Big[\mathcal{MS}_1\Big(\Phi_{\#}\mu,\Phi_{\#}\widehat{\mu}_n\Big)\cdot(2A)^{p-1}\Big]^{1/p}\\ &\le \Big\{\mathbb{E}\Big[\mathcal{MS}_1\Big(\Phi_{\#}\mu,\Phi_{\#}\widehat{\mu}_n\Big)\cdot(2A)^{p-1}\Big]\Big\}^{1/p}\\ &\le \left\{\frac{AC}{\sqrt{n}}\cdot(2A)^{p-1}\right\}^{1/p} = 2^{1-1/p}AC^{1/p}\cdot n^{-1/(2p)}.\end{aligned}$$

(II) For the second part, we re-write $\mathcal{KMS}_1(\mu,\widehat{\mu}_n)$ with $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ using the kantorovich dual reformulation of OT:

$$\mathcal{KMS}_1(\mu,\widehat{\mu}_n) = \sup_{\substack{f\in\mathcal{H}:\ \|f\|_{\mathcal{H}}\le 1,\\ g \text{ is 1-Lipschitz with } g(0)=0}} \left|\frac{1}{n}\sum_{i=1}^n\Big(g(f(x)) - \mathbb{E}_{x\sim\mu}[g(f(x))]\Big)\right|,$$

where the additional constraint $g(0)=0$ does not impact the optimal value of the OT problem. In other words, one can represent

$$\mathcal{KMS}_1(\mu,\widehat{\mu}_n) = \sup_{h\in\mathfrak{H}}\ \Big|\frac{1}{n}\sum_{i=1}^n h(x_i)\Big|,$$

where the function class

$$\mathfrak{H} = \Big\{x\mapsto g(f(x)) - \mathbb{E}_{x\sim\mu}[g(f(x))]:\ g \text{ is 1-Lipschitz with } g(0)=0, f\in\mathcal{H}, \|f\|_{\mathcal{H}}\le 1\Big\}.$$

Consequently, for any x,

$$|g(f(x))| = |g(f(x)) - g(0)| \le |f(x)| = |\langle f, K_x\rangle_{\mathcal{H}}| \le \|f\|_{\mathcal{H}}\|K_x\|_{\mathcal{H}} \le A.$$

One can apply Theorem 31 with $\mathcal{F}\equiv\mathfrak{H}$, $a_{i,h}\equiv -A - \mathbb{E}_{x\sim\mu}[g(f(x))]$, $b_{i,h}\equiv A -$

$\mathbb{E}_{x\sim\mu}[g(f(x))]$, where $h(x) = g(f(x)) - \mathbb{E}_{x\sim\mu}[g(f(x))]$, to obtain

$$\mathbb{P}\Big\{\mathcal{KMS}_1(\mu, \widehat{\mu}_n) \geq \mathbb{E}\left[\mathcal{KMS}_1(\mu, \widehat{\mu}_n)\right] + \delta\Big\} \leq \exp\left(-\frac{n\delta^2}{4(2A)^2}\right) = \exp\left(-\frac{n\delta^2}{16A^2}\right).$$

Or equivalently, the following relation holds with probability at least $1 - \alpha$:

$$\mathcal{KMS}_1(\mu, \widehat{\mu}_n) \leq \mathbb{E}\left[\mathcal{KMS}_1(\mu, \widehat{\mu}_n)\right] + 4An^{-1/2}\sqrt{\log\frac{1}{\alpha}} \leq An^{-1/2}\left(C + 4\sqrt{\log\frac{1}{\alpha}}\right).$$

By the relation (B.1), we find that with probability at least $1 - \alpha$,

$$\begin{aligned}
&\mathcal{KMS}_p(\mu, \widehat{\mu}_n) \\
\leq&\Big[An^{-1/2}\left(C + 4\sqrt{\log\frac{1}{\alpha}}\right) \cdot (2A)^{p-1}\Big]^{1/p} \\
=&2^{1-1/p}A\left(C + 4\sqrt{\log\frac{1}{\alpha}}\right)^{1/p} \cdot n^{-1/(2p)}.
\end{aligned}$$

□

We now complete the proof of Theorem 6. By the triangle inequality, with probability at least $1 - 2\alpha$, it holds that

$$\begin{aligned}
\mathcal{KMS}_p(\widehat{\mu}_n, \widehat{\nu}_n) &\leq \mathcal{KMS}_p(\mu, \widehat{\mu}_n) + \mathcal{KMS}_p(\nu, \widehat{\nu}_n) \\
&\leq 2 \cdot 2^{1-1/p}A\left(C + 4\sqrt{\log\frac{1}{\alpha}}\right)^{1/p} \cdot n^{-1/(2p)} \\
&\leq 4A\left(C + 4\sqrt{\log\frac{1}{\alpha}}\right)^{1/p} \cdot n^{-1/(2p)}.
\end{aligned}$$

Then, substituting α with $\alpha/2$ gives the desired result.

Proof of Corollary 1. It remains to show the type-II risk when proving this corollary. In

particular,

$$\begin{aligned}\text{Type-II Risk} &= \mathbb{P}_{H_1}\big\{\mathcal{KMS}_p(\widehat{\mu}_n,\widehat{\nu}_n) < \Delta(n,\alpha)\big\}\\ &= \mathbb{P}_{H_1}\big\{\mathcal{KMS}_p(\mu,\nu) - \mathcal{KMS}_p(\widehat{\mu}_n,\widehat{\nu}_n) \geq \mathcal{KMS}_p(\mu,\nu) - \Delta(n,\alpha)\big\}\\ &\leq \mathbb{P}_{H_1}\big\{\big|\mathcal{KMS}_p(\mu,\nu) - \mathcal{KMS}_p(\widehat{\mu}_n,\widehat{\nu}_n)\big| \geq \mathcal{KMS}_p(\mu,\nu) - \Delta(n,\alpha)\big\}\\ &\leq \frac{\mathbb{E}\big|\mathcal{KMS}_p(\mu,\nu) - \mathcal{KMS}_p(\widehat{\mu}_n,\widehat{\nu}_n)\big|}{\mathcal{KMS}_p(\mu,\nu) - \Delta(n,\alpha)},\end{aligned}$$

where the last relation is based on the Markov inequality and the assumption that $\mathcal{KMS}_p(\mu,\nu) - \Delta(n,\alpha) > 0$. Based on the triangular inequality, we can see that

$$\mathbb{E}\big|\mathcal{KMS}_p(\mu,\nu) - \mathcal{KMS}_p(\widehat{\mu}_n,\widehat{\nu}_n)\big| \leq \mathbb{E}[\mathcal{KMS}_p(\mu,\widehat{\mu}_n)] + \mathbb{E}[\mathcal{KMS}_p(\nu,\widehat{\nu}_n)] \leq 2AC^{1/p}\cdot n^{-1/(2p)}.$$

Combining these two upper bounds, we obtain the desired result. □

B.3 Sufficient Condition for Positive Definiteness of Matrix G

To implement our computational algorithm, one needs to assume the gram matrix

$$G = [K(x^n, x^n), -K(x^n, y^n); -K(y^n, x^n), K(y^n, y^n)]$$

to be strictly positive definite. By the Lemma on the Schur complement (see, e.g., [25, Lemma 4.2.1]), It can be showed that its necessary and sufficient condition should be

$$G' = [K(x^n, x^n), K(x^n, y^n); K(y^n, x^n), K(y^n, y^n)]$$

is strictly positive definite. By Wendland [317], this requires our data points $\{x_1, \ldots, x^n, y_1, \ldots, y_n\}$ are pairwise distinct and $K(x, y)$ is of the form $K(x, y) = \Phi(x - y)$, with $\Phi(\cdot)$ being continuous, bounded, and its Fourier transform is non-negative and non-vanishing. For instance, Gaussian kernel $K(x, y) = e^{-\|x-y\|_2^2/\sigma^2}$ or Bessel kernel $K(x, y) = (c^2 + \|x\|_2^2)^{-\beta}, x \in$

$\mathbb{R}^d, \beta > d/2$ satisfies our requirement.

B.4 Reformulation for 2-KMS Wasserstein Distance in (KMS)

In this section, we derive the reformulation for computing 2-KMS Wasserstein distance:

$$\max_{f\in\mathcal{H},\ \|f\|_{\mathcal{H}}^2\le 1} \left\{ \min_{\pi\in\Gamma_n} \sum_{i,j\in[n]} \pi_{i,j} |f(x_i) - f(y_j)|^2 \right\}. \tag{B.2}$$

Based on the expression of f in (3.7), we reformulate the problem above as

$$\max_{a_x,a_y\in\mathbb{R}^n} \left\{ \min_{\pi\in\Gamma_n} \sum_{i,j\in[n]} \pi_{i,j} \left| \sum_{l\in[n]} a_{x,l} K(x_i,x_l) - \sum_{l\in[n]} a_{y,l} K(y_j,y_l) \right|^2 \right\}, \tag{B.3a}$$

subject to the constraint

$$\begin{aligned} &\left\| \sum_{i=1}^n a_{x,i} K(\cdot,x_i) - \sum_{i=1}^n a_{y,i} K(\cdot,y_i) \right\|_{\mathcal{H}}^2 \\ =&\langle \sum_{i=1}^n a_{x,i} K(\cdot,x_i) - \sum_{i=1}^n a_{y,i} K(\cdot,y_i) \sum_{i=1}^n a_{x,i} K(\cdot,x_i) - \sum_{i=1}^n a_{y,i} K(\cdot,y_i) \rangle \\ =&\sum_{i,j\in[n]} a_{x,i} a_{x,j} \langle K(\cdot,x_i), K(\cdot,x_j)\rangle + \sum_{i,j\in[n]} a_{y,i} a_{y,j} \langle K(\cdot,y_i), K(\cdot,y_j)\rangle \\ &- 2 \sum_{i,j\in[n]} a_{x,i} a_{y,j} \langle K(\cdot,x_i), K(\cdot,y_j)\rangle \le 1. \end{aligned} \tag{B.3b}$$

One can re-write (B.3) in compact matrix form. If we define

$$\begin{aligned} s &= [a_x; a_y], \\ M'_{i,j} &= [(K(x_i,x_l) - K(y_i,x_l))_{l\in[n]}; (K(y_j,y_l) - K(x_i,y_l))_{l\in[n]}], \\ G &= [K(x^n,x^n), -K(x^n,y^n); -K(y^n,x^n), K(y^n,y^n)] \in \mathbb{R}^{2n\times 2n}, \end{aligned}$$

Problem (B.3) can be reformualted as

$$\max_{s\in\mathbb{R}^{2n}} \left\{ \min_{\pi\in\Gamma_n} \sum_{i,j\in[n]} \pi_{i,j} \left| s^{\mathrm{T}} M'_{i,j} \right|^2 : \quad s^{\mathrm{T}} G s \le 1 \right\}. \tag{B.4}$$

Take Cholesky decomposition $G^{-1} = UU^{\mathrm{T}}$ and use the change of variable approach to take $\omega = U^{-1}s$, Problem (B.4) can be further reformulated as

$$\max_{\omega\in\mathbb{R}^{2n}} \left\{ \min_{\pi\in\Gamma_n} \sum_{i,j\in[n]} \pi_{i,j} \big(\langle \omega, U^{\mathrm{T}} M'_{i,j} \rangle \big)^2 : \quad \omega^{\mathrm{T}} \omega \le 1 \right\}. \tag{B.5}$$

After defining $M_{i,j} = U^{\mathrm{T}} M'_{i,j}$ and observing that the inequality constraint $\omega^{\mathrm{T}}\omega \le 1$ will become tight, we obtain the desired reformulation as in (3.9).

B.5 Proof of Theorem 8

The general procedure of NP-hardness proof is illustrated in the following diagram: Problem (3.9) contains the **(Fair PCA with rank-1 data)** as a special case, whereas this special problem further contains **(Partition)** (which is known to be NP-complete) as a special case. After building these two reductions, we finish the proof of Theorem 8.

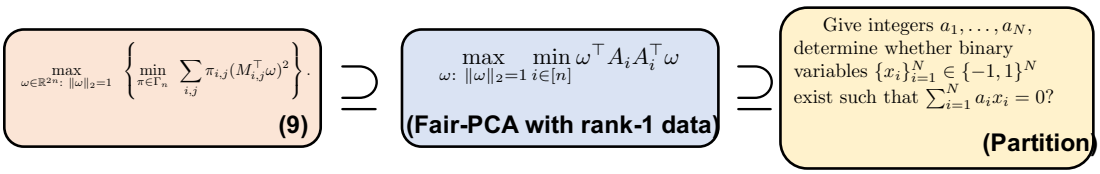

Figure B.1: Proof outline of Theorem 8

Claim 1. Problem (3.9) contains Problem (**Fair PCA with rank-1 data**).

Proof of Claim 1. Given vectors $A_1, \ldots, A_n$, we specify

$$M_{1,:} \triangleq \{M_{1,1}, M_{1,2}, \ldots, M_{1,n}\} = \{A_1, \ldots, A_n\},$$

and $M_{i,:} \triangleq \{M_{i,1}, M_{i,2}, \ldots, M_{i,n}\}, i = 2, \ldots, n$ is specified by circularly shifting the elements in $M_{1,:}$ by $i-1$ positions. For instance, $M_{2,:} = \{A_n, A_1, \ldots, A_{n-1}\}$. For the inner OT problem in (3.9), it suffices to consider deterministic optimal transport π, i.e.,

$$\pi_{i,j} = \begin{cases} 1/n, & \text{if } j = \sigma(i), \\ 0, & \text{otherwise} \end{cases}$$

for some bijection mapping $\sigma : [n] \to [n]$. The cost matrix for the inner OT is actually a circulant matrix:

$$\left((M_{i,j}^{\mathrm{T}}\omega)^2\right)_{i,j} = \begin{pmatrix} (A_1\omega)^2 & (A_2\omega)^2 & \cdots & (A_n\omega)^2 \\ (A_n\omega)^2 & (A_1\omega)^2 & \cdots & (A_{n-1}\omega)^2 \\ \vdots & \vdots & \ddots & \vdots \\ (A_2\omega)^2 & (A_3\omega)^2 & \cdots & (A_1\omega)^2 \end{pmatrix}.$$

When considering the feasible circularly shifting bijection mapping (e.g., $\sigma(i) = (i+j) \mod n, \forall i \in [n]$ for $j = 0, 1, \ldots, n-1$), we obtain the upper bound on the optimal value of the inner OT problem in (3.9):

$$\min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} (M_{i,j}^{\mathrm{T}}\omega)^2 \leq \min_{i \in [n]} (A_i^{\mathrm{T}}\omega)^2 = \min_{i \in [n]} \omega^{\mathrm{T}} A_i A_i^{\mathrm{T}} \omega.$$

On the other hand, for any bijection mapping σ, the objective of the inner OT problem in (3.9) can be written as a convex combination of $(A_1^{\mathrm{T}}\omega)^2, \ldots, (A_n^{\mathrm{T}}\omega)^2$, and thus,

$$\min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} (M_{i,j}^{\mathrm{T}}\omega)^2 \geq \min_{\alpha \in \mathbb{R}_n^+, \sum_i \alpha_i = 1} \left\{ \sum_i \alpha_i (A_i^{\mathrm{T}}\omega)^2 \right\} \geq \min_{i \in [n]} (A_i^{\mathrm{T}}\omega)^2.$$

Since the upper and lower bounds match with each other, we obtain

$$\min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} (M_{i,j}^{\mathrm{T}}\omega)^2 = \min_{i \in [n]} \omega^{\mathrm{T}} A_i A_i^{\mathrm{T}} \omega,$$

and consequently,

$$\max_{\omega:\ \|\omega\|_2=1} \left\{ \min_{\pi\in\Gamma_n} \sum_{i,j} \pi_{i,j} (M_{i,j}^{\mathrm{T}}\omega)^2 \right\} = \max_{\omega:\ \|\omega\|_2=1} \left\{ \min_{i\in[n]} \omega^{\mathrm{T}} A_i A_i^{\mathrm{T}} \omega \right\},$$

which justifies Problem (3.9) contains Problem (**Fair PCA with rank-1 data**). □

Claim 2. Problem (**Fair PCA with rank-1 data**) contains Problem (**Partition**).

It is noteworthy that Claim 2 has previously been proved by [283]. For the sake of completeness, we provide the proof here.

Proof of Claim 2. Consider the norm minimization problem

$$P = \min_{\omega} \left\{ \|\omega\|_2^2 :\ \min_{i\in[n]} (\omega^{\mathrm{T}} A_i)^2 \geq 1 \right\}. \tag{B.6}$$

and the scaled problem

$$\max_{\omega} \left\{ \min_{i\in[n]} (\omega^{\mathrm{T}} A_i)^2 :\ \|\omega\|_2^2 = P \right\}. \tag{B.7}$$

We can show that Problem (**Fair PCA with rank-1 data**) is equivalent to (B.7), whereas (B.7) is equivalent to (B.6). Indeed,

- For the first argument, for any optimal solution from Problem (**Fair PCA with rank-1 data**), denoted as ω^*, one can do the scaling to consider $\tilde{\omega}^* = \sqrt{P}\omega$, which is also optimal to (B.7), and vise versa.

- For the second argument, let ω_1, ω_2 be optimal solutions from (B.6), (B.7), respectively. Since P is the optimal value of (B.6), one can check that ω_1 is a feasible solution to (B.7). Since $\min_{i\in[n]} (\omega_1^{\mathrm{T}} A_i)^2 \geq 1$, by the optimality of ω_2, it holds that $\min_{i\in[n]} (\omega_2^{\mathrm{T}} A_i)^2 \geq 1$, i.e., ω_2 is a feasible solution to (B.6). Since $\|\omega_2\|_2^2 = P$, ω_2 is an optimal solution to (B.6). Reversely, one can show ω_1 is an optimal solution to (B.7): suppose on the contrary that there exists $\bar{\omega}_1$ such that $\|\bar{\omega}_1\|_2^2 = P$ and

$\min_{i\in[n]} (\bar{\omega}_1^{\mathrm{T}} A_i)^2 > \min_{i\in[n]} (\omega_1^{\mathrm{T}} A_i)^2 \geq 1$, then one can do a scaling of $\bar{\omega}_1$ such that $\min_{i\in[n]} (\bar{\omega}_1^{\mathrm{T}} A_i)^2 = 1$ whereas $\|\bar{\omega}_1\|_2^2 > P$, which contradicts to the optimality of P. Combining both directions, we obtain the equivalence argument.

Thus, it suffices to show (B.6) contains Problem (**Partition**). Define $a = (a_i)_{i\in[n]}, Q = I_n + aa^{\mathrm{T}}$, and assume Q admits Cholesky factorization $Q = S^{\mathrm{T}}S$. Then we create the vector $A_i = S^{-\top} e_i$, where e_i is the i-th unit vector of length n. Then, it holds that

$$
\begin{aligned}
&\text{(B.6)} \\
&= \min_{\omega} \ \left\{ \|\omega\|_2^2 : \ \min_{i\in[n]} ((S^{-1}\omega)^{\mathrm{T}} e_i)^2 \geq 1 \right\} \\
&= \min_{\omega} \ \left\{ \|Sx\|_2^2 : \ \min_{i\in[n]} (x^{\mathrm{T}} e_i)^2 \geq 1 \right\} \\
&= \min_{\omega} \ \left\{ x^{\mathrm{T}} Q x : \ x_i^2 \geq 1 \right\} \\
&= \min_{\omega} \ \left\{ \sum_{i=1}^n x_i^2 + \left(\sum_{i=1}^n a_i x_i \right)^2 : \ x_i^2 \geq 1 \right\} \qquad (*)
\end{aligned}
$$

where the second equality is by the change of variable $x = S^{-1}\omega$, the third equality is by the definitions of S and e_i, and the last equality is by the definition of Q. The solution to Problem (**Partition**) exists if and only if the optimal value to Problem (*) equals n. Thus, we finish the proof of Claim 2. □

B.6 Algorithm that Finds Near-optimal Solution to Optimal Transport

In this section, we present the algorithm that returns ϵ-optimal solution to the following OT problem:

$$
\min_{\pi\in\Gamma_n} \sum_{i,j} \pi_{i,j} c_{i,j}, \tag{B.8}
$$

where $\{c_{i,j}\}_{i,j}$ is the given cost matrix. Define $\|C\|_\infty = \max_{i,j} c_{i,j}$. In other words, we find $\widehat{\pi} \in \Gamma_n$ such that

$$\texttt{optval(B.8)} \leq \sum_{i,j} \widehat{\pi}_{i,j} c_{i,j} \leq \texttt{optval(B.8)} + \epsilon.$$

Entropy-Regularized OT. The key to the designed algorithm is to consider the entropy regularized OT problem

$$\min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} c_{i,j} + \eta \sum_{i,j} \pi_{i,j} \log(\pi_{i,j}),$$

whose dual problem is

$$\min_{v \in \mathbb{R}^n} \left\{ G(v) = \frac{1}{n} \sum_{i=1}^{n} h_i(v) \right\}, \tag{B.9}$$

where

$$h_i(v) = \eta \log \sum_j \exp\left(\frac{v_j - c_{i,j} - \eta}{\eta} \right) - \frac{1}{n} \sum_j v_j + \eta(1 + \log n).$$

Given the dual variable $v \in \mathbb{R}^n$, one can recover the primal variable π using

$$\pi(v) = \frac{\frac{1}{n} \exp\left(\frac{v_j - c_{i,j} - \eta}{\eta} \right)}{\sum_{j' \in [n]} \exp\left(\frac{v_{j'} - c_{i,j'} - \eta}{\eta} \right)}$$

Algorithm 12 essentailly optimizes the dual formulation (B.9) with light computational speed.

Theorem 32 ([325, Theorem 3])**.** *Suppose we specify $T_{out} = \mathcal{O}(\frac{\|C\|_\infty \sqrt{\ln n}}{\epsilon}), T = n$, the number of total iterations (including outer and inner iterations) of Algorithm 12 is $\mathcal{O}(\frac{n\|C\|_\infty \sqrt{\ln n}}{\epsilon})$ with per-iteration cost $\mathcal{O}(n)$. Therefore, the number of arithmetic operations of Algorithm 12 for finding ϵ-optimal solution is $\mathcal{O}(\frac{n^2\|C\|_\infty \sqrt{\ln n}}{\epsilon})$*

Algorithm 12 Stochastic Gradient-based Algorithm with Katyusha Momentum for solving OT [325]

Input: Accuracy $\epsilon > 0$, $\eta = \frac{\epsilon}{8 \log n}$, $\epsilon' = \frac{\epsilon}{6 \max_{i,j} c_{i,j}}$, maximum outer iteration T_{out}, and maximum inner iteration T.
Take $(y_0, z_0, \widetilde{\lambda}_0, \lambda_0, C_0, D_0) = (0, 0, 0, 0, 0, 0)$
for $t = 0, \ldots, T_{\text{out}} - 1$ **do**
  $\tau_{1,t} = \frac{2}{t+4}, \gamma_t = \frac{\eta}{9\tau_{1,t}}$
  $u_t = \nabla \phi(\widetilde{\lambda}_t)$
  for $j = 0, \ldots, T - 1$ **do**
    $k = j + tT$
    $\lambda_{k+1} = \tau_{1,t} z_k + \frac{1}{2}\widetilde{\lambda}_t + (\frac{1}{2} - \tau_{1,t}) y_k$
    Sample i uniformly from $[n]$, and construct

$$H_{k+1} = u_t + \Big(\nabla h_i(\lambda_{k+1}) - \nabla h_i(\widetilde{\lambda}_t) \Big)$$

    Update $z_{k+1} = z_k - \gamma_t \cdot H_{k+1}/2$ and $y_{k+1} = \lambda_{k+1} - \eta H_{k+1}/9$
  end for
  Update $\widetilde{\lambda}_{t+1} = \frac{1}{T} \sum_{j=1}^{T} y_{tT+j}$
  Sample $\widehat{\lambda}_t$ uniformly from $\{\lambda_{tT+1}, \ldots, \lambda_{tT+T}\}$ and take $D_t = D_t + \mathsf{vec}(\pi(\widehat{\lambda}_t))/\tau_{1,t}$
  $C_t = C_t + 1/\tau_{1,t}$
  $\pi_{t+1} = D_t / C_t$
end for
Query Algorithm 13 to Round $\widetilde{\pi} := \pi_{T_{\text{out}}}$ to $\widehat{\pi}$ such that $\widehat{\pi}\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n$ and $\widehat{\pi}^{\mathrm{T}}\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n$
Return $\widehat{\pi}$

B.7 Proof of Theorem 9

To analyze the complexity of Algorithm 3, we first derive the bias and computational cost of the supgradient estimator $v(S)$ in (3.12).

Lemma 16 (Bias and Computational Cost)**.** *The following results hold:* (I) ***(Bias)*** $v(S)$ *corresponds to the gradient of* $\widehat{F}(S) = \sum_{i,j} \widehat{\pi}_{i,j} \langle M_{i,j}^{\mathrm{T}} M_{i,j}, S \rangle$*, where* $\widehat{\pi}$ *is defined in* (3.12) *and* $|F(S) - \widehat{F}(S)| \leq \epsilon$*;*

(II) ***(Cost)*** *The cost for computing* (3.12) *is* $\mathcal{O}\left(C \cdot n^2 \sqrt{\log n} \epsilon^{-1}\right)$*, with* $\mathcal{O}(\cdot)$ *hiding some universal constant.*

Next, we analyze the error of the inexact mirror ascent framework in Algorithm 3.

Algorithm 13 Round to Γ_n ([6, Algorithm 2])

1: **Input:** $\pi \in \mathbb{R}_+^{n\times n}$
2: $X = \mathrm{diag}(x_1, \ldots, x_n)$, with $x_i = \min\{1, \frac{1}{nr_i(\pi)}\}$. Here $r_i(\pi)$ denotes the i-th row sum of π.
3: $\pi' = X\pi$.
4: $Y = \mathrm{diag}(y_1, \ldots, y_n)$, with $y_j = \min\{1, \frac{1}{nc_i(\pi')}\}$. Here $c_j(\pi')$ denotes the j-th column sum of π'.
5: $\pi'' = \pi' Y$.
6: $\mathbf{e}_r = \frac{1}{n}\mathbf{1}_n - r(\pi''), \mathbf{e}_c = \frac{1}{n}\mathbf{1}_n - c(\pi'')$, where

$$r(\pi'') = (r_i(\pi''))_{i\in[n]}, c(\pi'') = (c_j(\pi''))_{j\in[n]}.$$

7: **Return** $\pi'' + \mathbf{e}_r\mathbf{e}_c^{\mathrm{T}}/\|\mathbf{e}_r\|_1$.

Lemma 17 (Error Analysis of Algorithm 3)**.** *When taking the stepsize* $\gamma = \frac{\log(2n)}{C\sqrt{T}}$*, the output* $\widehat{S}_{1:T}$ *from Algorithm 3 satisfies*

$$0 \le F(S^*) - F(\widehat{S}_{1:T}) \le 2\epsilon + 2C\sqrt{\frac{\log(2n)}{T}}.$$

Combining Lemmas 16 and 17, we obtain the complexity for solving (SDR).

Proof of Lemma 16. For the first part, it is noteworthy that $v(S)$ is associated with the objective

$$\widehat{F}(S) = \sum_{i,j} \widehat{\pi}_{i,j}\langle M_{i,j}^{\mathrm{T}} M_{i,j}, S\rangle,$$

where $\widehat{\pi}_{i,j}$ is the ϵ-optimal solution to

$$F(S) = \min_{\pi\in\Gamma_n} \sum_{i,j} \pi_{i,j}\langle M_{i,j}^{\mathrm{T}} M_{i,j}, S\rangle.$$

By definition, it holds that

$$0 \le \widehat{F}(S) - F(S) \le \epsilon.$$

The second part follows from Theorem 32. □

The proof of Lemma 17 replies on the following technical result.

Lemma 18 ([233]). *Let $\{S_k\}_{k=1}^T$ be the updating trajectory of mirror ascent aiming to solve the maximization of $G(S)$ with $S \in \mathcal{S}_{2n}$, i.e.,*

$$S_{k+1} = \arg\max_{S \in \mathcal{S}_{2n}} \gamma \langle v(S_k), S \rangle + V(S, S_k), \quad k = 1, \ldots, T-1.$$

Here $v(S)$ is a supgradient of $G(S)$, and we assume there exists $M_* > 0$ such that

$$\|v(S)\|_{Tr} := \textit{Trace}(v(S)) \leq M_*, \qquad \forall S \in \mathcal{S}_{2n}.$$

Let $\widehat{S}_{1:T} = \frac{1}{T}\sum_{k=1}^T S_k$, and S^* be a maximizer of $G(S)$. Define the diameter

$$D^2_{\mathcal{S}_{2n}} = \max_{S \in \mathcal{S}_{2n}} h(S) - \min_{S \in \mathcal{S}_{2n}} h(S) = \log(2n).$$

For constant step size

$$\gamma = \frac{D^2_{\mathcal{S}_{2n}}}{M_* \sqrt{T}} = \frac{\log(2n)}{M_* \sqrt{T}},$$

it holds that

$$0 \leq G(S^*) - G(\widehat{S}_{1:T}) \leq M_* \sqrt{\frac{4\log(2n)}{T}}.$$

Proof of Lemma 17. Let S^* and $\widehat{S}^*$ be maximizers of the objective $F(\cdot)$ and $\widehat{F}(\cdot)$, then we have the following error decomposition:

$$\begin{aligned}
&F(S^*) - F(\widehat{S}_{1:T}) \\
=&\big[F(S^*) - \widehat{F}(S^*)\big] + \big[\widehat{F}(S^*) - \widehat{F}(\widehat{S}^*)\big] + \big[\widehat{F}(\widehat{S}^*) - \widehat{F}(\widehat{S}_{1:T})\big] + \big[\widehat{F}(\widehat{S}_{1:T}) - F(\widehat{S}_{1:T})\big] \\
\leq& 2\epsilon + \big[\widehat{F}(S^*) - \widehat{F}(\widehat{S}^*)\big] + \big[\widehat{F}(\widehat{S}^*) - \widehat{F}(\widehat{S}_{1:T})\big] \\
\leq& 2\epsilon + \big[\widehat{F}(\widehat{S}^*) - \widehat{F}(\widehat{S}_{1:T})\big],
\end{aligned}$$

where the first inequality is because $\|F - \widehat{F}\|_\infty \leq \epsilon$ and

$$|\big[F(S^*) - \widehat{F}(S^*)\big]| \leq \epsilon, |\big[\widehat{F}(\widehat{S}_{1:T}) - F(\widehat{S}_{1:T})\big]| \leq \epsilon;$$

and the second inequality is because $\widehat{F}(\widehat{S}^*) - \widehat{F}(\widehat{S}_{1:T}) \leq 0$. It remains to bound $\big[\widehat{F}(\widehat{S}^*) - \widehat{F}(\widehat{S}_{1:T})\big]$. It is worth noting that

$$\|v(S)\|_{\text{Tr}} = \sum_{i,j} \pi_{i,j} \|M_{i,j} M_{i,j}^{\text{T}}\|_{\text{Tr}} \leq \sum_{i,j} \pi_{i,j} \cdot C = C.$$

Therefore, the proof can be finished by querying Lemma 18 with $M_* = C$ and stepsize $\gamma = \frac{\log(2n)}{C\sqrt{T}}$. □

Proof of Theorem 9. The proof can be finished by taking hyper-parameters such that

$$2\epsilon \leq \frac{\delta}{2}, \qquad 2C\sqrt{\frac{\log(2n)}{T}} \leq \frac{\delta}{2}.$$

In other words, we take $\epsilon = \frac{\delta}{4}$ and $T = \lceil \frac{16C^2 \log(2n)}{\delta^2} \rceil$. We follow the proof of Lemma 17 to take stepsize $\gamma = \frac{\log(2n)}{C\sqrt{T}}$. □

B.8 Proof of Theorems 10 and 12

We rely on the following two technical results when proving Theorem 10.

Theorem 33 (Birkhoff-von Neumann Theorem [37])**.** *Consider the discrete OT problem*

$$\min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} c_{i,j}, . \tag{B.10}$$

There exists an optimal solution π that has exactly one entry of $1/n$ in each row and each column with all other entries 0.

Theorem 34 (Rank Bound, Adopted from [198, Theorem 2] and [195, Lemma 1])**.** *Consider*

the domain set

$$\mathcal{D} = \Big\{ S \in \mathbb{S}_m^+ : \ \textit{Trace}(S) = 1 \Big\}$$

and the intersection of N linear inequalities:

$$\mathcal{E} = \Big\{ S \in \mathbb{R}^{m \times m} : \ \langle S, A_i \rangle \geq b_i, i \in [N] \Big\}.$$

Then, any feasible extreme point in $\mathcal{D} \cap \mathcal{E}$ has a rank at most $1 + \lceil \sqrt{2N + 9/4} - 3/2 \rceil$. Such a rank bound can be strengthened by replacing N by the number of binding constraints in $\mathcal{E}$.

Proof of Theorem 10. By taking the dual of inner OT problem, we find (SDR) can be reformulated as

$$\max_{\substack{S \in \mathcal{S}_{2n} \\ f,g \in \mathbb{R}^n}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (f_i + g_i) : \quad f_i + g_j \leq \langle M_{i,j} M_{i,j}^{\mathrm{T}}, S \rangle, \quad \forall i, j \in [n] \right\}. \tag{B.11}$$

Let S^* be the optimal solution of variable S to the optimization problem above. Then for fixed S^*, according to Theorem 33 and complementary slackness of OT, there exists optimal solutions (f^*, g^*) such that only n constraints out of n^2 constraints in (B.11) are binding. Moreover, an optimal solution to (SDR) can be obtained by finding a feasible solution to the following intersection of constraints:

$$\text{Find } S \in \mathcal{S}_{2n} \bigcap \mathcal{E} \triangleq \Big\{ S : \ f_i^* + g_j^* \leq \langle M_{i,j} M_{i,j}^{\mathrm{T}}, S \rangle, \ \ i, j \in [n] \Big\}.$$

By Theorem 34, any feasible extreme point from $\mathcal{S}_{2n} \cap \mathcal{E}$ has rank at most $1 + \Big\lfloor \sqrt{2n + \frac{9}{4}} - \frac{3}{2} \Big\rfloor$. Thus, we pick such a feasible extreme point to satisfy the requirement of Theorem 10. $\square$

Proof of Theorem 12. Recall that

$$\text{(KMS)} = \max_{\substack{S \succeq 0, \text{Trace}(S)=1, \text{rank}(S)=1, \\ f,g \in \mathbb{R}^n}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (f_i + g_i) : \quad f_i + g_j \leq \langle M_{i,j} M_{i,j}^{\mathrm{T}}, S \rangle, \quad \forall i, j \in [n] \right\}$$

and

$$(\mathrm{SDR}) = \max_{\substack{S\succeq 0, \mathrm{Trace}(S)=1 \\ f,g\in\mathbb{R}^n}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(f_i+g_i) : \quad f_i+g_j \le \langle M_{i,j}M_{i,j}^{\mathrm{T}}, S\rangle, \quad \forall i,j\in[n] \right\}.$$

It is easy to see Optval(**KMS**) $\le \mathrm{Optval}(\mathrm{SDR})$. On the other hand, let $(\widehat{S},\widehat{f},\widehat{g})$ be an optimal solution to (**SDR**) such that $\mathrm{rank}(\widehat{S}) \le k \triangleq 1+\left\lfloor \sqrt{2n+\frac{9}{4}}-\frac{3}{2}\right\rfloor$. Next, take

$$\zeta \sim \mathcal{N}(0,\widehat{S}), \qquad \xi = \frac{\zeta}{\|\zeta\|_2}, \qquad \widetilde{S} = \xi\xi^{\mathrm{T}}.$$

It can be seen that $\widetilde{S}\succeq 0, \mathrm{Trace}(\widetilde{S})=1, \mathrm{rank}(\widetilde{S})=1$. Then, for any $\varepsilon\in(0,1]$ and $\mu>0$, it holds that

$$\begin{aligned}
&\Pr\Big\{\langle M_{i,j}M_{i,j}^{\mathrm{T}},\widetilde{S}\rangle \ge \varepsilon\cdot\big(\widehat{f}_i+\widehat{g}_j\big), \quad \forall i,j\in[n]\Big\}\\
\ge&\Pr\Big\{\langle M_{i,j}M_{i,j}^{\mathrm{T}},\widetilde{S}\rangle \ge \varepsilon\cdot\langle M_{i,j}M_{i,j}^{\mathrm{T}},\widehat{S}\rangle, \quad \forall i,j\in[n]\Big\}\\
=&\Pr\Big\{\langle M_{i,j}M_{i,j}^{\mathrm{T}},\widetilde{S}\rangle \ge \varepsilon\cdot\mathbb{E}[\langle M_{i,j}M_{i,j}^{\mathrm{T}},\zeta\zeta^{\mathrm{T}}\rangle], \quad \forall i,j\in[n]\Big\}\\
=&\Pr\Big\{\langle M_{i,j}M_{i,j}^{\mathrm{T}},\zeta\zeta^{\mathrm{T}}\rangle\cdot\|\zeta\|_2^{-2} \ge \varepsilon\cdot\mathbb{E}[\langle M_{i,j}M_{i,j}^{\mathrm{T}},\zeta\zeta^{\mathrm{T}}\rangle], \quad \forall i,j\in[n]\Big\}\\
\ge&\Pr\Big\{\langle M_{i,j}M_{i,j}^{\mathrm{T}},\zeta\zeta^{\mathrm{T}}\rangle \ge \frac{\varepsilon}{\mu}\cdot\mathbb{E}[\langle M_{i,j}M_{i,j}^{\mathrm{T}},\zeta\zeta^{\mathrm{T}}\rangle], \quad \forall i,j\in[n], \quad \|\zeta\|_2^{-2}\ge\mu\Big\}\\
\ge&1-\sum_{(i,j)\in[n]}\Pr\Big\{\langle M_{i,j}M_{i,j}^{\mathrm{T}},\zeta\zeta^{\mathrm{T}}\rangle < \frac{\varepsilon}{\mu}\cdot\mathbb{E}[\langle M_{i,j}M_{i,j}^{\mathrm{T}},\zeta\zeta^{\mathrm{T}}\rangle]\Big\}-\Pr\Big\{\|\zeta\|_2>\frac{1}{\sqrt{\mu}}\Big\}\\
\ge&1-n^2\cdot\sqrt{\frac{\varepsilon}{\mu}}-\mu.
\end{aligned}$$

As long as we take $\mu=33/100$, $\varepsilon=4\cdot(33/100)^3/n^4$, it holds that

$$\Pr\Big\{\langle M_{i,j}M_{i,j}^{\mathrm{T}},\widetilde{S}\rangle \ge \varepsilon\cdot\big(\widehat{f}_i+\widehat{g}_j\big), \quad \forall i,j\in[n]\Big\} \ge 1/100.$$

In other words, there exists ξ such that

$$(\varepsilon\cdot\widehat{f}_i)+(\varepsilon\cdot\widehat{g}_j) \le \langle M_{i,j}M_{i,j}^{\mathrm{T}},\xi\xi^{\mathrm{T}}\rangle, \quad \forall i,j\in[n]^2, \qquad \|\xi\|_2=1.$$

Algorithm 14 Rank reduction algorithm for (SDR)

1: Run Algorithm 3 to obtain δ-optimal solution to (SDR), denoted as $\widehat{S}$.

`// Step 2: Find` n `binding constraints`

2: Run Hungarian algorithm [184] to solve OT (3.11) with $S \equiv \widehat{S}$, and obtain an optimal assignment $\sigma : [n] \to [n]$ together with dual optimal solutoin (f^*, g^*).

`// Step 3-9: Calibrate low-rank solution using a greedy algorithm`

3: Initialize $\delta^* = 1$
4: **while** $\delta^* > 0$ **do**
5: Perform eigendecomposition $\widehat{S} = Q\Lambda Q^{\mathrm{T}}$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$ with $\mathrm{rank}(\widehat{S}) = r$
6: Find a direction $Y = Q\Delta Q^{\mathrm{T}}$, where $\Delta \in \mathcal{S}^r$ is some nonzero matrix satisfying

$$\mathrm{Trace}(\Delta) = 0, \langle Q^{\mathrm{T}} M_{i,\sigma(i)} M_{i,j}^{\mathrm{T}} Q, \Delta \rangle = 0, \quad \forall i \in [n].$$

7: **If** such Y does not exist, **then** break the loop.
8: Take new solution $\widehat{S}(\delta^*) := \widehat{S} + \delta^* Y$, where

$$\delta^* = \arg\max_{\delta \geq 0} \Big\{ \delta : \ \lambda_{\min}(\Lambda + \delta\Delta) \geq 0 \Big\}.$$

9: Update $\widehat{S} = \widehat{S}(\delta^*)$
10: **end while**
11: **Return** $\widehat{S}$

This indicates that $(\xi\xi^{\mathrm{T}}, \varepsilon \cdot \widehat{f}, \varepsilon \cdot \widehat{g})$ is a feasible solution to (KMS), and consequently,

$$\mathrm{Optval}(\mathrm{KMS}) \geq \frac{\varepsilon}{n} \sum_{i=1}^{n} (\varepsilon \cdot \widehat{f}_i + \varepsilon \cdot \widehat{g}_i) = \varepsilon \cdot (\mathrm{KMS}).$$

□

B.9 Rank Reduction Algorithm

In this section, we develop a rank reduction algorithm that, based on the near-optimal solution (denoted as $\widehat{S}$) returned from Algorithm 3, finds an alternative solution of the same (or smaller) optimality gap while satisfying the desired rank bound in Theorem 10.

Step (i): Find n binding constraints. First, we fix $S \equiv \widehat{S}$ in (3.14) and finds the optimal

solution (f^*, g^*) such that only n constraints out of n^2 constraints are binding. It suffices to apply the Huangarian algorithm [184] to solve the following balanced discrete OT problem

$$\max_{f,g\in\mathbb{R}^n}\left\{\frac{1}{n}\sum_{i=1}^n(f_i+g_i):\ \ f_i+g_j\le c_{i,j}\right\}=\min_{\pi\in\Gamma_n}\left\{\sum_{i,j=1}^n\pi_{i,j}c_{i,j}\right\}$$

where $c_{i,j}=\langle M_{i,j}M_{i,j}^{\mathrm{T}},\widehat{S}\rangle$. The output of the Huangarian algorithm is a *deterministic* optimal transport that moves n probability mass points from the left marginal distribution of π to the right, which is denoted as a bijection σ that permutes $[n]$ to $[n]$. Thus, these n binding constraints are denoted as

$$f_i^*+g_{\sigma(i)}^*\le\langle M_{i,j}M_{i,j}^{\mathrm{T}},S\rangle,\quad i\in[n].$$

We denote by the intersection of these n constraints as $\mathcal{E}_n$.

Step (ii): Calibrate low-rank solution using a greedy algorithm. Second, let us assume $\widehat{S}$ is not an extreme point of $\mathcal{S}_{2n}\cap\mathcal{E}_n$, since otherwise one can terminate the algorithm to output $\widehat{S}$ following the proof of Theorem 10. We run the following greedy rank reduction procedure:

(I) We find a direction $Y\neq 0$, along which $\widehat{S}$ remains to be feasible, and the null space of $\widehat{S}$ is non-decreasing.

(II) Then, we move $\widehat{S}$ along the direction Y until its smallest non-zero eigenvalue vanishes. We update $\widehat{S}$ to be such a new boundary point.

(III) We terminate the iteration until no movement is available.

To achieve (I), denote the eigendecomposition of $\widehat{S}$ with $\mathrm{rank}(\widehat{S})=r$ as

$$\widehat{S}=\begin{pmatrix}Q & 0\end{pmatrix}\begin{pmatrix}\Lambda & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}Q^{\mathrm{T}} & 0\end{pmatrix}=Q\Lambda Q^{\mathrm{T}}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$ with $\lambda_1 \geq \cdots \geq \lambda_r > 0$ and $Q \in \mathbb{R}^{2n \times r}$. To ensure $\widehat{S} + \delta Y \in \mathcal{S}_{2n} \cap \mathcal{E}_n$ while $\mathrm{Null}(\widehat{S} + \delta Y) \supseteq \mathrm{Null}(\widehat{S})$, for some stepsize $\delta > 0$, it suffices to take

$$Y = \begin{pmatrix} Q & 0 \end{pmatrix} \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Q^{\mathrm{T}} & 0 \end{pmatrix} = Q \Delta Q^{\mathrm{T}},$$

where $\Delta \in \mathcal{S}^r \setminus \{0\}$ is a symmetric matrix satisfying

$$\mathrm{Trace}(\Delta) = 0, \quad \langle M_{i,j} M_{i,j}^{\mathrm{T}}, Q \Delta Q^{\mathrm{T}} \rangle = 0, \quad i \in [n].$$

To achieve (II), it suffices to solve the one-dimensional optimization

$$\delta^* = \arg\max_{\delta \geq 0} \Big\{ \delta : \ \lambda_{\min}(\Lambda + \delta \Delta) \geq 0 \Big\}, \tag{B.12}$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a given matrix. the optimization above admits closed-form solution update. Let eigenvalues of Δ be $\lambda'_1 \geq \cdots \geq \lambda'_r$. It suffices to solve

$$\delta^* = \arg\max_{\delta \geq 0} \Big\{ \delta : \ \min_{i \in [r]} (\lambda_i + \delta \lambda'_i) \geq 0 \Big\}.$$

As long as $\lambda'_r \geq 0$, we return $\delta^* = 0$. Otherwise, let i be an index such that $\lambda'_i \geq 0 > \lambda'_{i+1}$. We take $\delta^* = \max_{i<j \leq r} -\frac{\lambda_j}{\lambda'_j}$ as the desired optimal solution.

The overall algorithm is summarized in Algorithm 14. Its performance guarantee is summarized in Propositions 13, 14, and Theorem 11.

B.10 Proof of Theorem 11

The proof of this theorem is separated into two parts.

Proposition 13. *The rank of iteration points in Algorithm 14 strictly decreases. Thus, Algorithm 14 is guaranteed to terminate within* $2n$ *iterations.*

Proof of Proposition 13. Assume on the contrary that $\mathrm{rank}(\widehat{S}(\delta^*)) = \mathrm{rank}(\widehat{S}) = r$. Since

$\widehat{S}(\delta^*) = Q(\Lambda + \delta^*\Delta)Q^{\mathrm{T}}$, the positive eigenvalues of $\widehat{S}(\delta^*)$ are those of the matrix $\Lambda + \delta^*\Delta$. According to the solution structure of (B.12), this could happen only when $\Lambda + \delta^*\Delta \succ 0$, i.e., either $\delta^* = 0$ or $\Delta \succeq 0$. For the first case, this algorithm terminates. For the second case, since Trace$(\Delta) = 0, \Delta \in \mathcal{S}^r$, it implies that $\Delta = 0$, which is a contradiction.

Thus, the rank of the iteration point reduces by at least 1 in each iteration. □

Proposition 14. *Let S^* be the output of Algorithm 14. Then, it holds that*

(I) *S^* is a δ-optimal solution to (SDR).*

(II) *The rank of S^* satisfies*

$$\textit{rank}(S^*) \leq 1 + \left\lfloor \sqrt{2n + \frac{9}{4}} - \frac{3}{2} \right\rfloor.$$

Proof. Recall the solution $\widehat{S}$ obtained from Step 1 of Algorithm 14 satisfies

$$\begin{aligned} F(\widehat{S}) &= \min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} \langle M_{i,j} M_{i,j}^{\mathrm{T}}, \widehat{S} \rangle \\ &= \max_{f,g \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} (f_i + g_i) : \quad f_i + g_j \leq \langle M_{i,j} M_{i,j}^{\mathrm{T}}, \widehat{S} \rangle \right\} \geq \texttt{objval}(SDR) - \delta. \end{aligned}$$

Since Step 2 of Algorithm 14 solves the OT problem exactly, we obtain

$$\frac{1}{n} \sum_{i=1}^{n} (f_i^* + g_i^*) = F(\widehat{S}) \geq \texttt{objval}(SDR) - \delta$$

Since Step 3-7 always finds feasible solutions to the n binding constraints

$$f_i^* + g_{\sigma(i)}^* \leq \langle M_{i,j} M_{i,j}^{\mathrm{T}}, S \rangle, \quad i \in [n],$$

for any iteration points from Step 3-7, denoted as $\widetilde{S}$, the pair $(\widetilde{S}, f^*, g^*)$ is guaranteed to be the δ-optimal solution to (3.14), a reformulation of (SDR). Hence we finish the proof of Part (I).

For the second part, assume on the contrary that $r = \text{rank}(S^*) \geq 1 + \left\lfloor \sqrt{2n + \frac{9}{4}} - \frac{3}{2} \right\rfloor$. It implies $n + 1 < r(r+1)/2$. Recall that Step 6 of Algorithm 14 essentially solves a linear system with $n+1$ constraints and $r(r+1)/2$ variables, so a nonzero matrix Δ is guaranteed to exist. Thus, one can pick a sufficiently small $\delta > 0$ such that $\lambda_{\min}(\Lambda + \delta\Delta) \geq 0$, which contradicts to the termination condition $\delta^* = 0$. Thus, we finish the proof of Part (II). □

Combining both parts, we start to prove Theorem 11.

Proof. Algorithm 14 satisfies the requirement of Theorem 11. For computational complexity, the computational cost of Step 2 of Algorithm 14 is $\mathcal{O}(n^3)$. In each iteration from Step 3-7, the most computationally expansive part is to solve Step 6, which essentially solves a linear system with $n+1$ constraints and $r(r+1)/2$ variables. The conservative bound $r \leq 2n$. Hence, the worst-case computational cost of Step 6 (which can be achieved using Gaussian elimination) is

$$\mathcal{O}((n + 1 + r(r+1)/2) \cdot (n+1)^2) = \mathcal{O}(n^4).$$

Since Algorithm 14 terminates within at most $2n$ iterations, the overall complexity of it is $\mathcal{O}(n^5)$. □

B.11 Numerical Implementation Details

B.11.1 Setup for Computing KMS Wasserstein Distance

When implementing our mirror ascent algorithm, for small sample size ($n \leq 200$), we use the exact algorithm adopted from https://pythonot.github.io/ to solve the inner OT; whereas for large sample size, we use the approximation algorithm adopted from https://github.com/YilingXie27/PDASGD to solve this subproblem. For the baseline `BCD` approach, we implement it using the code from github.com/WalterBabyRudin/KPW_Test/tree/main.

B.11.2 Setup for High-dimensional Hypothesis Testing

For baselines `Sinkhorn Div, SW, MS`, we implement them by calling the well-established package `POT` [113]. For the `ME` baseline, we implement it using the code from https://github.com/wittawatj/interpretable-test.

Next, we outline how to generate the datasets for two-sample testing experiments in Fig. 3.4:

(I) The Gaussian covariance shift dataset was generated by taking

$$\mu = \mathcal{N}(0, I_d), \nu = \mathcal{N}(0, I_d + \rho E),$$

where the dimension $d = 200$, the sample size $n = 200$, and E is the all-one matrix. We vary the hyper-parameter ρ from 0 to 0.06.

(II) The Gaussian mixture dataset was generated by taking

$$\mu = \frac{1}{2}\mathcal{N}(0_d, I_d) + \frac{1}{2}\mathcal{N}(0.5 \cdot 1_d, I_d), \quad \nu = \frac{1}{2}\mathcal{N}(0_d, I_d + 0.05E) + \frac{1}{2}\mathcal{N}(0.5 \cdot 1_d, I_d + 0.05E),$$

where the dimension $d = 40$, and we vary the sample size n from 100 to 700.

(III) For MNIST or CIFAR10 examples, we take μ as the uniform distribution subsampled from the target dataset, denoted as $\mu = p_{\text{data}}$. We take ν as the one having a change of abundance, i.e., $\nu = 0.85 p_{\text{data}} + 0.15 p_{\text{data of label 1}}$, where $p_{\text{data of label 1}}$ corresponds to the distribution of a subset of the class with label 1.

Finally, we outline the procedure for practically using the KMS Wasserstein distance for two-sample testing, based on the train-test split method:

(I) We first do the 50%-50% training-testing data split such that $x^n = x^{\text{Tr}} \cup x^{\text{Te}}$ and $y^n = y^{\text{Tr}} \cup y^{\text{Te}}$.

(II) Then we compute the optimal nonlinear projector f for training data $(x^{\text{Tr}}, y^{\text{Tr}})$. We specify the testing statistic as the Wasserstein distance between projected testing data $(f_{\#}x^{\text{Te}}, f_{\#}y^{\text{Te}})$.

(III) Then we do the permutation bootstrap strategy that shuffles $(x^{\text{Te}}, y^{\text{Te}})$ for many times, e.g., $L = 500$ times. For each time t, the permuted samples are $(x^{\text{Te}}_{(t)}, y^{\text{Te}}_{(t)})$, and we obtain the testing statistic as the Wasserstein distance between projected samples (using the estimated projector f, denoted as $\mathcal{W}(f_{\#}x^{\text{Te}}_{(t)}, f_{\#}y^{\text{Te}}_{(t)})$.

(IV) Finally, we obtain the threshold as $(1-\alpha)$-quantile of the testing statistics for permuted samples, where the type-I error $\alpha = 0.05$.

B.11.3 Setup for Human Activity Detection

The MSRC-12 Kinect gesture dataset contains sequences of human body movements recorded by 20 sensors collected from 30 users performing 12 different gestures. We pre-process the dataset by extracting 10 different users such that in the first 600 timeframes, they are performing the throwing action, and in the remaining 300 timeframes, they are performing the lifting action.

B.11.4 Setup for Generative Modeling

We follow Deshpande et al. [97] to design the optimization algorithm that solves the problem $\min_\theta \mathcal{D}\big(p_{\text{data}}, (f_\theta)_{\#}p_{\text{noise}}\big)$, where p_{data} denotes the empirical distribution of MNIST dataset, p_{noise} denotes the Gaussian noise, and $(f_\theta)_{\#}p_{\text{noise}}$ represents the distribution of the fake image dataset. We specify f_θ as a 4-layer feed-forward neural-net with leaky relu activation, and θ denotes its weight parameters. We train the optimization algorithm in 30 epoches. From Fig. 3.6, we observe that the KMS Wasserstein distance provides fake images that are more closer to the ground truth compared with the Sliced Wasserstein distance.

B.12 Additional Numerical Study

B.12.1 Experiment on Significance Level

Here we add the experiment in the following table to show that as long as we take $\alpha = 0.05$, the practical type-I error of KMS Wasserstein test is guaranteed to be controlled within 0.05.

Table B.2: Type-I Error for two-sample testing with Gaussian mixture dataset.

Method	**20**	**40**	**80**	**160**	**180**	**200**
KMS	0.061 ± 0.015	0.055 ± 0.012	0.043 ± 0.014	0.059 ± 0.011	0.042 ± 0.014	0.062 ± 0.015

B.12.2 Rank Reduction Algorithm

Recall that Theorem 10 provides the rank bound regarding some optimal solution from SDR. In this part, we compare the rank of the matrix estimated from Algorithm 3 with our theoretical rank bound based on the CIFAR10 dataset. For a given positive semidefinite matrix, we calculate the rank as the number of eigenvalues greater than the tolerance `1e-6`. The numerical performance is summarized in Table B.3. In these cases, one can run our rank reduction algorithm to obtain a low-rank solution. Fig. B.2 illustrates the procedure by showing the probability mass values associated with the eigenvectors of the estiamted solution $\widehat{S}$. From the plot, we find that our rank reduction algorithm is capable of producing low-rank solutions even though the matrix size is large.

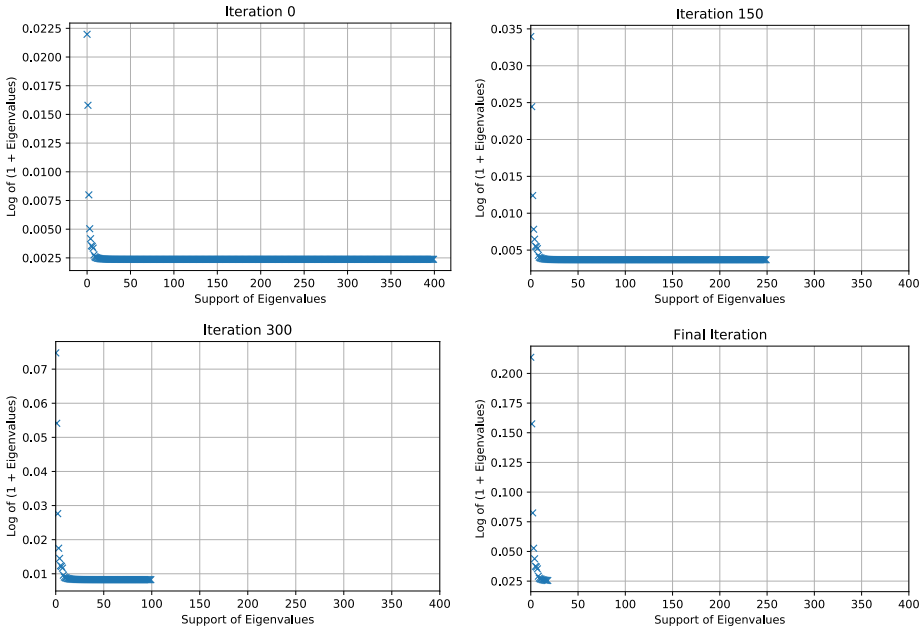

Figure B.2: Procedure of rank reduction algorithm for CIFAR10 example with sample size $n = 200$.

Table B.3: Numerical performance on rank for `CIFAR10` dataset

Sample Size n	**Rank Obtained from Algorithm 3**	**Rank Bound from Theorem 10**
200	400	19
250	500	21
300	600	24
350	700	26
400	800	27
450	900	29
500	1000	31

APPENDIX C

PROOFS AND ADDITIONAL DETAILS OF CHAPTER 4

C.1 ADMM for Solving SDP Problem (4.21)

Define the domain sets

$$\mathcal{C} = \Big\{ Z \in \mathbb{S}^+_{D+1} : \ Z^{(0,0)} = 1, \mathrm{Tr}(Z) = 2 \Big\},$$
$$\mathcal{B} = \left\{ (Z,q) : \sum_{j\in[D]} \left(Z^{(i,j)}\right)^2 \le Z^{(i,i)} q^{(i)}, \left(\sum_{j\in[D]} |Z^{(i,j)}| \right)^2 \le d Z^{(i,i)} q^{(i)}, \forall i \in [D], q \in \overline{\mathcal{Q}} \right\}.$$

Let $\mathcal{I}_{\mathcal{A}}(\cdot)$ denote the indicator function of set $\mathcal{A}$. Then problem (4.21) can be reformulated as

$$\min_{Z,q} \ \Big\{ -\langle \tilde{A}, Z\rangle + \mathcal{I}_{\mathcal{C}}(Z) + \mathcal{I}_{\mathcal{B}}(Z,q) \Big\}.$$

By introducing a new variable Y, the problem above can be written as

$$\min_{Z,Y,q} \ \Big\{ -\langle \tilde{A}, Z\rangle + \mathcal{I}_{\mathcal{C}}(Z) + \mathcal{I}_{\mathcal{B}}(Y,q) : \ Z = Y \Big\}. \tag{C.1}$$

The augmented Lagrangian function for problem (C.1) is defined as

$$\mathcal{L}_\mu(Z,Y,q;\Lambda) = -\langle \tilde{A}, Z\rangle + \mathcal{I}_{\mathcal{C}}(Z) + \mathcal{I}_{\mathcal{B}}(Y,q) - \langle \Lambda, Z - Y\rangle + \frac{1}{2\tau}\|Z-Y\|_F^2,$$

where $\tau > 0$ is a penalty parameter. The ADMM approach produces the following iterations:

$$Z_{k+1} = \arg\min_{Z} \ \mathcal{L}_\mu(Z, Y_k, q_k; \Lambda_k), \tag{C.2a}$$

$$(Y_{k+1}, q_{k+1}) = \arg\min_{Y,q} \ \mathcal{L}_\mu(Z_{k+1}, Y, q; \Lambda_k), \tag{C.2b}$$

$$\Lambda_{k+1} = \Lambda_k - \frac{1}{\tau}\big[Z_{k+1} - Y_{k+1}\big]. \tag{C.2c}$$

The ADMM algorithm terminates at iteration k if for some tolerance parameter $\texttt{tol} > 0$, it holds that

$$\frac{\|Z_{k+1} - Y_{k+1}\|}{1 + \|\tilde{A}\|_1} \leq \texttt{tol}.$$

The advantage of ADMM is that, based on the variable splitting trick, the subproblems (C.2a) and (C.2b) are easier to solve than the original SDP problem.

Specifically, the subproblem (C.2a) reduces to

$$Z_{k+1} = \arg\min_{Z \in \mathcal{C}} \ \left\| Z - (Y_k + \tau\tilde{A} + \tau\Lambda_k) \right\|_F^2, \tag{C.3}$$

which amounts to solving an eigenvalue problem. See the detailed algorithm design in Remark 31.

Next, the subproblem (C.2b) reduces to

$$(Y_{k+1}, q_{k+1}) = \arg\min_{(Y,q) \in \mathcal{B}} \ \|Y - (Z_{k+1} - \tau\Lambda_k)\|_F^2, \tag{C.4}$$

which amounts to solving a large-scale second-order cone program. See the detailed algorithm design in Remark 32.

Remark 31. *Given a symmetric matrix $X \in \mathbb{R}^{(D+1)\times(D+1)}$, consider the optimization problem*

$$\min_{Z \in \mathcal{C}} \ \|Z - X\|_F^2.$$

Since this problem is unitary-invariant, its optimal solution is given by $Z^* = U\,\mathrm{diag}(a^*)U^{\mathrm{T}}$ *for some vector* $a^* \in \mathbb{R}^{D+1}$*, where the matrix* X *admits eigendecomposition* $X = U\,\mathrm{diag}(b)U^{\mathrm{T}}$*. The vector* a^* *can be obtained by solving the following problem:*

$$a^* = \arg\min \left\{ \|a-b\|_2^2 : \quad a \geq 0, a^{(0)} = 1, \sum_{i=0}^{D} a^{(i)} = 2 \right\}. \tag{C.5}$$

Such a problem is a variant of the projection problem onto the simplex in Euclidean space. We adopt the algorithm in [277] with complexity $O(D\log D)$ *to solve this problem. See details in Algorithm 15.*

Algorithm 15 An $O(D\log D)$-complexity algorithm to solving problem (C.5)

1: Sort $b^{(1:D)}$ to $\hat{b}$ such that $\hat{b}^{(1)} \leq \cdots \leq \hat{b}^{(D)}$.

2: Find smallest index $\hat{j}$ such that $\hat{b}^{(j)} - \frac{1}{D-j+1}\left(\sum_{i=j}^{D}\hat{b}^{(i)} - 1\right) > 0$.

3: Compute $\theta = \frac{1}{D-\hat{j}+1}\left(\sum_{i=\hat{j}}^{D}\hat{b}^{(i)} - 1\right)$

4: **Return** vector a such that $a^{(0)} = 1$ and $a^{(i)} = \max\{0, b^{(i)} - \theta\}, i \in [D]$.

Remark 32. *Given a matrix* $X \in \mathbb{R}^{(D+1)\times(D+1)}$*, consider the optimization problem*

$$\min_{(Y,q)\in\mathcal{B}} \|Y - X\|_F^2.$$

It can be reformulated as a second-order cone program that could be solved efficiently based on some off-the-shelf solver:

$$\begin{aligned}
\min_{Y,q\in\overline{\mathcal{Q}},A_i,i\in[D]} \quad & \|vec(Y) - vec(X)\|_2^2 \\
s.t. \quad & \|Y^{(i,:)}\|_1 \leq A_i, i \in [D], \\
& (2A_i, Y^{(i,i)} - dq^{(i)}, Y^{(i,i)} + dq^{(i)}) \in \mathcal{C}_3, Y^{(i,i)} \geq 0, i \in [D], \\
& (Y^{(i,1)}, \ldots, Y^{(i,i)} - \frac{1}{2}q^{(i)}, \ldots, Y^{(i,D)}, \frac{1}{2}q^{(i)}) \in \mathcal{C}_{D+1}, i \in [D],
\end{aligned}$$

where $\mathcal{C}_{D+1}$ denotes the second-order cone of dimension $D+1$:

$$\mathcal{C}_{D+1} = \{(x,t) : \ x \in \mathbb{R}^D, t \in \mathbb{R}, \|x\|_2 \leq t\}.$$

C.2 Proof of Example 2

Proof of Example 2. Note that the population version of the objective in (4.12) becomes

$$F(z) = \mathrm{MMD}^2(\mathcal{N}(0,1), \mathcal{N}(0,(1+\epsilon)^2); k_1) - \lambda \sigma^2_{\mathcal{H}_1}(\mathcal{N}(0,1), \mathcal{N}(0,(1+\epsilon)^2); k_1), \quad \text{if } z^{(1)} \neq 0,$$

when $z = \widehat{z}$ and otherwise $F(z) = 0$. Therefore, taking the variance regularization

$$\lambda \in \left[0, \frac{\mathrm{MMD}^2(\mathcal{N}(0,1), \mathcal{N}(0,(1+\epsilon)^2); k_1)}{\sigma^2_{\mathcal{H}_1}(\mathcal{N}(0,1), \mathcal{N}(0,(1+\epsilon)^2)}\right),$$

achieves the desired result. It remains to compute $\mathrm{MMD}^2(\mathcal{N}(0,1), \mathcal{N}(0,(1+\epsilon)^2); k_1)$ and $\sigma^2_{\mathcal{H}_1}(\mathcal{N}(0,1), \mathcal{N}(0,(1+\epsilon)^2)$ to finish the proof. According to the definition, it holds that

$$\begin{aligned}
&\mathrm{MMD}^2(\mathcal{N}(0,1), \mathcal{N}(0,(1+\epsilon)^2); k_1) \\
=&\mathbb{E}_{x,x' \sim \mathcal{N}(0,1)}\left[k_1(x,x')\right] + \mathbb{E}_{y,y' \sim \mathcal{N}(0,(1+\epsilon)^2)}\left[k_1(y,y')\right] - 2\mathbb{E}_{x \sim \mathcal{N}(0,1), y \sim \mathcal{N}(0,(1+\epsilon)^2)}\left[k_1(x,y)\right] \\
=&\mathbb{E}_{x,x' \sim \mathcal{N}(0,1)}\left[k_1(x,x') + k_1((1+\epsilon)x, (1+\epsilon)x') - 2k_1(x,(1+\epsilon)y)\right] \\
=&\sqrt{\frac{\tau_1^2}{\tau_1^2+2}} + \sqrt{\frac{\tau_1^2}{\tau_1^2+2(1+\epsilon)^2}} - 2\sqrt{\frac{\tau_1^2}{\tau_1^2+1+(1+\epsilon)^2}},
\end{aligned}$$

where the last step is by substituting the expression $k_1(x,y) = e^{-(x-y)^2/(2\tau_1^2)}$ and calculating several integral of exponential functions. Also, we have that

$$\sigma^2_{\mathcal{H}_1}(\mathcal{N}(0,1), \mathcal{N}(0,(1+\epsilon)^2); k_1) = 4\mathbb{E}[H_{1,2}H_{1,3}] - 4\mathrm{MMD}^4(\mathcal{N}(0,1), \mathcal{N}(0,(1+\epsilon)^2); k_1).$$

According to the definition of $H_{i,j}$ in (4.3), it holds that

$$\begin{aligned}
\mathbb{E}[H_{1,2}H_{1,3}] &= \mathbb{E}_{x_1,x_2,x_3,x_4\sim\mathcal{N}(0,1)}\Big[k_1(x_1,x_2)k_1(x_1,x_3) + 2k_1(x_1,x_2)k_1((1+\epsilon)x_3,(1+\epsilon)x_4) \\
&\quad - 4k_1(x_1,x_2)k(x_1,(1+\epsilon)x_3) - 4k_1(x_1,(1+\epsilon)x_2)k_1((1+\epsilon)x_3,(1+\epsilon)x_4) \\
&\quad + 4k_1(x_1,(1+\epsilon)x_2)k_1(x_1,(1+\epsilon)x_3) + k_1((1+\epsilon)x_1,(1+\epsilon)x_2)k_1((1+\epsilon)x_1,(1+\epsilon)x_3)\Big] \\
&= \sqrt{\frac{\tau_1^4}{(\tau_1^2+1)(3+\tau_1^2)}} + \sqrt{\frac{4\tau_1^4}{(\tau_1^2+2)(\tau_1^2+2(1+\epsilon)^2)}} \\
&\qquad - \sqrt{\frac{16\tau_1^4}{2\tau_1^2+1+(1+\epsilon)^2+(1+\tau_1^2)((1+\epsilon)^+\tau_1^2)}} \\
&\quad - \sqrt{\frac{16\tau_1^4}{(\tau_1^2+1+(1+\epsilon)^2)(\tau_1^2+2(1+\epsilon)^2)}} + \sqrt{\frac{16\tau_1^4}{(\tau_1^2+(1+\epsilon)^2)(\tau_1^2+(1+\epsilon)^2+2)}} \\
&\quad + \sqrt{\frac{\tau_1^4}{(\tau_1^2+(1+\epsilon)^2)(\tau_1^2+3(1+\epsilon)^2)}}.
\end{aligned}$$

The proof is completed.

□

C.3 Proofs of Technical Results in Section 4.4

Proof of Theorem 13. A natural combinatorial reformulation of (STRS) is

$$\max_{\substack{S\subseteq[D]:\ |S|\le d\\ z\in\mathbb{R}^D}} \left\{z^{\mathrm{T}}Az + z^{\mathrm{T}}t:\ \|z\|_2 = 1, z^{(k)} = 0, \forall k \notin S\right\}. \tag{C.6}$$

Given a size-d set $S \subseteq [D]$, and problem parameters $A \in \mathbb{S}_D, t \in \mathbb{R}^D$, it holds that

$$\max_{z\in\mathbb{R}^D} \left\{z^{\mathrm{T}}Az + z^{\mathrm{T}}t:\ \|z\|_2 = 1, z^{(k)} = 0, \forall k \notin S\right\} = \max_{z\in\mathbb{R}^d} \left\{z^{\mathrm{T}}A^{(S,S)}z + z^{\mathrm{T}}t^{(S)}:\ \|z\|_2 = 1\right\}. \tag{C.7}$$

Next, we linearize the problem (C.7) using the auxiliary variable defined as

$$Z = \begin{pmatrix}1\\ z\end{pmatrix}\begin{pmatrix}1\\ z\end{pmatrix}^{\mathrm{T}} = \begin{pmatrix}1 & z^{\mathrm{T}}\\ z & zz^{\mathrm{T}}\end{pmatrix}$$

and the matrix

$$\tilde{A}^{(S,S)} = \begin{pmatrix}0 & \frac{1}{2}(t^{(S)})^{\mathrm{T}}\\ \frac{1}{2}t^{(S)} & A^{(S,S)}\end{pmatrix}.$$

Assume the index of Z and $\tilde{A}^{(S,S)}$ is over $\{0, 1, \ldots, d\}^2$. Then we equivalently reformulate the problem (C.7) as

$$\begin{aligned}\max_{Z\in\mathbb{S}^+_{d+1}}\quad & \langle\tilde{A}^{(S,S)}, Z\rangle\\ \text{s.t.}\quad & \mathrm{rank}(Z) = 1,\\ & Z^{(0,0)} = 1, \mathbf{Tr}(Z) = 2.\end{aligned} \tag{C.8}$$

In particular, constraints $Z \succeq 0, \mathrm{rank}(Z) = 1, Z^{(0,0)} = 1$ together imply that

$$Z = \begin{pmatrix}1 & z^{\mathrm{T}}\\ z & zz^{\mathrm{T}}\end{pmatrix}$$

for some vector $z \in \mathbb{R}^d$, and the condition $\mathrm{Tr}(Z) = 2$ implies $\|z\|_2 = 1$. By [195, Corollary 3], we further obtain the following equivalent reformulation of problem (C.7) when dropping the nonconvex rank constraint $\mathrm{rank}(Z) = 1$:

$$\begin{aligned} \max_{Z \in \mathbb{S}^+_{d+1}} \quad & \langle \tilde{A}^{(S,S)}, Z \rangle \\ \text{s.t.} \quad & Z^{(0,0)} = 1, \mathrm{Tr}(Z) = 2. \end{aligned} \tag{C.9}$$

In summary, we obtain the following reformulation of (STRS):

$$\begin{aligned} \max_{Z \in \mathbb{S}^+_{d+1}, S \subseteq [D]:\ |S| \le d} \quad & \langle \tilde{A}^{(S,S)}, Z \rangle \\ \text{s.t.} \quad & Z^{(0,0)} = 1, \mathrm{Tr}(Z) = 2. \end{aligned} \tag{C.10}$$

It remains to show the equivalence between formulations (4.16) and (C.10). We only need to show for any feasible $q \in \mathcal{Q}$ with its support $S := \{k :\ q^{(k)} = 1\}$, it holds that

$$\begin{aligned} & \max_{Z \in \mathbb{S}^+_{D+1}} \left\{ \langle \tilde{A}, Z \rangle :\ Z_{i,i} \le q^{(i)}, i \in [D], Z^{(0,0)} = 1, \mathrm{Tr}(Z) = 2 \right\} \\ = & \max_{Z \in \mathbb{S}^+_{d+1}} \left\{ \langle \tilde{A}^{(S,S)}, Z \rangle :\ Z_{0,0} = 1, \mathrm{Tr}(Z) = 2 \right\}. \end{aligned} \tag{C.11}$$

Since $Z \in \mathbb{S}^+_{D+1}$ is a positive semi-definite matrix, the condition $Z^{(i,i)} = 0$ for $i \in [D] \setminus S$ implies

$$Z^{(i,j)} = 0, \quad \forall (i,j) \notin S \times S.$$

Leveraging this property, we check the relation (C.11) indeed holds true.

□

Proof of Corollary 2. It suffices to verify the following two valid inequalities hold for

problem (4.16):

$$\sum_{j\in[D]} (Z^{(i,j)})^2 \le Z^{(i,i)} q^{(i)}, \quad \forall i \in [D], \tag{C.12}$$

$$\left(\sum_{j\in[D]} |Z^{(i,j)}| \right)^2 \le d Z^{(i,i)} q^{(i)}, \quad \forall i \in [D]. \tag{C.13}$$

This verification step follows a similar argument in [194, Lemma 2].

□

Proof of Theorem 14. We first re-write $f(q)$ as the optimal value to the following optimization problem:

$$\begin{aligned}
\max_{Z\in\mathbb{S}^+_{D+1}, U\ge 0, Y, y, t} \quad & \langle \tilde{A}, Z \rangle \\
& Z^{(0,0)} = 1, && [\lambda_0] \\
& \mathrm{Tr}(Z) = 2, && [\lambda] \\
& \sum_j U^{(i,j)} \le y^{(i)}, \quad \forall i \in [D], && [\beta^{(i)}] \\
& -U^{(i,j)} \le Z^{(i,j)} \le U^{(i,j)}, \quad \forall i,j \in [D], && [W_1^{(i,j)}, W_2^{(i,j)}] \\
& \|(y_i; t_i)\|_2 \le \frac{1}{2} Z^{(i,i)} + \frac{d}{2} q^{(i)}, \quad \forall i \in [D], && [\nu_1^{(i)}] \\
& t_i = \frac{1}{2} Z^{(i,i)} - \frac{d}{2} q^{(i)}, \quad \forall i \in [D], && [\nu_2^{(i)}] \\
& Y^{(i,:)} = Z^{(i,:)} - \frac{1}{2} q^{(i)} e_i, \quad \forall i \in [D], && [\Lambda^{(i,:)}] \\
& \left\| Y^{(i,:)} \right\|_2 \le \frac{1}{2} q^{(i)}, \quad \forall i \in [D]. && [\mu^{(i)}]
\end{aligned}$$

Here, we associate dual variables with primal constraints in brackets. In detail, constraints corresponding to $[\beta^{(i)}], [W_1^{(i,j)}, W_2^{(i,j)}], [\nu_1^{(i)}], [\nu_2^{(i)}]$ are reformulation of the valid inequality (C.13), and constraints corresponding to $[\Lambda^{(i,:)}]$ and $[\mu^{(i)}]$ are second-order conic reformulation of the valid inequality (C.12).

Its Lagrangian dual reformulation becomes

$$\min_{\substack{\lambda,\lambda_0,\nu_2,\Lambda\\ \beta,W_1,W_2,\nu_1,\mu\geq 0}} \max_{Z\in\mathbb{S}^+_{D+1},U\geq 0,Y,y,t} \quad \langle \tilde{A}, Z\rangle + \lambda_0\big(1-Z^{(0,0)}\big) + \lambda\big(2-\mathrm{Tr}(Z)\big) + \sum_i \beta^{(i)}\big[y^{(i)} - \sum_j U^{(i,j)}\big]$$
$$+ \sum_{i,j} W_1^{(i,j)}\big[U^{(i,j)} + Z^{(i,j)}\big] + \sum_{i,j} W_2^{(i,j)}\big[U^{(i,j)} - Z^{(i,j)}\big] + \sum_i \nu_1^{(i)}\left[\frac{1}{2}Z^{(i,i)} + \frac{d}{2}q^{(i)} - \|(y_i;t_i)\|_2\right]$$
$$+ \sum_i \nu_2^{(i)}\left(\frac{1}{2}Z^{(i,i)} - \frac{d}{2}q^{(i)} - t_i\right) + \sum_i \Lambda^{(i,:)}\left[Z^{(i,:)} - \frac{1}{2}q^{(i)}e_i - Y^{(i,:)}\right] + \sum_i \mu^{(i)}\left(\frac{1}{2}q^{(i)} - \left\|Y^{(i,:)}\right\|_2\right).$$

Or equivalently, it can be written as

$$\min_{\substack{\lambda,\lambda_0,\nu_2,\Lambda\\ \beta,W_1,W_2,\nu_1,\mu\geq 0}} \left\{\lambda_0 + 2\lambda + \frac{d}{2}(\nu_1-\nu_2)^{\mathrm{T}}q + \frac{1}{2}(\mu - \mathrm{diag}(\Lambda))^{\mathrm{T}}q\right.$$
$$+ \max_{Z\in\mathbb{S}^+_{D+1}} \left\{\langle \tilde{A}, Z\rangle - \lambda_0 Z^{(0,0)} - \lambda\mathrm{Tr}(Z) + \sum_{i,j}(W_1^{(i,j)} - W_2^{(i,j)} + \Lambda^{(i,j)})Z^{(i,j)} + \frac{1}{2}\sum_i(\nu_1^{(i)} + \nu_2^{(i)})Z^{(i,i)}\right\}$$
$$+ \max_{U\geq 0} \left\{-\sum_i \beta^{(i)}\sum_j U^{(i,j)} + \sum_{i,j}(W_1^{(i,j)} + W_2^{(i,j)})U^{(i,j)}\right\} + \max_Y \left\{-\sum_i \Lambda^{(i,:)}Y^{(i,:)} - \sum_i \mu^{(i)}\|Y^{(i,:)}\|_2\right\}$$
$$\left. + \max_{y,t} \left\{\sum_i \beta^{(i)}y^{(i)} - \sum_i \nu_1^{(i)}\|(y_i;t_i)\|_2 - \sum_i \nu_2^{(i)}t_i\right\}\right\}.$$

The inner maximization over Z can be simplified into the constraint

$$\begin{pmatrix} -\lambda_0 & \frac{1}{2}t^{\mathrm{T}} \\ \frac{1}{2}t & A - \lambda I_D + W_1 - W_2 + \Lambda + \frac{1}{2}\,\mathrm{diag}(\nu_1+\nu_2)\end{pmatrix} \preceq 0.$$

The inner maximization over U can be simplified as

$$W_1 + W_2 - \mathrm{diag}(\beta) \leq 0.$$

The inner maximization over Y can be simplified as

$$\sum_j (\Lambda^{(i,j)})^2 \leq (\mu^{(i)})^2, \quad i\in[D].$$

The inner maximization over (y, t) can be simplified as

$$(\beta^{(i)})^2 + (\nu_2^{(i)})^2 \leq (\nu_1^{(i)})^2, \quad i \in [D].$$

Combining those relations, we arrive at the dual problem

$$\begin{aligned}
\min_{\substack{\lambda,\lambda_0,\nu_2,\Lambda \\ \beta,W_1,W_2,\nu_1,\mu\geq 0}} \quad & \lambda_0 + 2\lambda + q^{\mathrm{T}} \left[\frac{d}{2}(\nu_1 - \nu_2) + \frac{1}{2}(\mu - \operatorname{diag}(\Lambda))\right] \\
& \begin{pmatrix} -\lambda_0 & \frac{1}{2}t^{\mathrm{T}} \\ \frac{1}{2}t & A - \lambda I_D + W_1 - W_2 + \Lambda + \frac{1}{2}\operatorname{diag}(\nu_1 + \nu_2) \end{pmatrix} \preceq 0, \\
& W_1 + W_2 - \operatorname{diag}(\beta) \leq 0, \\
& \sum_j (\Lambda^{(i,j)})^2 \leq (\mu^{(i)})^2, \quad i \in [D], \\
& (\beta^{(i)})^2 + (\nu_2^{(i)})^2 \leq (\nu_1^{(i)})^2, \quad i \in [D].
\end{aligned}$$

□

Proof of Theorem 15. The left-hand-side relation is easy to show. The proof for the right-hand-side relation is separated into two parts:

- optval(4.21) $\leq \|t\|_2 + d \cdot \{$optval(4.16) $- \min_k |t[k]|\}$;
- optval(4.21) $\leq \|t\|_2 + D/d \cdot$ optval(4.16).

(I) For any feasible solution (q, Z) to (4.21), we find

$$\begin{aligned}
\sum_i t^{(i)} Z^{(0,i)} &\leq \sum_i |t^{(i)}||Z^{(0,i)}| \leq \sum_i |t^{(i)}| \sqrt{Z^{(0,0)} Z^{(i,i)}} \\
&= \sum_i |t^{(i)}| \sqrt{Z^{(i,i)}} \leq \left(\sum_i |t^{(i)}|^2\right)^{1/2} \left(\sum_i Z^{(i,i)}\right)^{1/2} = \|t\|_2,
\end{aligned}$$

where the first inequality is due to taking absolute values, the second inequality is

because $Z \succeq 0$ and $|Z^{(0,i)}| \leq \sqrt{Z^{(0,0)} Z^{(i,i)}}$, and the last inequality is by the Cauchy-Schwarz inequality.

As a consequence, for any feasible solution (q, Z) to (4.21), it holds that

$$\langle \tilde{A}, Z \rangle = \sum_{i,j} A^{(i,j)} Z^{(i,j)} + \sum_{i} t^{(i)} Z^{(0,i)} \leq \sum_{i,j} \left|A^{(i,j)}\right| \left|Z^{(i,j)}\right| + \|t\|_2. \tag{C.14}$$

On the other hand, it is easy to verify that for any $i \in [D]$, the following is a feasible solution to (4.16):

$$Z_i = \begin{pmatrix} 1 \\ e_i \end{pmatrix} \begin{pmatrix} 1 \\ e_i \end{pmatrix}^{\mathrm{T}} \quad \text{or} \quad Z_i = \begin{pmatrix} 1 \\ -e_i \end{pmatrix} \begin{pmatrix} 1 \\ -e_i \end{pmatrix}^{\mathrm{T}},$$

where e_i is a basis vector with the i-th element being 1. This yields

$$\mathsf{optval}(4.16) \geq \max\left\{A^{(i,i)} + t^{(i)}, A^{(i,i)} - t^{(i)}\right\} = A^{(i,i)} + |t^{(i)}|, \quad \forall i \in [D].$$

Therefore, we obtain

$$A^{(i,i)} \leq \mathsf{optval}(4.16) - |t^{(i)}| \leq \mathsf{optval}(4.16) - \min_{i \in [D]} |t^{(i)}|,$$

and $|A^{(i,j)}| \leq \sqrt{A^{(i,i)} A^{(j,j)}} \leq \mathsf{optval}(4.16) - \min_{i \in [D]} |t^{(i)}|$ for any $i, j \in [D]$. Combining this expression with (C.14) implies that

$$\langle \tilde{A}, Z \rangle \leq \sum_{i,j} \left|Z^{(i,j)}\right| \cdot \left(\max_{i,j} \left|A^{(i,j)}\right|\right) + \|t\|_2 \leq \sum_{i,j} \left|Z^{(i,j)}\right| \cdot \left(\mathsf{optval}(4.16) - \min_{i \in [D]} |t^{(i)}|\right) + \|t\|_2. \tag{C.15}$$

Also, because of the valid inequality $\left(\sum_j \left|Z^{(i,j)}\right|\right)^2 \leq d Z^{(i,i)} q^{(i)}$, it holds that

$$\sum_{i,j} \left|Z^{(i,j)}\right| \leq \sqrt{d} \sum_{i} \sqrt{Z^{(i,i)} q^{(i)}} \leq \sqrt{d} \left(\sum_i Z^{(i,i)}\right)^{1/2} \left(\sum_i q^{(i)}\right)^{1/2} = d.$$

Combining this relation with (C.15) gives the desired result.

(II) For any feasible solution (Z, q) in (4.16), we enforce $Z^{(0,i)} = Z^{(i,0)} = 0$ for $i \in [D]$, then the updated solution is still feasible, with the associated objective value

$$\langle Z^{([D],[D])}, A\rangle.$$

Therefore, we obtain the relation

$$\mathsf{optval}(4.16) \geq \max_{Z\in\mathbb{S}^+_D, q\in\mathcal{Q}} \left\{\langle Z, A\rangle :\ Z^{(i,i)} \leq q^{(i)}, i \in [D], \mathrm{Tr}(Z) = 1\right\} \geq d/D\cdot\lambda_{\max}(A), \tag{C.16}$$

where the last inequality is due to [194, Proposition 2 and proof of Theorem 5].

For any feasible solution (Z, q) in (4.21), according to Part (I), it holds that $\sum_i t^{(i)} Z^{(0,i)} \leq \|t\|_2$, and therefore

$$\begin{aligned}\langle \tilde{A}, Z\rangle &= \sum_{i,j} A^{(i,j)} Z^{(i,j)} + \sum_i t^{(i)} Z^{(0,i)} \leq \langle A, Z^{([D],[D])}\rangle + \|t\|_2 \\ &\leq \max_{Z\succeq 0, \mathrm{Tr}(Z)=1} \langle A, Z\rangle + \|t\|_2 = \lambda_{\max}(A) + \|t\|_2.\end{aligned}$$

Combining this relation with (C.16) gives the desired result.

□

Proof of Theorem 16(I). Let $z_* = \sum_i y^{(i)} e_i$ be the optimal solution of (STRS), where e_i is

the i-th basis vector. Then it holds that

$$\begin{aligned}
\textsf{optval(STRS)} &= \sum_i y^{(i)} \left[e_i^{\mathrm{T}} (A z_* + t) \right] \\
&\leq \sqrt{\sum_i (y^{(i)})^2} \sqrt{\sum_i (e_i^{\mathrm{T}} (A z_* + t))^2} \\
&\leq \sqrt{d} \max_i \; e_i^{\mathrm{T}} (A z_* + t) \\
&\leq \sqrt{d} \max_i \; \left\{ \max_{z \in \mathcal{Z}} \; e_i^{\mathrm{T}} (A z + t) \right\} \\
&= \sqrt{d} \max_i \; \left\{ e_i^{\mathrm{T}} (A \hat{z}_i + t) \right\},
\end{aligned}$$

where the last equality is because

$$\hat{z}_i = \arg\max_{z \in \mathcal{Z}} \; e_i^{\mathrm{T}} (A z) = \arg\max_{z \in \mathcal{Z}} \; e_i^{\mathrm{T}} (A z + t).$$

Based on the observation above, one can assert that there exists $i \in [D]$ such that

$$\sqrt{d} e_i^{\mathrm{T}} (A \hat{z}_i + t) \geq \textsf{optval(STRS)}. \tag{C.17}$$

Next, we provide the lower bound for $V_{(\mathrm{I})}$:

$$\begin{aligned}
V_{(\mathrm{I})} &= \max_i \; \max \left(e_i^{\mathrm{T}} A e_i + e_i^{\mathrm{T}} t, \hat{z}_i^{\mathrm{T}} A \hat{z}_i + \hat{z}_i^{\mathrm{T}} t \right) \\
&\geq \max_i \; \left\{ \max \left(e_i^{\mathrm{T}} A e_i, \hat{z}_i^{\mathrm{T}} A \hat{z}_i \right) + \min \left(e_i^{\mathrm{T}} t, \hat{z}_i^{\mathrm{T}} t \right) \right\} \\
&\geq \max_i \; \left\{ e_i^{\mathrm{T}} A \hat{z}_i + \min \left(e_i^{\mathrm{T}} t, \hat{z}_i^{\mathrm{T}} t \right) \right\} \\
&= \max_i \; \left\{ e_i^{\mathrm{T}} (A \hat{z}_i + t) + \min \left(0, (\hat{z}_i - e_i)^{\mathrm{T}} t \right) \right\} \\
&\geq \frac{1}{\sqrt{d}} \textsf{optval(STRS)} + \max_i \; \min \left(0, (\hat{z}_i - e_i)^{\mathrm{T}} t \right) \\
&\geq \frac{1}{\sqrt{d}} \textsf{optval(STRS)} - 2 \|t\|_{(d+1)},
\end{aligned}$$

where the second inequality is because $A \succeq 0$ and $0 \leq (e_i - \hat{z}_i)^{\mathrm{T}} A (e_i - \hat{z}_i) = (e_i^{\mathrm{T}} A e_i +$

$\hat{z}_i^{\mathrm{T}} A\hat{z}_i) - 2e_i^{\mathrm{T}} A\hat{z}_i$, i.e.,

$$e_i^{\mathrm{T}} A\hat{z}_i \leq \frac{1}{2}(e_i^{\mathrm{T}} Ae_i + \hat{z}_i^{\mathrm{T}} A\hat{z}_i) \leq \max\left(e_i^{\mathrm{T}} Ae_i, \hat{z}_i^{\mathrm{T}} A\hat{z}_i\right),$$

the third inequality is due to (C.17), and the last inequality is because $\hat{z}_i - e_i$ is a $(d+1)$-sparse vector with $\|\hat{z}_i - e_i\|_2 = 2$, and

$$\max_i \ \min\left(0, (\hat{z}_i - e_i)^{\mathrm{T}} t\right) \geq -\max_i \max_{a:\ \|a\|_0 \leq d+1, \|a\|_2 \leq 2} a^{\mathrm{T}} t = -2\|t\|_{(d+1)}.$$

The proof is completed.

□

Proof of Theorem 16(II). By [1, Theorem 1.1], the primal-dual pair (v, λ) of the trust region subproblem satisfies the following:

$$\begin{cases} (A - \lambda I)v = -t \\ A \preceq \lambda I \\ \|v\|_2 \leq 1 \\ \lambda(1 - \|v\|_2) = 0 \end{cases}$$

Let $\bar{z}$ be the d-sparse truncation of v. Then it holds that

$$\begin{aligned} z^{\mathrm{T}} Av + t^{\mathrm{T}} z &= z^{\mathrm{T}} Av + z^{\mathrm{T}}(-A + \lambda I)v \\ &= \lambda z^{\mathrm{T}} v = \lambda z^{\mathrm{T}} \bar{z} = \lambda \|\bar{z}\|_2 \geq \lambda\sqrt{\frac{d}{D}}. \end{aligned}$$

On the other hand,

$$z^{\mathrm{T}} Av + t^{\mathrm{T}} z \leq \sqrt{z^{\mathrm{T}} Az}\sqrt{v^{\mathrm{T}} Av} + t^{\mathrm{T}} x \leq \sqrt{z^{\mathrm{T}} Az} \cdot \left(\lambda - t^{\mathrm{T}} v\right)^{1/2} + t^{\mathrm{T}} z,$$

where the last inequality is because $(A - \lambda I)v = -t$ and therefore

$$v^{\mathrm{T}} A v + t^{\mathrm{T}} v = \lambda \|v\|_2^2 \leq \lambda.$$

By re-arrangement, it holds that

$$\sqrt{\frac{d}{D}}\lambda \leq \sqrt{z^{\mathrm{T}} A z} \cdot \left(\lambda - t^{\mathrm{T}} v\right)^{1/2} + t^{\mathrm{T}} z.$$

Or equivalently, the dual multiplier λ satisfies

$$\frac{d}{D}\lambda^2 - \left[2\sqrt{\frac{d}{D}} z^{\mathrm{T}} t + z^{\mathrm{T}} A z\right] \lambda + (z^{\mathrm{T}} t)^2 + (z^{\mathrm{T}} A z)(v^{\mathrm{T}} t) \leq 0.$$

Consequently,

$$\frac{d}{D}\lambda^2 - \left[2\sqrt{\frac{d}{D}} \|t\|_{(d)} + z^{\mathrm{T}} A z\right] \lambda - (z^{\mathrm{T}} A z)\|t\| \leq 0.$$

The determinant of the quadratic function on the left-hand-side above is non-negative:

$$\Delta := \left[2\sqrt{\frac{d}{D}} \|t\|_{(d)} + z^{\mathrm{T}} A z\right]^2 + \frac{4d}{D} z^{\mathrm{T}} A z \|t\| \geq 0.$$

On the other hand,

$$\sqrt{\Delta} \leq z^{\mathrm{T}} A z + 2\sqrt{\frac{d}{D}} \|t\|_{(d)} + \frac{2d}{D}\|t\|_2.$$

Hence, we find the upper bound of λ:

$$\begin{aligned}
\lambda &\leq \frac{2\sqrt{\frac{d}{D}} \|t\|_{(d)} + z^{\mathrm{T}} A z + \sqrt{\Delta}}{2\frac{d}{D}} \\
&\leq \frac{D}{d} z^{\mathrm{T}} A z + \|t\|_2 + \sqrt{\frac{D}{d}} \|t\|_{(d)} \\
&\leq \frac{D}{d} V_{(\mathrm{II})} + \|t\|_2 + \left(\sqrt{\frac{D}{d}} + \frac{D}{d}\right) \|t\|_{(d)}.
\end{aligned}$$

This, together with the fact that $\lambda \geq v^{\mathrm{T}}Av + t^{\mathrm{T}}v \geq$ optval(STRS) completes the proof.

□

Proof of Theorem 17. Define the following two sets:

$$T_d := \{z \in \mathbb{R}^D : \ \|z\|_2 \leq 1, \|z\|_1 \leq \sqrt{d}\},$$
$$S_d := \{z \in \mathbb{R}^D : \ \|z\|_2 \leq 1, \|z\|_0 \leq d\}.$$

It has been shown in [100, Lemma 1] that there exists a factor $\rho \in (1, 1 + \sqrt{d/(d+1)}]$ such that

$$T_d \subseteq \rho \cdot \mathrm{conv}(S_d).$$

It follows that

$$\begin{aligned}\text{optval(4.22)} \leq & \max_{z \in \rho \cdot \mathrm{conv}(S_d)} \{z^{\mathrm{T}}Az + z^{\mathrm{T}}t\} = \max_{z \in \mathrm{conv}(S_d)} \{\rho^2 z^{\mathrm{T}}Az + \rho z^{\mathrm{T}}t\} \\ \leq & \max_{z \in \mathrm{conv}(S_d)} \{\rho^2 z^{\mathrm{T}}Az + \rho^2 z^{\mathrm{T}}t\} = \rho^2 \cdot \max_{z \in \mathrm{conv}(S_d)} \{z^{\mathrm{T}}Az + z^{\mathrm{T}}t\} \\ = & \rho^2 \text{optval(STRS)},\end{aligned}$$

where the second inequality follows from the fact that in the optimal solution to the problem $\max_{z \in \mathrm{conv}(S_d)} \{\rho^2 z^{\mathrm{T}}Az + \rho z^{\mathrm{T}}t\}$, we have that $z^{\mathrm{T}}t \geq 0$. The last equality is because the objective function $z^{\mathrm{T}}Az + z^{\mathrm{T}}t$ is a convex function.

□

Proof of Proposition 2. The proof of this proposition is a simple extension from [100].

□

C.4 Extension of Technical Results in Section 4.5 for a Generic Kernel

We first make the following assumptions regarding the kernel choice $K_z(\cdot,\cdot)$, variance regularization value λ, and data distributions μ, ν.

Assumption 10. *The kernel $K_z(\cdot,\cdot)$ is uniformly bounded and satisfies the Lipshitz continuous condition, i.e., for any $z, z' \in \mathcal{Z}, x, y \in \Omega$, it holds that $|K_z(x,y)| \leq M$ and $|K_z(x,y) - K_{z'}(x,y)| \leq L\|z - z'\|_2$.*

Assumption 11. *Under the alternative hypothesis $\mathcal{H}_1 : \mu \neq \nu$, there exists $\lambda \geq 0$ such that for some $\zeta \in \mathcal{Z}$, it holds that $\mathrm{MMD}^2(\mu,\nu;K_\zeta) > 0$ and*

$$\Delta_\zeta \triangleq \mathrm{MMD}^2(\mu,\nu;K_\zeta) - \lambda\big[\max_{z\in\mathcal{Z}} \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z) - \min_{z\in\mathcal{Z}} \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\big] > 0. \tag{C.18}$$

Here $\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)$ denotes the population version of the empirical variance estimator defined in (4.11)*.*

Assumption 10 is a standard assumption used in the statistical analysis of kernel-based testing in literature. Assumption 11 is imposed to ensure the expected value of the testing statistic is strictly positive. It is worth noting that this assumption is not too restrictive. In the following, we demonstrate that, under mild conditions, our proposed kernels in (4.5)-(4.7) indeed satisfy Assumptions 10 and 11.

Proposition 15 (Sufficient Condition of Assumptions 10 and 11)**.** (I) *(Linear Kernel) For the kernel in* (4.5)*, Assumption 10 is guaranteed to hold with $M = \sqrt{d}$ and $L = \sqrt{2d}$. As long as there exists $s^* \in [D]$ such that $\mathrm{Proj}_{s^*\#}\mu \neq \mathrm{Proj}_{s^*\#}\nu$, Assumption 11 is guaranteed to hold with*

$$\lambda \in \left[0, \frac{\max_{z\in\mathcal{Z}} \sum_{s\in[D]} z^{(s)} \mathrm{MMD}^2(\mathrm{Proj}_{s\#}\mu, \mathrm{Proj}_{s\#}\nu; k_s)}{16d}\right).$$

(II) *(Quadratic Kernel) For the kernel in* (4.6)*, Assumption 10 is guaranteed to hold with*

$M = 2c^2 + 2d$ *and* $L = 4d + 2c\sqrt{2d}$*. As long as there exists* $s^* \in [D]$ *such that* $\mathrm{Proj}_{s^*\#}\mu \neq \mathrm{Proj}_{s^*\#}\nu$ *or* $(\mathcal{A}(\mu,\nu))^{(s^*,s^*)} > 0$ *holds, Assumption 11 is guaranteed to hold with*

$$\lambda \in \left[0, \frac{\max_{z\in\mathcal{Z}} z^{\mathrm{T}}\mathcal{A}(\mu,\nu)z + z^{\mathrm{T}}\mathcal{T}(\mu,\nu)}{16(2d+2c^2)^2}\right).$$

(III) *(Gaussian Kernel) For the kernel in* (4.7)*, if additionally assuming that* $\Omega \subseteq \{x \in \mathbb{R}^D : \|x\|_\infty \leq R\}$*, Assumption 10 is guaranteed to hold with* $M = 1$ *and* $L = \frac{2R}{\sigma\sqrt{e}}$*. As long as there exists* $S \subseteq [D]$ *with* $|S| \leq d$ *such that* $\mathrm{Proj}_{S\#}\mu \neq \mathrm{Proj}_{S\#}\nu$*, Assumption 11 is guaranteed to hold with*

$$\lambda \in \left[0, \frac{\max_{z\in\mathcal{Z}} \mathrm{MMD}^2(\mu,\nu;K_z)}{16}\right).$$

Proposition 16 (Non-asymptotic Concentration Properties)**.** *Under Assumption 10, with probability at least* $1-\delta$*, (i) the bias approximation error can be bounded as*

$$\begin{aligned}\sup_{z\in\mathcal{Z}}\left|S^2(\mathbf{x}^n,\mathbf{y}^n;K_z) - \mathrm{MMD}^2(\mu,\nu;K_z)\right| &\leq \epsilon^1_{n,\delta} \triangleq \frac{8}{\sqrt{n}}\left[M\sqrt{2\log\binom{D}{d}\frac{2}{\delta} + 2d\log(4\sqrt{n})} + L\right] \\ &= O\left(\frac{1}{\sqrt{n}}\left[M\cdot\left(d(\log n + \log\frac{D}{d} + \log\frac{1}{\delta})\right)^{1/2} + L\right]\right),\end{aligned}$$

where $O(\cdot)$ *hides constants that are independent to parameters* D, d, n, M, L*.*
(ii) and the variance approximation error can be bounded as

$$\begin{aligned}\sup_{z\in\mathcal{Z}}\left|\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z) - \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| &\leq \epsilon^2_{n,\delta} \triangleq \frac{64}{\sqrt{n}}\left[7\sqrt{2\log\binom{D}{d}\frac{2}{\delta} + 2d\log(4\sqrt{n})} + \frac{18M^2}{\sqrt{n}} + 8LM\right] \\ &= O\left(\frac{1}{\sqrt{n}}\cdot\left[LM + \frac{M^2}{\sqrt{n}} + \left(d(\log n + \log\frac{D}{d} + \log\frac{1}{\delta})\right)^{1/2}\right]\right),\end{aligned}$$

where $\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z) \triangleq \mathbb{E}_{\mathbf{x}^n\sim\mu,\mathbf{y}^n\sim\nu}[\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z)]$.

Proof of the proposition above follows similar covering number arguments in [202, Theorem 6]. The main difference is that when applying union bound on the set $\mathcal{Z}$, the corresponding error bound is sharper because the covering number of sparse-constrained set

$\mathcal{Z}$ is much smaller.

C.5 Proofs of Technical Results in Section 4.5

Proof of Proposition 15. (I) We first verify the boundness and Lipschitz continuity conditions. Specifically, it holds that

$$|K_z(x,y)| \leq \sum_{s\in[D]} |z^{(s)}| \leq \max_{z\in\mathcal{Z}} \sum_{s\in[D]} |z^{(s)}| \leq \max_{z\in\mathbb{R}^d:\ \|z\|_2=1} \sum_{s\in[d]} |z^{(s)}| \leq \sqrt{d},$$

where the first inequality is because $|k_s(x,y)| \leq 1$ for any $x, y \in \mathbb{R}$, and the third inequality is because any vector in $\mathcal{Z}$ only has at most d non-zero entries. Next, we find

$$\begin{aligned} |K_z(x,y) - K_{z'}(x,y)| &= \left| \sum_{s\in[D]} (z^{(s)} - (z')^{(s)}) k_s(x^{(s)}, y^{(s)}) \right| \\ &\leq \sum_{s\in[D]} |z^{(s)} - (z')^{(s)}| = \|z - z'\|_1 \\ &\leq \sqrt{2d}\|z - z'\|_2, \end{aligned}$$

where the first inequality is because $|k_s(x,y)| \leq 1$ for any $x, y \in \mathbb{R}, s \in [D]$, the second inequality is because the vector $z - z'$ only has at most $2d$ non-zero entries. The remaining of Part (I) can be proved by noting that

$$\mathrm{MMD}^2(\mu, \nu, K_{\bar{z}}) \leq \max_{z\in\mathcal{Z}} \sum_{s\in[D]} z^{(s)} \mathrm{MMD}^2(\mathrm{Proj}_{s\#}\mu, \mathrm{Proj}_{s\#}\nu; k_s)$$

and

$$\max_{z\in\mathcal{Z}} \left|\sigma^2_{\mathcal{H}_1}(\mu, \nu; K_z)\right| \leq 8d.$$

(II) For quadratic kernel, we find

$$|K_z(x,y)| \le 2\left(\sum_{s\in[D]} z^{(s)} k_s(x^{(s)}, y^{(s)})\right)^2 + 2c^2 \le 2d + 2c^2,$$

where the first inequality is based on the relation $(a+b)^2 \le 2a^2 + 2b^2$, and the second inequality is because in Part (I) we have shown that $|\sum_{s\in[D]} z^{(s)} k_s(x^{(s)}, y^{(s)})| \le \sqrt{d}$.

Besides, it holds that

$$|K_z(x,y) - K_{z'}(x,y)| = \left|\sum_{s\in[D]} (z^{(s)} - (z')^{(s)}) k_s(x^{(s)}, y^{(s)})\right| \left|\sum_{s\in[D]} (z^{(s)} + (z')^{(s)}) k_s(x^{(s)}, y^{(s)}) + 2c\right|.$$

Recall the first term on the right-hand-side can be bounded by $\sqrt{2d}\|z - z'\|_2$, and the second term can be upper bounded by a constant:

$$\begin{aligned}
&\left|\sum_{s\in[D]} (z^{(s)} + (z')^{(s)}) k_s(x^{(s)}, y^{(s)}) + 2c\right| \\
\le &\sum_{s\in[D]} |z^{(s)} + (z')^{(s)}||k_s(x^{(s)}, y^{(s)})| + 2c \\
\le &\sum_{s\in[D]} |z^{(s)} + (z')^{(s)}| + 2c \\
\le &\max_{v:\ \|v\|_0 \le 2d, \|v\|_2 \le 2} \|v\|_1 + 2c \le 2\sqrt{2d} + 2c.
\end{aligned}$$

Combining those two relations gives the desired result. The remaining of Part (II) can be proved by noting that

$$\mathrm{MMD}^2(\mu, \nu, K_{\bar{z}}) \le \max_{z\in\bar{\mathcal{Z}}} \max_{z\in\bar{\mathcal{Z}}} z^{\mathrm{T}} \mathcal{A}(\mu,\nu) z + z^{\mathrm{T}} \mathcal{T}(\mu,\nu)$$

and

$$\max_{z\in\bar{\mathcal{Z}}} \left|\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| \le 8(2d + 2c^2)^2.$$

(III) The boundness of the Gaussian kernel is easy to check. The Lipscthiz continuity condition of the Gaussian kernel follows from [202, Lemma 20]. The remaining of Part (III) can be proved by noting that

$$\max_{z\in\mathcal{Z}} \left|\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| \leq 8.$$

□

Before showing the proof of Theorem 16, we list two useful technical lemmas.

Lemma 19 ([138, Theorem 10])**.** *Assume the kernel $K_z(\cdot,\cdot)$ is uniformly bounded, i.e., for any $z\in\mathcal{Z}, x, y\in\Omega$, it holds that $|K_z(x,y)|\leq M$. For fixed $z\in\mathcal{Z}$, with probability at least $1-\delta$,*

$$\left|S^2(\mathbf{x}^n,\mathbf{y}^n;K_z)-\mathrm{MMD}^2(\mu,\nu;K_z)\right| \leq \frac{16M}{\sqrt{2n}}\sqrt{\log\frac{2}{\delta}}.$$

Lemma 20 ([202, Lemma 17 and 18])**.** *Assume the kernel $K_z(\cdot,\cdot)$ is uniformly bounded, i.e., for any $z\in\mathcal{Z}, x, y\in\Omega$, it holds that $|K_z(x,y)|\leq M$. For fixed $z\in\mathcal{Z}$, with probability at least $1-\delta$,*

$$\left|\hat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z)-\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| \leq 448\sqrt{\frac{2}{n}\log\frac{2}{\delta}}+\frac{1152M^2}{n}.$$

Proof of Theorem 16. We first consider an ϵ-cover of $\mathcal{Z}$, denoted as $\{z_i\}_{i\in[T]}$. According to the definition of $\mathcal{Z}$, it can be shown that $T\leq\binom{D}{d}(4/\epsilon)^d$. Applying the union bound regarding the concentration inequality in Lemma 19, we obtain with probability at least $1-\delta$,

$$\max_{z\in\{z_i\}_{i\in[T]}}\left|S^2(\mathbf{x}^n,\mathbf{y}^n;K_z)-\mathrm{MMD}^2(\mu,\nu;K_z)\right| \leq \frac{16M}{\sqrt{2n}}\sqrt{\log\frac{2T}{\delta}}.$$

For any $z\in\mathcal{Z}$, there exists z' from $\{z_i\}_{i\in[T]}$ such that $\|z-z'\|_2\leq\epsilon$. Based on the Lipschitz assumption regarding the kernel function, we find with probability at least $1-\delta$, it holds

that

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}} \left|S^2(\mathbf{x}^n,\mathbf{y}^n;K_z) - \mathrm{MMD}^2(\mu,\nu;K_z)\right| \\
\leq &\max_{z\in\{z_i\}_{i\in[T]}} \left|S^2(\mathbf{x}^n,\mathbf{y}^n;K_z) - \mathrm{MMD}^2(\mu,\nu;K_z)\right| + 8L\epsilon \\
\leq &\frac{16M}{\sqrt{2n}}\sqrt{\log\frac{2T}{\delta}} + 8L\epsilon \leq \frac{16M}{\sqrt{2n}}\sqrt{\log\binom{D}{d}\frac{2}{\delta} + d\log\frac{4}{\epsilon}} + 8L\epsilon.
\end{aligned}$$

Setting $\epsilon = 1/\sqrt{n}$ gives the desired result.

Next, applying the union bound regarding the concentration inequality in Lemma 20, we obtain with probability at least $1-\delta$,

$$\max_{z\in\{z_i\}_{i\in[T]}} \left|\hat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z) - \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| \leq 448\sqrt{\frac{2}{n}\log\frac{2T}{\delta}} + \frac{1152M^2}{n}.$$

Similar as in the first part, we find with probability at least $1-\delta$, it holds that

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}} \left|\hat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z) - \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| \\
\leq &\max_{z\in\{z_i\}_{i\in[T]}} \left|\hat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z) - \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| + 512LM\epsilon \\
\leq &448\sqrt{\frac{2}{n}\log\frac{2T}{\delta}} + \frac{1152M^2}{n} + 512LM\epsilon \\
\leq &448\sqrt{\frac{2}{n}\log\binom{D}{d}\frac{2}{\delta} + \frac{2}{n}d\log\frac{4}{\epsilon}} + \frac{1152M^2}{n} + 512LM\epsilon.
\end{aligned}$$

Also, setting $\epsilon = 1/\sqrt{n}$ gives the desired result. □

Proof of Theorem 4. To simplify notation, let us define the population version of the objective in (4.12) as follows:

$$F^*(z) = \mathrm{MMD}^2(\mu,\nu;K_z) - \lambda\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z).$$

We first derive the lower bound of $F^*(\hat{z}_{\mathrm{Tr}})$ in terms of $F^*(\bar{z})$ with $\bar{z}$ defined in Assumption 11

using concentration analysis. It is clear that $|F^*(z) - F(z)| \leq \epsilon_{n,\delta/2}$ with probability at least $1-\delta$. As a consequence, with probability at least $1-\delta$, it holds that

$$F^*(\hat{z}_{\mathrm{Tr}}) \geq F(\hat{z}_{\mathrm{Tr}}) - \epsilon_{n_{\mathrm{Tr}},\delta/4} \geq F(\bar{z}) - \epsilon_{n_{\mathrm{Tr}},\delta/4} \geq F^*(\bar{z}) - 2\epsilon_{n_{\mathrm{Tr}},\delta/4}, \tag{C.19}$$

where we use this observation in the first and last inequalities, and the second inequality is because of the sub-optimality of $\bar{z}$. Now we are ready to show part (I) of this theorem. By definition, we find

$$\begin{aligned}\mathbb{E}[T_{n_{\mathrm{Te}}}] &= \mathrm{MMD}^2(\mu,\nu;K_{\hat{z}_{\mathrm{Tr}}}) \\ &= F^*(\hat{z}_{\mathrm{Tr}}) + \lambda\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_{\hat{z}_{\mathrm{Tr}}}) \geq F^*(\hat{z}_{\mathrm{Tr}}) + \lambda \min_{z\in\mathcal{Z}} \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z).\end{aligned}$$

Combining the relation above and (C.19) implies that, with probability at least $1-\delta$, it holds that

$$\mathbb{E}[T_{n_{\mathrm{Te}}}] \geq F^*(\bar{z}) - 2\epsilon_{n_{\mathrm{Tr}},\delta/4} + \lambda \min_{z\in\mathcal{Z}} \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z) = \Delta_{\bar{z}} - 2\epsilon_{n_{\mathrm{Tr}},\delta/4}.$$

The second part of this theorem follows from [138, Theorem 12].

□

Lemma 21 (Asymptotics of Inverse Error Function [106])**.** *Denote by $S(x)$ the inverse of the error function*

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\,\mathrm{d}t.$$

As $x \to 1$, it holds that

$$S(x) \to \sqrt{\mathcal{LW}\left(\frac{1}{2\pi(x-1)^2}\right)},$$

where $\mathcal{LW}(x)$ denotes the function Lambert $W(x)$ admitting the series expansion

$$\mathcal{LW}(x) = \sum_{n\geq 1} \frac{(-1)^{n-1}}{n!} x^n.$$

Specifically, $\mathcal{LW}(x) \to \ln(x) - \ln\ln(x)$ *as* $x \to \infty$.

Proof of Theorem 19. It is worth noting that

$$\begin{aligned}\mathbb{P}(T_{n_{\mathrm{Te}}} > t_{\mathrm{thres}}) = \mathbb{P}\left(T_{n_{\mathrm{Te}}} > \frac{\tau}{n_{\mathrm{Te}}}\right) &= 1 - \mathbb{P}\left(\frac{\sqrt{n_{\mathrm{Te}}}(T_{n_{\mathrm{Te}}} - \mathbb{E}T_{n_{\mathrm{Te}}})}{\sigma_{\mathcal{H}_1}} \leq \frac{\tau}{\sigma_{\mathcal{H}_1}\sqrt{n_{\mathrm{Te}}}} - \frac{\sqrt{n_{\mathrm{Te}}}\mathbb{E}T_{n_{\mathrm{Te}}}}{\sigma_{\mathcal{H}_1}}\right) \\ &\geq 1 - \Phi\left(\frac{\tau}{\sigma_{\mathcal{H}_1}\sqrt{n_{\mathrm{Te}}}} - \frac{\sqrt{n_{\mathrm{Te}}}\mathbb{E}T_{n_{\mathrm{Te}}}}{\sigma_{\mathcal{H}_1}}\right) - \frac{C\rho}{\sigma_{\mathcal{H}_1}^3\sqrt{n_{\mathrm{Te}}}},\end{aligned}$$

where for the inequality above we apply the Berry–Esseen theorem to argue that the distribution of $\sqrt{n_{\mathrm{Te}}}(T_{n_{\mathrm{Te}}} - \mathbb{E}T_{n_{\mathrm{Te}}})/\sigma_{\mathcal{H}_1}$ can be approximated as the normal distribution with residual error $O(1/\sqrt{n_{\mathrm{Te}}})$. Therefore, as long as we ensure that

$$\frac{\tau}{\sigma_{\mathcal{H}_1}\sqrt{n_{\mathrm{Te}}}} - \frac{\sqrt{n_{\mathrm{Te}}}\mathbb{E}T_{n_{\mathrm{Te}}}}{\sigma_{\mathcal{H}_1}} \leq \Phi^{-1}(\epsilon) \Longleftrightarrow \mathbb{E}T_{n_{\mathrm{Te}}} \geq \frac{\tau}{n_{\mathrm{Te}}} + \frac{\Phi^{-1}(1-\epsilon)}{\sqrt{n_{\mathrm{Te}}}},$$

it holds that the testing power is lower bounded:

$$\mathbb{P}(T_{n_{\mathrm{Te}}} > t_{\mathrm{thres}}) \geq 1 - \epsilon - \frac{C\rho}{\sigma_{\mathcal{H}_1}^3\sqrt{n_{\mathrm{Te}}}}.$$

Taking $\epsilon = 1/\sqrt{n_{\mathrm{Te}}}$ and applying the asymptotic formula on the inverse cdf $\Phi^{-1}(\cdot)$ in Lemma 21 gives the desired result. The type-I risk upper bound follows a similar argument.

□

APPENDIX D

PROOFS AND ADDITIONAL DETAILS OF CHAPTER 5

D.1 Detailed Experiment Setup

Unless stated otherwise, we solved the SAA, Wasserstein DRO, and KL-divergence DRO baseline models exactly using the off-the-shelf solver Mosek [7]. Optimization hyperparameters, such as step size, maximum iterations, and number of levels, were tuned to minimize training error after 10 outer iterations. We use RT-MLMC subgradient estimator to solve the Sinkhorn DRO model. We employed the *warm starting* strategy during the iterative procedure: we set the initial guess of parameter θ at the beginning of outer iteration as the one obtained from the SAA approach. At other outer iterations, the initial guess of parameter θ is set to be the final obtained solution θ at the last outer iteration. The following subsections outline some special reformulations, optimization algorithms used to solve the baseline models.

D.1.1 Setup for Newsvendor Problem and Running Time

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

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

D.1.2 Setup for Mean-risk Portfolio Optimization

From [226, Eq. (27)] we can see that the 1-Wasserstein DRO formulation with radius ρ for the portfolio optimization problem becomes

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

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

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

In particular, the inner subproblem has the following reformulation:

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

Hence, the 2-Wasserstein DRO can be reformulated as

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

D.1.3 Setup for Adversarial Multi-class Logistic Regression

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

(I) For a given classifer B and data sample $(x, \boldsymbol{y})$, the ℓ_p-norm ($p \in \{1, 2\}$) adversarial attack based on projected gradient method [212] iterates as follows: $x_0 \leftarrow x$ and

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

We perform the gradient update above for 15 steps with initial learning rate $\alpha = 1$. When $p = 1$, the radius of attack $\xi \in \{\texttt{0}, \texttt{3e-3}, \texttt{6e-3}, \texttt{9e-3}, \texttt{1.2e-2}\} \cdot \varrho$; and when $p = 2$, the radius $\xi \in \{\texttt{0}, \texttt{8e-3}, \texttt{1.6e-2}, \texttt{2.4e-2}, \texttt{3.2e-2}\} \cdot \varrho$.

(II) For a given feature vector x, the perturbed feature using white Laplacian noise becomes $x + \xi \cdot \zeta$, where the random vector ζ follows the isotropic Laplace distribution with zero mean and unit variance. The ratio $\xi \in \{\texttt{0}, \texttt{2e-3}, \texttt{4e-3}, \texttt{6e-3}, \texttt{8e-3}\} \cdot \varrho$. Similarly, the perturbed feature using white Gaussian noise becomes $x + \xi \cdot \zeta$, with ζ being the isotropic Gaussian distribution with zero mean and unit variance. In this case, the ratio $\xi \in \{\texttt{0}, \texttt{5e-2}, \texttt{1e-1}, \texttt{1.5e-1}, \texttt{2e-1}\} \cdot \varrho$.

In this example, we use stochastic gradient methods to solve the SAA formulation and all penalized DRO formulations. We terminate the training of SAA or DRO models when the number of epoches, i.e., the number of times for processes each training sample, exceeds 30. In is worth mentioning that the Wasserstein DRO model with a fixed Lagrangian multiplier λ using samples $\{x_i, \boldsymbol{y}_i\}_{i=1}^n$ can be reformulated as

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

D.2 Additional Validation Experiments

D.2.1 Comparison of Optimization Algorithms: Linear Regression

To examine the performance of different (sub)gradient estimators, we study the problem of distributionally robust linear regression (see the setup in Example 4). We take the nominal distribution $\widehat{\mathbb{P}}$ as the empirical one based on samples $\{(a_i, b_i)\}_{i=1}^n$. As a consequence, the inner objective function in (5.12) 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 $\Theta = \{\theta : \ \|\theta\|_2^2 \leq 0.999 \cdot \frac{\lambda}{2}\}$. Similar to the setup in [192, Section 5.1], we examine the performance using three LIBSVM regression real world datasets [67]: housing, mg, and mpg.

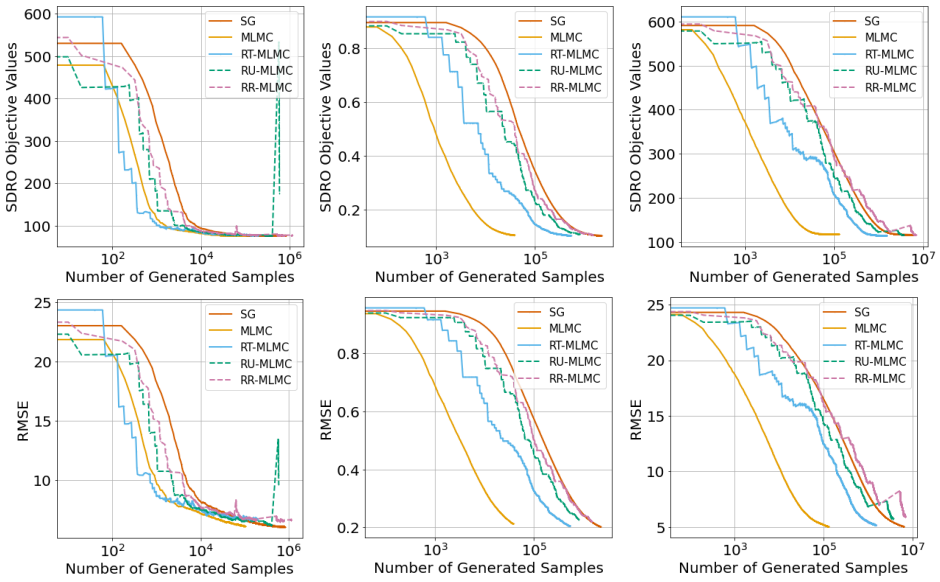

Figure D.1: Comparison results of SG, (V-)MLMC, RT-MLMC, RU-MLMC, and RR-MLMC on robust linear regression problem in terms of sample complexties from $\widehat{\mathbb{P}}$ and $\mathbb{Q}_{x,\epsilon}$. From left to right, the figures correspond to three different regression datasets: (a) housing; (b) mg; and (c) mpg. From top to bottom, the figures correspond to plots of (a) Sinkhorn DRO objective values; and (b) RMSE of obtained solutions.

The quality of proposed gradient estimators is examined in a single BSMD step with

specified hyper-parameters $(\lambda, \epsilon) = (10^3, 10^{-1})$. For baseline comparison, we examine the SG, RT-MLMC estimators together with the (V-)MLMC, RU-MLMC, and RR-MLMC estimators that have been proposed in [156]. We have validated in Theorem 21 that both SG and RT-MLMC estimators have convergence guarantees for smooth and nonsmooth loss functions, whereas SG estimator has slower convergence rate. The (V-)MLMC estimator only have convergence guarantees for smooth loss functions, and RU-MLMC/RR-MLMC estimators do not have convergence guarantees as their (sub)gradient second-order moments are unbounded.

For a given solution θ, we quantify its performance using the corresponding Sinkhorn DRO objective value. Besides, we report its root-mean-square error (RMSE) on training data. Thus, the smaller those two performance criteria are, the smaller the solution's optimization performance has. Fig. D.1 shows the performance of various gradient estimators in terms of the number of generated samples from $\widehat{\mathbb{P}}$ and $\mathbb{Q}_{x,\epsilon}, x \in \mathrm{supp}\widehat{\mathbb{P}}$ based on these criteria. The results demonstrate that the SG scheme does not perform competitively, as expected from our theoretical analysis, which shows that SG has the worst complexity order. In contrast, using other four types of MLMC methods lead to faster convergence behavior. While the RU-MLMC and RR-MLMC schemes exhibit competitive performance, the optimization procedure shows some oscillations. One possible explanation is that the variance values of those gradient estimators are unbounded, making these two approaches unstable.

D.2.2 Comparison of Optimization Algorithms: Portfolio Optimization

In this subsection, we validate the competitive performance of RT-MLMC gradient estimator on the case where the loss is convex and nonsmooth, and we try to solve the 2-SDRO formulation. We consider the portfolio optimization problem, and specify instances $(n, d) = (50, 50), (100, 100), (400, 400)$. We quantify the performance of obtained solution using the Sinkhorn DRO objective value. Since in this problem setup no analytical expression of the objective value is available, we estimate the objective value using (5.18) with hyper-

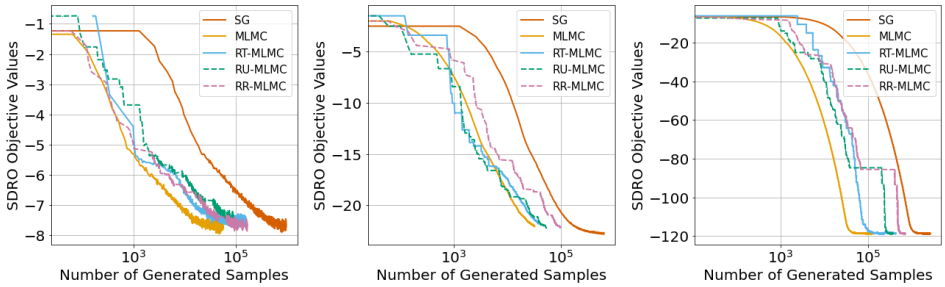

Figure D.2: Comparison results of SG, (V-)MLMC, RT-MLMC, RU-MLMC, and RR-MLMC on portfolio optimization problem. From left to right, plots correspond to three different instances of $(n, d) \in \{(50, 50), (100, 100), (400, 400)\}$.

parameters $L = 8$ and $n_L^o = 10^3$. Fig. D.2 shows the performance in terms of the number of generated samples based on this criterion. The results demonstrate that even for nonsmooth loss function, those listed MLMC-based gradient estimators have better performance than the SG estimator. Besides, the proposed RT-MLMC and standard (V-)MLMC schemes have comparable performance, and in some cases (V-)MLMC estimator even has better performance. It is an open question that whether the (V-)MLMC estimator will have the similar performance guarantees as the RT-MLMC estimator for convex nonsmooth optimization, which can be a topic for future study.

D.2.3 Comparison of Running Time for Different Baselines

The computational time for the newsvendor problem in Section 5.5.1 is reported in Table D.1. We observe that the training time of 2-Wasserstein DRO model increases quickly as the sample size increases, while the training time of other DRO models increases mildly in the training sample size.

The computational time for the portfolio optimization problem in Section 5.5.2 is reported in Table D.2. We observe that the computational time of 1- or 2-SDRO model increases mildly as the problem input size increases. Also, SDRO models do not have the smallest computational time in general. The reason is that in this example, other DRO models have tractable finite-dimensional conic programming formulations so that off-the-shelf software can solve them efficiently. In contrast, Sinkhorn DRO models do not have

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

Model	Exponential			Gamma			Gaussian Mixture		
	$n = 10$	$n = 30$	$n = 100$	$n = 10$	$n = 30$	$n = 100$	$n = 10$	$n = 30$	$n = 100$
SAA	4.11e-3	4.66e-3	4.67e-3	3.96e-3	4.57e-3	5.81e-3	3.82e-3	4.60e-3	4.79e-3
KL-DRO	6.92e-3	8.17e-3	1.15e-2	8.07e-3	8.24e-3	1.16e-2	7.77e-3	8.47e-3	1.12e-2
1-SDRO	8.77e-2	8.88e-2	1.03e-1	2.76e-2	3.40e-2	4.72e-2	2.90e-2	3.13e-2	4.50e-2
2-WDRO	1.68e00	5.67e00	2.71e01	1.72e00	5.63e00	2.77e01	1.51e00	5.47e00	2.84e01
2-SDRO	3.16e-2	3.77e-2	5.92e-2	2.64e-2	2.95e-2	5.02e-2	2.57e-2	3.10e-2	4.87e-2

special reformulation, but they can still be solved in a reasonable amount of time.

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

(n, d) Values	SAA	KL-DRO	1-WDRO	1-SDRO	2-WDRO	2-SDRO
$(30, 30)$	6.76e-03	1.42e-02	7.80e-03	4.91e-02	8.95e-03	5.00e-02
$(50, 30)$	7.31e-03	1.84e-02	8.33e-03	1.87e-01	1.11e-02	5.88e-02
$(100, 30)$	8.99e-03	2.95e-02	1.03e-02	2.78e-01	1.12e-02	6.00e-02
$(150, 30)$	1.12e-02	4.14e-02	1.21e-02	2.80e-01	1.22e-02	6.95e-02
$(200, 30)$	1.12e-02	5.66e-02	1.35e-02	2.99e-01	1.48e-02	7.67e-02
$(400, 30)$	1.89e-02	6.45e-02	2.09e-02	2.99e-01	2.30e-02	1.62e-01
$(100, 5)$	5.76e-03	1.46e-02	6.79e-03	1.05e-01	7.62e-03	5.40e-02
$(100, 10)$	6.18e-03	1.70e-02	7.70e-03	1.08e-01	8.73e-03	5.55e-02
$(100, 20)$	7.43e-03	1.82e-02	8.41e-03	1.12e-01	9.44e-03	5.58e-02
$(100, 40)$	9.87e-03	3.25e-02	1.13e-02	1.16e-01	1.18e-02	5.70e-02
$(100, 80)$	1.31e-02	6.48e-02	1.56e-02	1.19e-01	1.68e-02	5.72e-02
$(100, 100)$	1.54e-02	7.00e-02	1.87e-02	1.22e-01	1.93e-02	5.73e-02

The computational time of adversarial multi-class classification problem in Section 5.5.3 is reported in Table D.4, with the basic statistics of classification datasets presented in Table D.3. The results indicate that Sinkhorn DRO models have shorter computational time than Wasserstein DRO models in general. Note that we solve all baseline methods with stochastic algorithms. For large-scale datasets optimizing the log-sum-exp type loss for Sinkhorn DRO seems to be more efficient than solving the minimax game formulation for Wasserstein DRO.

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

	MNIST	CIFAR-10	tinyImageNet	STL-10
Image Size (before pre-processing)	784	3072	12288	27648
Feature Dimension (after pre-processing)	512	512	512	512
# of classes	10	10	200	10
Training Size	50000	50000	90000	5000
Testing Size	10000	10000	10000	8000

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

Dataset	SAA	KL-DRO	1-WDRO	1-SDRO	2-WDRO	2-SDRO
MNIST	37.2	60.1	154	94.1	166	84.0
CIFAR-10	31.6	51.7	133	98.3	140	80.6
tinyImageNet	58.1	102	248	153	259	143
STL-10	3.42	5.15	13.5	10.1	14.2	8.61

D.2.4 Coefficient of Prescriptiveness for Different Parameter(s) Combination

In this subsection, we report the coefficient of prescriptiveness for different parameter(s) combination on instances which are omitted in the main content. Specifically,

- Fig. D.3 and D.4 correspond to the omitted experiment results in Section 5.5.1.
- Fig. D.5 and D.6 correspond to the omitted experiment results in Section 5.5.2.
- Fig. D.7 corresponds to the omitted experiment results in Section 5.5.3.

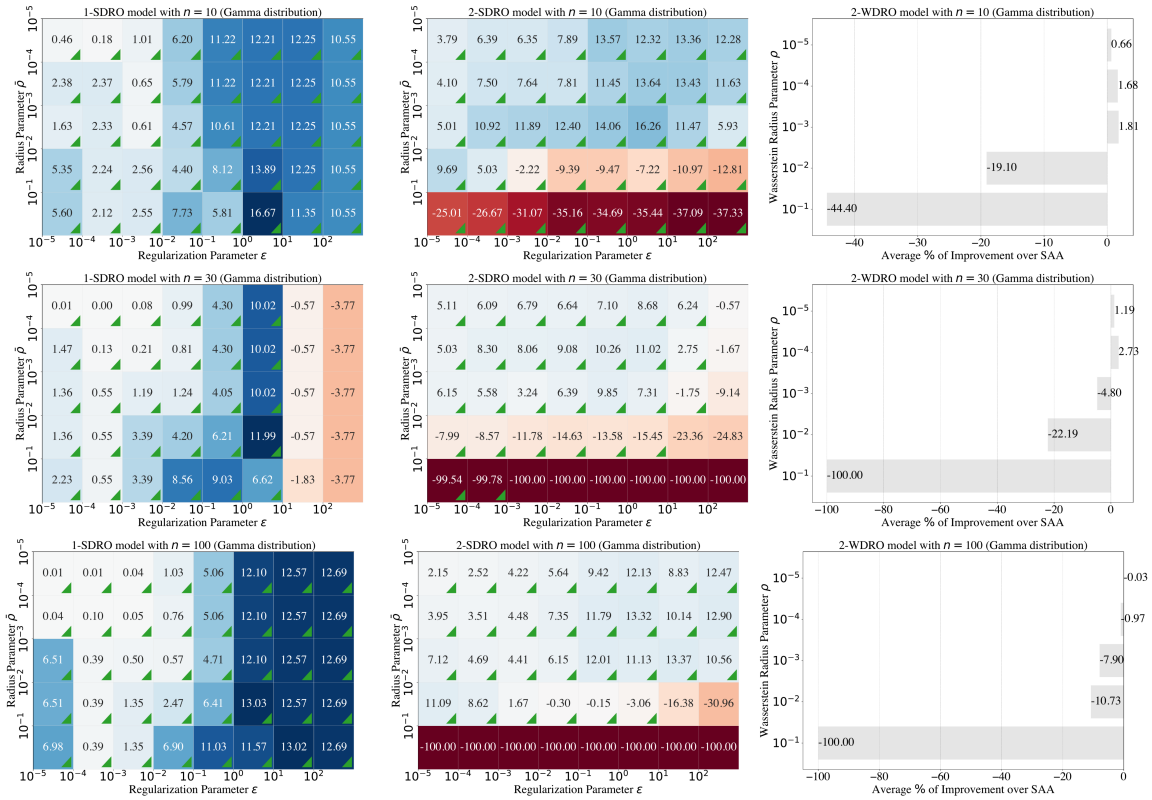

Figure D.3: Experiment results of the newsvendor model for gamma data distribution. Details of these subplots follow the same setup from Figure 5.4.

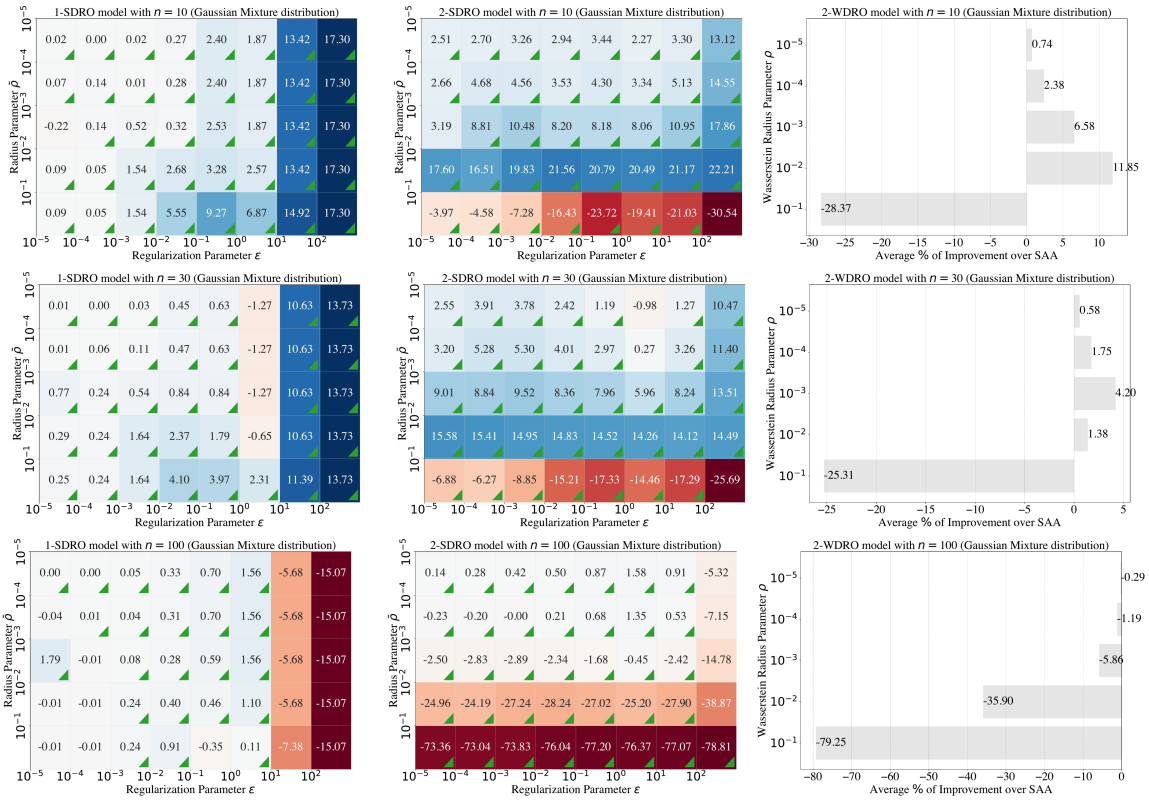

Figure D.4: Experiment results of the newsvendor model for the mixture of truncated normal distributions. Details of these subplots follow the same setup from Figure 5.4.

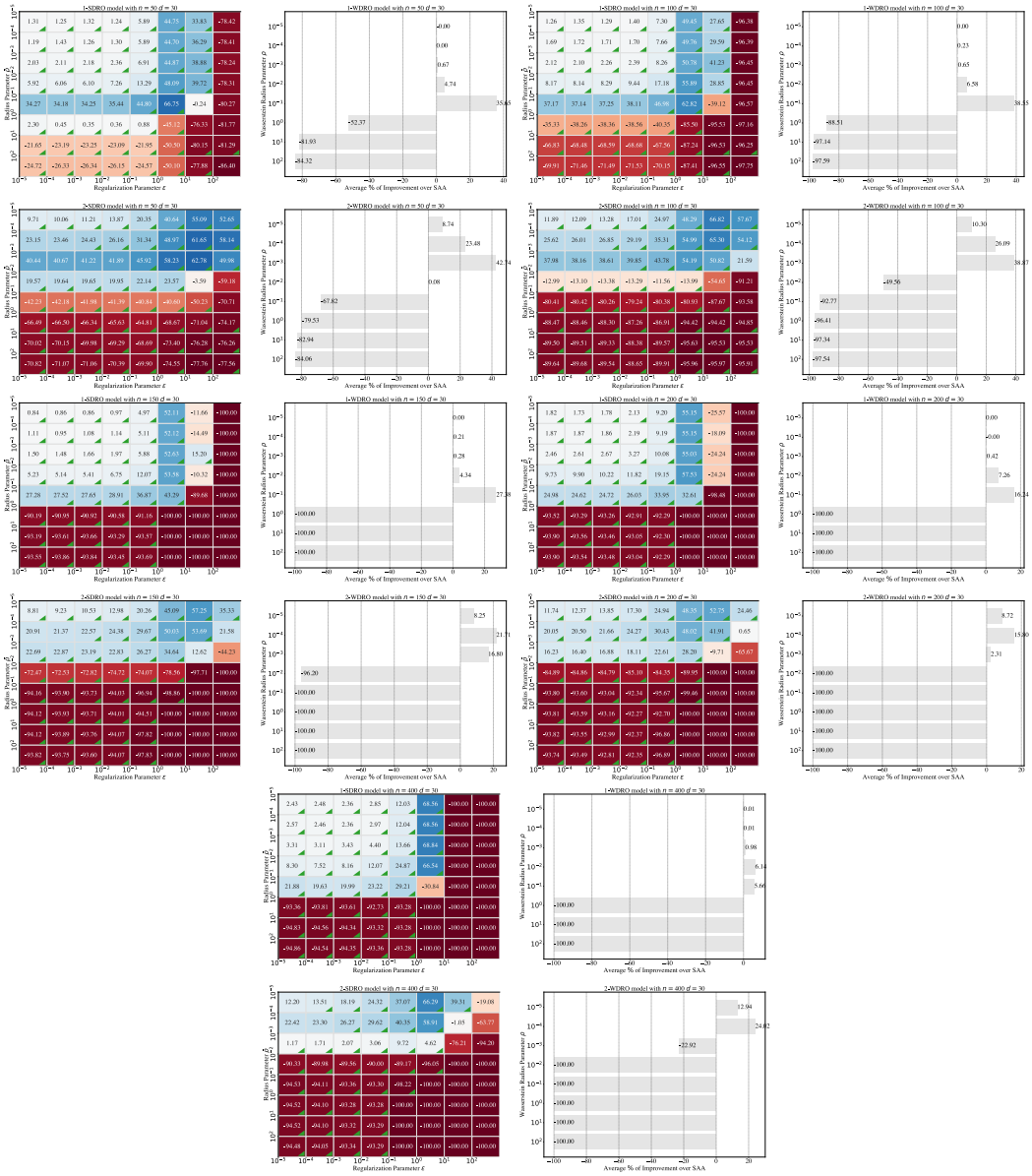

Figure D.5: Additional experiment results of the portfolio optimization model for different data dimensions in heatmaps. Here we fix the data dimension $d = 30$ and vary the sample size $n \in \{50, 100, 150, 200, 400\}$. Details of these subplots follow the same setup from Fig. 5.6. (a) $(n, d) = (50, 30)$; (b) $(n, d) = (100, 30)$; (c) $(n, d) = (150, 30)$; (d) $(n, d) = (200, 30)$; (e) $(n, d) = (400, 30)$.

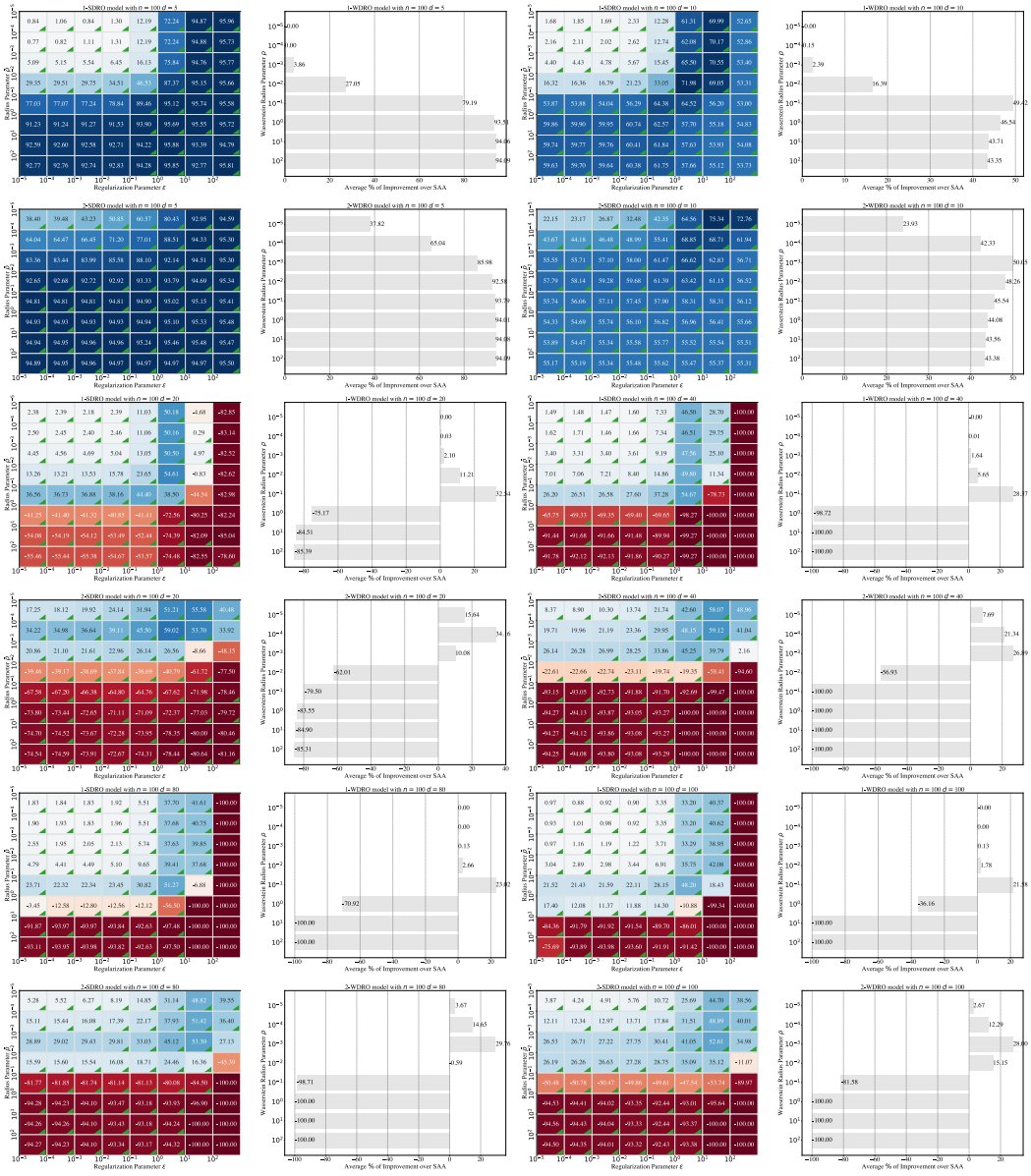

Figure D.6: Additional experiment results of the portfolio optimization model for different data dimensions in heatmaps. Here we fix the sample size $n = 100$ and vary the data dimension $d \in \{5, 10, 20, 40, 80, 100\}$. Details of these subplots follow the same setup from Fig. 5.6. (a) $(n, d) = (100, 5)$; (b) $(n, d) = (100, 10)$; (c) $(n, d) = (100, 20)$; (d) $(n, d) = (100, 40)$; (e) $(n, d) = (100, 80)$; (f) $(n, d) = (100, 100)$.

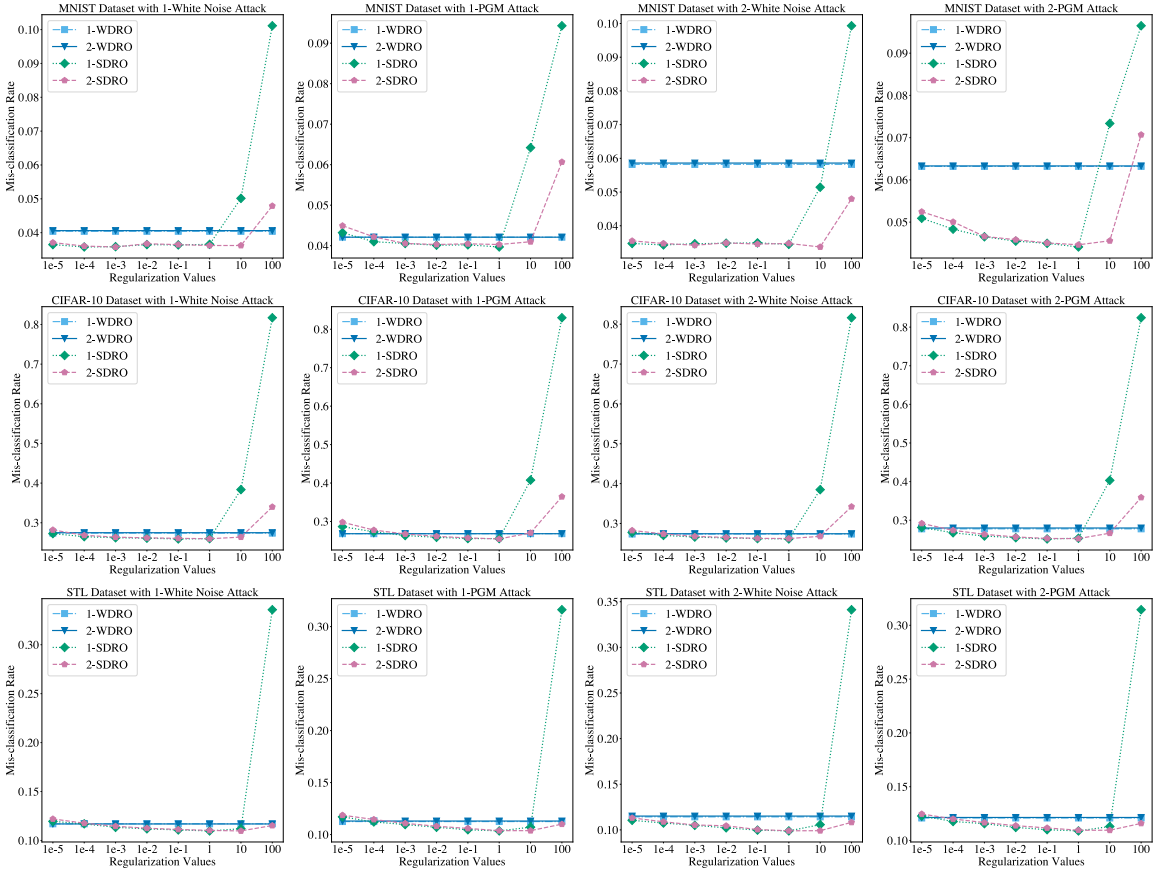

Figure D.7: Additional experiment results of the adversarial classification problem for different datasets and different types of perturbations. Details of these subplots follow the same setup from Fig. 5.8.

D.3 Sufficient Condition for Condition 1

Proposition 17. *Condition 1 holds if there exists $p \geq 1$ so that the following conditions are satisfied:*

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

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

(III) *Assumption 4*(III) *holds, and there exists $\lambda > 0$ such that $\mathbb{E}_{z\sim\nu}\left[e^{f(z)/(\lambda\epsilon)} e^{-2^{1-p}c(\overline{x},z)/\epsilon}\right] < \infty$.*

We make some remarks for the sufficient conditions listed above. The first condition can be satisfied by taking the transport cost as the p-th power of the metric defined on $\mathcal{Z}$ for any $p \geq 1$. The second condition requires the nominal distribution $\widehat{\mathbb{P}}$ is finite almost surely, e.g., it can be a subguassian distribution with respect to the transport cost c. We first present an useful technical lemma before showing the proof of Proposition 17.

Lemma 22. *Under the first condition of Proposition 17, for any $x \in \mathcal{Z}$, it holds that*

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

Proof of Lemma 22. Based on the inequality $(a + b)^p \leq 2^{p-1}(a^p + b^p)$, we can see that

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

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

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

The proof is completed. □

Proof of Proposition 17. One can see that for any $x \in \operatorname{supp} \widehat{\mathbb{P}}$, it holds that

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

where the first inequality is based on the lower bound in Lemma 22, the second inequality is based on the triangular inequality $c(x,z) \geq 2^{1-p}c(\overline{x},z) - c(x,\overline{x})$. Note that almost surely for all $x \in \operatorname{supp}\widehat{\mathbb{P}}$, $c(x,\overline{x}) < \infty$. Moreover, $0 < \mathbb{E}_{z\sim\nu}\left[e^{-2^{p-1}c(\overline{x},z)/\epsilon}\right] \leq \mathbb{E}_{z\sim\nu}\left[e^{-c(\overline{x},z)/\epsilon}\right] < \infty$, where the lower bound is because $c(\overline{x},z) < \infty$ almost surely for all z, the upper bound is because $c(\overline{x},z) \geq 0$ almost surely for all z. Based on these observations, we have that

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

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

D.4 Proofs of Technical Results in Section 5.3.2 and 5.3.3

Proof of Remark 13. Recall the dual objective function in (5.1) is

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

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

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

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

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

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

Proof of Example 4. In this example, the dual objective becomes

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

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

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

The term (I) can be simplified using the integral of exponential functions method:

$$\text{(I)} = \begin{cases} \det\left(I - \dfrac{2\theta\theta^{\mathrm{T}}}{\lambda}\right)^{-1/2} \exp\left(2\dfrac{(\theta^{\mathrm{T}}a - b)^2}{\lambda^2\epsilon}\theta^{\mathrm{T}}A^{-1}\theta\right), & \text{when } \|\theta\|_2^2 < \dfrac{\lambda}{2}, \\ \infty, & \text{otherwise,} \end{cases}$$

where the matrix $A = I - \frac{2\theta\theta^{\mathrm{T}}}{\lambda}$. Finally, we obtain that if $\|\theta\|_2^2 < \frac{\lambda}{2}$,

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

Substituting this expression into (D.2) gives the desired result. □

Proof of Corollary 3. We now introduce the epi-graphical variables $s_i, i = 1, \ldots, n$ to reformulate V_{D} as

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

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

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

where the second constraint set can be formulated as $(\lambda\epsilon, a_{i,\ell}, f(z_\ell) - s_i) \in \mathcal{K}_{\exp}$. Substituting this expression into V_{D} completes the proof. □

D.5 Proofs of Technical Results in Section 5.3.4

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

Lemma 23. *([161, Section 2.1] or [279]) For fixed τ and a reference measure $\nu \in \mathcal{M}(\mathcal{Z})$, consider the optimization problem*

$$v(\tau) = \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{z \sim \mathbb{P}} \left[f(z) - \tau \log \left(\frac{\mathrm{d}\mathbb{P}(z)}{\mathrm{d}\nu(z)} \right) \right] \right\}. \tag{D.3}$$

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

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

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

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

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

it holds that

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

and $\lim_{\tau \downarrow 0} v(\tau) = v(0)$. The optimal solution in (D.3) *has the expression*

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

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

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

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

Lemma 24 (Measurability of $v_x(\lambda)$)**.** *Assume Assumptions 4*(I)*, 4*(II)*, 4*(III) *hold. For fixed $\lambda \geq 0$, define the function $v_x(\lambda) : \ \operatorname{supp} \widehat{\mathbb{P}} \to \mathbb{R} \cup \{+\infty\}$ as*

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

The function $v_x(\lambda)$ is measurable with respect to $x \sim \widehat{\mathbb{P}}$ regardless of the choice of $\lambda \geq 0$.

Proof of Lemma 24. When $\lambda = 0$, by Lemma 23, it holds that

$$v_x(\lambda) = \operatorname{ess\,sup}_{\nu}(f),$$

which is a constant independent of x, which is clearly measurable. When $\lambda > 0$ and satisfies Condition 1, by Lemma 23, it holds that

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

As loss function f and cost function c are both measurable, by conditioning Lemma [173, Lemma 2.11], $v_x(\lambda)$ is measurable. When $\lambda > 0$ such that the event

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

satisfies $\widehat{\mathbb{P}}(E) > 0$, by Lemma 23, it holds that

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

For fixed $\alpha \in \mathbb{R}$, the level set

$$\{x :\ v_x(\lambda) \geq \alpha\} = \{x \in E^c :\ v_x(\lambda) \geq \alpha\} \cup E = \big\{x \in E^c :\ \lambda\epsilon \log \mathbb{E}_{z\sim\nu}\big[e^{(f(z)-\lambda c(x,z))/(\lambda\epsilon)}\big] \geq \alpha\big\} \cup E,$$

which is clearly a measurable set, and therefore $v_x(\lambda)$ is measurable. The proof is completed.

□

Proof of Lemma 4. Recall from (5.6) that

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

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

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

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

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

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

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

□

In the remaining of this subsection, we provide the full proof of Theorem 20. We first show that the dual minimizer exists.

Lemma 25 (Existence of Dual Minimizer)**.** *Suppose $\overline{\rho} > 0$ and Condition 1 is satisfied, then the dual minimizer λ^* exists, which either equals to 0 or satisfies Condition 1.*

Proof of Lemma 25. We first show that $\lambda^* < \infty$. Denote by $v(\lambda)$ the objective function for the dual problem:

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

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

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

where the first equality follows from the change-of-variable technique with $\beta = 1/\lambda$, the second equality follows from the definition of derivative, the third and the last equality follows from the dominated convergence theorem. As a consequence, as long as $\overline{\rho} > 0$, we have $\lim_{\lambda\to\infty} v(\lambda) = \infty$. We can take λ satisfying Condition 1 and then $v(\lambda) < \infty$. This, toegther with the fact that $v(\cdot)$ is continuous, guarantees the existence of the dual minimizer. Hence $\lambda^* < \infty$, which implies that either $\lambda^* = 0$ or λ^* satisfies Condition 1. □

Next, we establish first-order optimality condition for cases $\lambda^* > 0$ or $\lambda^* = 0$, corresponding to whether the Sinkhorn distance constraint in (`Primal`) is binding or not. Lemma 26 below presents a necessary and sufficient condition for the dual minimizer $\lambda^* = 0$, corresponding to the case where the Sinkhorn distance constraint in (`Primal`) is not binding.

Lemma 26 (Necessary and Sufficient Condition for $\lambda^* = 0$)**.** *Suppose $\overline{\rho} > 0$ and Condition 1 is satisfied, then the dual minimizer $\lambda^* = 0$ if and only if all the following conditions hold:*

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

(II) $\overline{\rho}' = \overline{\rho} + \epsilon\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\big[\log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[1_A(z)]\big] \geq 0$, *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 $\overline{\rho}$.

Proof of Lemma 26. Suppose the dual minimizer $\lambda^* = 0$, then taking the limit of the dual objective function gives

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

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

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

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

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

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

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

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

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

Then we show that almost surely for all x,

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

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

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

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

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

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

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

We can see that

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

and

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

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

exists $\overline{\lambda} > 0$ such that

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

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

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

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

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

The gradient of $v(\lambda)$ becomes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof of Lemma 5. Recall that $v(\lambda)$ denotes the objective function for the dual problem. The optimality condition can be derived by taking $\nabla_\lambda\, v(\lambda)\mid_{\lambda=\lambda^*} = 0$. To show the uniqueness of λ^*, we find that

$$\begin{aligned}&\nabla^2_\lambda v(\lambda)\\ =&\frac{1}{\lambda^3\epsilon}\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\left(\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[e^{f(z)/(\lambda\epsilon)}]\right)^{-2}\cdot\left(\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[e^{f(z)/(\lambda\epsilon)}f^2(z)]\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[e^{f(z)/(\lambda\epsilon)}] - \left\{\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[e^{f(z)/(\lambda\epsilon)}f(z)]\right\}^2\right)\right].\end{aligned}$$

It can be shown by the Cauchy-Schwarz inequality that $\nabla^2_\lambda v(\lambda) \geq 0$ for any $\lambda > 0$, and the equality holds if and only if $f(\cdot)$ is a constant. If it is the case, the dual objective $v(\lambda)$ has the unique minimizer $\lambda^* = 0$, which contradicts to our assumption. Hence, strict convexity holds for the dual objective and it implies the uniquess of λ^*. □

Proof of Theorem 20. Recall the feasibility result in Theorem 20(I) can be easily shown by considering the reformulation of V in Lemma 4 and the non-negativity of KL-divergence.

When $\overline{\rho} = 0$, one can see that

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

Therefore, the strong duality result holds in this case. Theorem 20(IV) can be shown by Lemma 26. It remains to show the strong duality result for $\overline{\rho} > 0$, which can be further separated to two cases: Condition 1 holds or not.

- When Condition 1 holds, by Lemma 25, the dual minimizer λ^* exists. The proof for $\lambda^* > 0$ can be found in main context. When $\lambda^* = 0$, the optimality condition in Lemma 26 holds. We construct the primal (approximate) solution $\mathbb{P}_* = \mathrm{Proj}_{2\#}\gamma_*$, where γ_* satisfies

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

  We can verify easily that the primal solution is feasible based on the optimality condition $\overline{\rho}' \geq 0$ in Lemma 26. Moreover, we can check that the primal optimal value is lower bounded by the dual optimal value:

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

  where the second equality is because that $z \in A$ so that $f(z) = \operatorname*{ess\,sup}_{\nu} f$. This, together with the weak duality result, completes the proof in this part.

- When Condition 1 does not hold, we consider a sequence of real numbers $\{R_j\}_j$ such that $R_j \to \infty$ and take the objective function $f_j(z) = f(z)1\{f(z) \leq R_j\}$. Hence, there exists $\lambda > 0$ satisfying $\Pr_{x\sim\widehat{\mathbb{P}}}\left\{x : \ \mathbb{E}_{\mathbb{Q}_{x,\epsilon}}\left[e^{f_j(z)/(\lambda\epsilon)}\right] = \infty\right\} = 0$. According

to the necessary condition in Lemma 26, 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 20(III) to show that for sufficiently large j, it holds that

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

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

□

D.6 Proof of Theorem 21 in Section 5.4.2

In this section, we omit the dependence of λ when defining objective or subgradient terms, e.g., we write $F(\theta)$ for $F(\theta;\lambda)$. We first present some preliminaries that can be useful for developing the proof result in Section 5.4.2. As any two norms on a finite-dimensional vector space are equivalent, we impose the following assumption throughout Section 5.4 without loss of generality:

Assumption 12. *There exists* $\mathfrak{c}$ *and* $\mathfrak{d}$ *such that* $\mathfrak{c}\|\cdot\|_2 \leq \|\cdot\| \leq \mathfrak{d}\|\cdot\|_2$.

By Assumption 12, we obtain the bound regarding the dual norm $\|\cdot\|_*$:

$$\mathfrak{d}^{-1}\cdot\|\cdot\|_2 \leq \|\cdot\|_* \leq \mathfrak{c}^{-1}\cdot\|\cdot\|_2.$$

The complexity result of our proposed gradient estimators is summarized below.

Remark 33 (Complexity of Gradient Estimators [156, Appendix B])**.** *To generate the SG estimator* $v^{SG}(\theta)$*, one needs to generate* 1 *sample from* $\widehat{\mathbb{P}}$ *and* 2^L *samples from* $\mathbb{Q}_{x,\epsilon}$ *for some* $x \in \operatorname{supp}\widehat{\mathbb{P}}$*. To generate the RT-MLMC estimator* $v^{RT\text{-}MLMC}(\theta)$*, one needs to generate* 1 *sample from* $\widehat{\mathbb{P}}$ *and the required (expected) number of samples from* $\mathbb{Q}_{x,\epsilon}$ *for some* $x \in \operatorname{supp}\widehat{\mathbb{P}}$ *equals* $\frac{L}{2-2^{-L}} = \mathcal{O}(L)$. ♣

Next, we present some basic properties regarding the approximation function $F^\ell(\theta)$ defined in (5.14) in Lemma 27, which can be used to show Theorem 21. Recall that we defined the constant $K_{\lambda,\epsilon,B} = B/(\lambda\epsilon)$.

Lemma 27. (I) *Under Assumption 5*(III)*, it holds that*

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

(II) *Under Assumption 5*(III) *and 5*(II)*, it holds that*

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

(III) *Under Assumption 5*(II)*, it holds that*

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

Additionally when Assumption 5(III) *holds, it holds that*

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

Proof of Lemma 27. Recall that (5.12) is a special CSO problem in (5.13), by taking $H^1(\cdot) = \lambda\epsilon \log(\cdot)$ and $H^2(\cdot, z) = \exp(f_\cdot(z)/(\lambda\epsilon))$. Under the assumptions stated in Lemma 27, it can be shown that $H^2(\cdot, z)$ is $\exp(K_{\lambda,\epsilon,B})$-uniformly bounded, $\exp(K_{\lambda,\epsilon,B})L_f/(\lambda\epsilon)$-Lipschitz continuous. The function $H^1(\cdot)$ has the domain set $[1, \exp(K_{\lambda,\epsilon,B})]$, and is therefore $\lambda\epsilon$-Lipschitz continuous and $\lambda\epsilon$-smooth. Thus, the desired results hold by applying [154, Lemma 3.1] and [156, Proposition 4.1]. □

D.6.1 Proof of Theorem 21

We first study the convergence guarantees for solving a generic nonsmooth convex optimization $\min_{\theta\in\Theta} F(\theta)$. Let $\overline{F}(\theta)$ denote its approximation, with the approximation bias Δ_F satisfying

$$|\overline{F}(\theta) - F(\theta)| \leq \Delta_F, \quad \forall \theta \in \Theta.$$

Denote by $\nabla\overline{F}(\theta)$ a subgradient of $\overline{F}$ at θ. Suppose for a given θ, the subgradient estimate of $F(\theta)$, denoted as $v(\theta)$, satisfies

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

Let $\overline{\theta}^* \in \arg\min_{\theta\in\Theta} \overline{F}(\theta)$ and $\theta^* \in \arg\min_{\theta\in\Theta} F(\theta)$. We then establish the following result.

Lemma 28 (BSMD for Nonsmooth Convex Optimization)**.** *Under the assumptions stated above and with the initial guess $\theta_0 \in \Theta$, consider the BSMD algorithm that generates the following iteration:*

$$\theta_{t+1} = \mathrm{Prox}_{\theta_t}\big(hv(\theta_t)\big), \quad \theta_0 \in \Theta, \quad t = 0, \ldots, T-1,$$

where the stepsize parameter $h = \sqrt{\frac{2\kappa D_\omega(\theta_0, \overline{\theta}^*)}{TM_*^2}}$. Let the estimated optimal solution generated by BSMD algorithm be $\widehat{\theta} = \frac{1}{T}\sum_{t=1}^T \theta_t$. Then, the suboptimality gap satisfies:

$$\mathbb{E}\big[F(\widehat{\theta}) - F(\theta^*)\big] \leq 2\Delta_F + M_*\sqrt{\frac{2D_\omega(\theta_0, \overline{\theta}^*)}{\kappa T}}.$$

Remark 34. *If the approximation bias is zero (i.e., $\Delta_F = 0$), the BSMD algorithm reduces to the standard SMD studied in [233]. By [233, Section 2.3], the suboptimality gap in Lemma 28 is bounded by $M_*\sqrt{\frac{2D_\omega(\theta_0,\overline{\theta}^*)}{\kappa T}}$. For the case where $\Delta_F > 0$, the proof of Lemma 28 follows from the decomposition argument similar to [156, Eq. (9)]. However, our result generalizes to the BSMD algorithm with (potentially) nonsmooth loss functions,*

whereas [156] focuses only on the SGD algorithm for unconstrained optimization with smooth loss functions. ♣

Now we are ready to show complexity results for BSMD using SG and RT-MLMC estimators. Both estimators rely on the same approximation function $F^L(\theta)$ defined in (5.14). By Lemma 27(I), $\Delta_F = \lambda\epsilon \exp(2K_{\lambda,\epsilon,B}) \cdot 2^{-(L+1)}$. We now analyze each estimator separately.

SG. It can be shown from the first part of Lemma 27(III) that $\mathbb{E}\left[\|v^{\mathrm{SG}}(\theta)\|_*^2\right] \le \left(M_*^{\mathrm{SG}}\right)^2 := \mathfrak{c}^{-2}L_f^2$. To obtain δ-optimal solution for SG estimator, by Lemma 28, it suffices to ensure

$$2\Delta_F \le \frac{\delta}{2}, \quad M_*^{\mathrm{SG}}\sqrt{\frac{2D_\omega(\theta_0,\overline{\theta}^*)}{\kappa T}} \le \frac{\delta}{2}.$$

To satisfy these conditions, we specify the following hyper-parameters:

$$L = \left\lceil \frac{1}{\log 2}\left[\log\frac{2\lambda\epsilon\exp(2K_{\lambda,\epsilon,B})}{\delta}\right]\right\rceil, \quad T = \left\lceil \frac{8L_f^2 D_\omega(\theta_0,\overline{\theta}^*)}{\kappa\mathfrak{c}^2\delta^2}\right\rceil, \quad h = \sqrt{\frac{2\kappa\mathfrak{c}^2 D_\omega(\theta_0,\overline{\theta}^*)}{TL_f^2}}.$$

RT-MLMC. By the second part of Lemma 27(III) and basic calculation, we find

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

Similar to the case of SG, we ensure

$$2\Delta_F \le \frac{\delta}{2}, \quad M_*^{\mathrm{RT-MLMC}}\sqrt{\frac{2D_\omega(\theta_0,\overline{\theta}^*)}{\kappa T}} \le \frac{\delta}{2}.$$

To satisfy these conditions, we select the following hyper-parameters:

$$L = \left\lceil \frac{1}{\log 2} \left[\log \frac{2\lambda\epsilon \exp(2K_{\lambda,\epsilon,B})}{\delta} \right] \right\rceil, \quad T = \left\lceil \frac{16(L+1)L_f^2 D_\omega(\theta_0, \overline{\theta}^*) \exp(4K_{\lambda,\epsilon,B})}{\kappa \mathfrak{c}^2 \delta^2} \right\rceil,$$
$$h = \sqrt{\frac{2\kappa D_\omega(\theta_0, \overline{\theta}^*)}{T\big(M_*^{\mathrm{RT-MLMC}}\big)^2}}.$$

By Remark 33, when running BSMD with SG estimator, the sample complexity from $\widehat{\mathbb{P}}$ equals $\mathcal{O}(T)$ and that from $\mathbb{Q}_{x,\epsilon}$ equals $\mathcal{O}(T2^L)$; when running BSMD with RT-MLMC estimator, the sample complexity from $\widehat{\mathbb{P}}$ equals $\mathcal{O}(T)$ and that from $\mathbb{Q}_{x,\epsilon}$ equals $\mathcal{O}(TL)$. Substituting the expressions of T, L gives the desired result.

D.7 Proofs of Technical Results in Section 5.4.2

We first provide two technical lemmas that can be useful to show the main results in Section 5.4.2.

Lemma 29. *Under Assumption 5*(III)*, it holds that* $\mathbb{E}[(A^\ell(\theta, \zeta^\ell; \lambda))^2] \le \lambda^2\epsilon^2 \exp(2K_{\lambda,\epsilon,B}) \cdot 2^{-\ell}$.

Proof of Lemma 29. The proof follows the similar procedure from [156, Proposition 4.1]. □

Lemma 30 (Complexity of RT-MLMC-based Objective Estimator)**.** *Let error probability* $\alpha \in (0,1)$ *and accuracy level* $\delta > 0$*. Assume Assumption 5*(III) *holds and specify*

$$L = \left\lceil \frac{1}{\log 2} \left[\log \frac{\lambda\epsilon \exp(2K_{\lambda,\epsilon,B})}{\delta} \right] \right\rceil, \quad m' = \mathcal{O}(1) \frac{\lambda^2\epsilon^2 \exp(2K_{\lambda,\epsilon,B})(L+1)}{\delta^2} \cdot \log \frac{2}{\alpha}. \tag{D.4}$$

Then, the RT-MLMC estimator (5.18) *has an accuracy error* δ *with probability at least* $1-\alpha$*. Its sample complexity from* $\widehat{\mathbb{P}}$ *equals* $\mathcal{O}(m') = \widetilde{\mathcal{O}}(\lambda^2\epsilon^2 K_{\lambda,\epsilon,B} \exp(2K_{\lambda,\epsilon,B}) \cdot \delta^{-2})$ *and that*

from $\mathbb{Q}_{x,\epsilon}$ *equals* $\mathcal{O}(m' \cdot L) = \widetilde{\mathcal{O}}(\lambda^2\epsilon^2 K_{\lambda,\epsilon,B}^2 \exp(2K_{\lambda,\epsilon,B}) \cdot \delta^{-2})$. *Here* $\widetilde{\mathcal{O}}(\cdot)$ *hides constants linearly depending on* $(\log \frac{\lambda\epsilon}{\delta})^2$ *and* $\log \frac{1}{\alpha}$.

Proof of Lemma 30. We first specify L as in (D.4) such that $|F^L(\theta;\lambda) - F(\theta;\lambda)| \le \frac{\delta}{2}$. The RT-MLMC estimator (5.18) satisfies that

$$\begin{aligned}
&\mathbb{E}[\widehat{F}(\theta;\lambda)] = F^L(\theta;\lambda), \\
&\mathbb{V}\mathrm{ar}\left(\widehat{F}(\theta;\lambda)\right) \le \frac{1}{m'}\sum_{\ell=0}^{L}\frac{1}{p_\ell}\mathbb{E}[(A^\ell(\theta,\zeta^\ell))^2] \le \frac{1}{m'}\lambda^2\epsilon^2\exp(2K_{\lambda,\epsilon,B})\cdot(L+1).
\end{aligned}$$

Consequently, there exists $\delta' > 0$ such that

$$\begin{aligned}
&\mathrm{Pr}\left\{|F(\theta;\lambda) - \widehat{F}(\theta;\lambda)| > \delta\right\} \le \mathrm{Pr}\left\{|F^L(\theta;\lambda) - \widehat{F}(\theta;\lambda)| > \frac{\delta}{2}\right\} \\
\le& 2\exp\left(-\frac{\delta^2}{4(\delta'+2)\mathbb{V}\mathrm{ar}\left(\widehat{F}(\theta;\lambda)\right)}\right) \le 2\exp\left(-\frac{\delta^2 m'}{4(\delta'+2)\lambda^2\epsilon^2\exp(2K_{\lambda,\epsilon,B})(L+1)}\right),
\end{aligned}$$

where the second inequality is based on the Cramer's large deviation theorem [176], and the last inequality is by the upper bound on $\mathbb{V}\mathrm{ar}\left(\widehat{F}(\theta;\lambda)\right)$. To make the desired coverage probability, we take m' as in (D.4). The complexity results are derived by standard calculation similar to Remark 33. □

In the following, we provide the proof of Proposition 5.

Proof of Proposition 5. Denote by $\theta^* = \arg\min_{\theta\in\Theta} F(\theta;\lambda)$. The goal is to choose hyperparameters such that

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

On the one hand,

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

On the other hand,

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

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

$$\Pr\left\{\max_{i\in[m]}\ |F(\widehat{\theta}_i;\lambda)-\widehat{F}(\widehat{\theta}_i;\lambda)|\le\frac{\delta}{2}\right\}\ge 1-\frac{\eta}{2} \tag{D.5}$$

and

$$\Pr\left\{\min_{i\in[m]}\ F(\widehat{\theta}_i;\lambda)-F(\theta^*;\lambda)\le\frac{\delta}{2}\right\}\ge 1-\frac{\eta}{2}. \tag{D.6}$$

To ensure the relation (D.5), it suffices to apply Lemma 30 with error probability $\frac{\eta}{2m}$ and accuracy level $\delta/2$. It implies that the sample complexity from $\widehat{\mathbb{P}}$ at Step 3 of Algorithm 8 for each independent repetition is $\widetilde{\mathcal{O}}(\lambda^2\epsilon^2K_{\lambda,\epsilon,B}\exp(2K_{\lambda,\epsilon,B})\cdot\delta^{-2})$, and that from $\mathbb{Q}_{x,\epsilon}$ is $\widetilde{\mathcal{O}}(\lambda^2\epsilon^2K^2_{\lambda,\epsilon,B}\exp(2K_{\lambda,\epsilon,B})\cdot\delta^{-2})$. To ensure the relation (D.6), it suffices to take

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

By Markov's inequality, it suffices to ensure

$$\mathbb{E}[F(\widehat{\theta}_i;\lambda)-F(\theta^*;\lambda)]\le\frac{\delta}{2}\left(\frac{\eta}{2}\right)^{1/m},\quad\forall i\in[m]. \tag{D.7}$$

By Theorem 21(II) with accuracy level $\frac{\delta}{2}\left(\frac{\eta}{2}\right)^{1/m}$, the sample complexity from $\widehat{\mathbb{P}}$ at Step 2 of Algorithm 8 for each independent repetition is $\widetilde{\mathcal{O}}(K_{\lambda,\epsilon,B}\exp(4K_{\lambda,\epsilon,B})\cdot\delta^{-2}\eta^{-2/m})$, and that from $\mathbb{Q}_{x,\epsilon}$ is $\widetilde{\mathcal{O}}(K^2_{\lambda,\epsilon,B}\exp(4K_{\lambda,\epsilon,B})\cdot\delta^{-2}\eta^{-2/m})$. Therefore, the sample complexity from $\widehat{\mathbb{P}}$ of Algorithm 8 is

$$\begin{aligned}&m\cdot\left[\widetilde{\mathcal{O}}(K_{\lambda,\epsilon,B}\exp(4K_{\lambda,\epsilon,B})\cdot\delta^{-2}\eta^{-2/m})+\widetilde{\mathcal{O}}(\lambda^2\epsilon^2K_{\lambda,\epsilon,B}\exp(2K_{\lambda,\epsilon,B})\cdot\delta^{-2})\right]\\ =&\widetilde{\mathcal{O}}(H_{\lambda,\epsilon,B}K_{\lambda,\epsilon,B}\exp(2K_{\lambda,\epsilon,B})\delta^{-2}\cdot m(1+\eta^{-2/m}))\end{aligned}$$

and that from $\mathbb{Q}_{x,\epsilon}$ is

$$\begin{aligned}
&m \cdot \left[\widetilde{\mathcal{O}}(K_{\lambda,\epsilon,B}^2 \exp(4K_{\lambda,\epsilon,B}) \cdot \delta^{-2}\eta^{-2/m}) + \widetilde{\mathcal{O}}(\lambda^2\epsilon^2 K_{\lambda,\epsilon,B}^2 \exp(2K_{\lambda,\epsilon,B}) \cdot \delta^{-2})\right] \\
=&\widetilde{\mathcal{O}}(H_{\lambda,\epsilon,B} K_{\lambda,\epsilon,B}^2 \exp(2K_{\lambda,\epsilon,B})\delta^{-2} \cdot m(1+\eta^{-2/m})).
\end{aligned}$$

In the above deviation, we defined the constant $H_{\lambda,\epsilon,B} = \max(\exp(2K_{\lambda,\epsilon,B}), \lambda^2\epsilon^2)$. Hence, it suffices to specify m such that $\widetilde{\mathcal{O}}(m(1+\eta^{-2/m}))$ is minimized. One valid choice is $m = \lceil \log_2 \frac{2}{\eta} \rceil$, which leads to the desired complexity bounds. □

Finally, we show the proof of Theorem 22. A key technique is the following complexity result on bisection search with inexact oracles.

Lemma 31 (Complexity for Noisy Bisection)**.** *Let the accuracy level $\delta > 0$, and $\Psi : \mathbb{R} \to \mathbb{R}$ be a L_Ψ-Lipschitz continuous and convex function defined on the interval $[\lambda_l, \lambda_u]$. Assume there exists an oracle $\widehat{\Psi} : \mathbb{R} \to \mathbb{R}$ such that $|\widehat{\Psi}(\lambda) - \Psi(\lambda)| \leq \delta, \forall \lambda$. Let us run Algorithm 9 for $T' = \lceil \log_2 \left(\frac{L_\Psi(\lambda_u - \lambda_l)}{\delta}\right) \rceil$ iterations, then with at most $3 + 2T'$ calls to $\widehat{\Psi}$, Algorithm 9 outputs $\widehat{\lambda}$ so that*

$$\Psi(\widehat{\lambda}) - \min_{\lambda \in [\lambda_l, \lambda_u]} \Psi(\lambda) \leq 4\delta.$$

Proof of Lemma 31. The proof is straightforward by following [83, Lemma 33] □

Proof of Theorem 22. It can be verified that Ψ is a convex function with a subgradient

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

where $\theta_\lambda^* \in \arg\min_{\theta\in\Theta} F(\theta;\lambda)$. By Assumption 5 and $\lambda \in [\lambda_l, \lambda_u]$, this subgradient vector is bounded:

$$\left|\frac{\partial}{\partial\lambda}\Psi(\lambda)\right| \leq L_\Psi := \overline{\rho} + \frac{B}{\lambda_l}\left[1 + \exp(K_{\lambda_l,\epsilon,B})\right].$$

In summary, $\Psi(\lambda)$ is a L_Ψ-Lipschitz and convex function defined on $[\lambda_l, \lambda_u]$. Applying Lemma 31 with accuracy level $\delta/4$ 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 $\widehat{\Psi}$ for $3+2\left\lceil \log_2\left(\frac{4L_\Psi(\lambda_u-\lambda_l)}{\delta}\right)\right\rceil$ times. □

Lemma 32. *Under Assumption 5, the optimal multiplier λ^* to* (D) *satisfies $\lambda^* \leq \frac{B}{\overline{\rho}}$.*

Proof. It can be verified that

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\lambda}\Psi(\lambda)\bigg|_{\lambda=\lambda^*} \\
&= \overline{\rho} + \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\epsilon\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_{\theta^*_{\lambda^*}}(z)/(\lambda\epsilon)}\right]\right] - \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\frac{\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_{\theta^*_{\lambda^*}}(z)/(\lambda\epsilon)}f_{\theta^*_{\lambda^*}}(z)\right]}{\lambda^*\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_{\theta^*_{\lambda^*}}(z)/(\lambda\epsilon)}\right]}\right] \\
&\geq \overline{\rho} - \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\frac{\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_{\theta^*_{\lambda^*}}(z)/(\lambda\epsilon)}f_{\theta^*_{\lambda^*}}(z)\right]}{\lambda^*\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f_{\theta^*_{\lambda^*}}(z)/(\lambda\epsilon)}\right]}\right] \\
&\geq \overline{\rho} - \frac{B}{\lambda^*},
\end{aligned}
$$

where the two inequalities is based on the fact that $0 \leq f_\theta(z) \leq B$. The desired result holds directly. □

APPENDIX E

PROOFS AND ADDITIONAL DETAILS OF CHAPTER 6

E.1 Preliminaries on Projected Stochastic (Sub-)Gradient Descent

In the following, we present the convergence results on the standard projected stochastic (sub-)gradient descent algorithm with unbiased gradient estimates, which can be useful for the complexity analysis in Section 6.3.3.

Consider minimization of the objective function $F(\theta)$ over the constrained domain set Θ.

Nonsmooth Convex Optimization.

Let the objective $F(\theta)$ be a convex function in θ. Assume one can obtain stochastic estimate $G(\theta, \xi)$ such that for any $\theta \in \Theta$,

- $\mathbb{E}[G(\theta, \xi)] \in \partial F(\theta)$, where $\partial F(\theta)$ denotes the subgradient of F at θ;
- $\mathbb{E}\,\|G(\theta, \xi)\|^2 \leq M^2$.

Starting from an initial guess $\theta_1 \in \Theta$, the projected stochastic subgradient descent algorithm generates iterates

$$\theta_{t+1} = \text{Proj}_{\Theta}(\theta_t - \gamma G(\theta_t, \xi_t)), \quad t = 1, \ldots, T-1. \tag{Projected-SGD}$$

where $\gamma > 0$ is a constant step size, and $\xi_1, \ldots, \xi_{T-1}$ are i.i.d. copies of ξ. We take the average of all iterates $\{\theta_1, \ldots, \theta_T\}$ as the estimated optimal solution, denoted as $\widetilde{\theta}$.

Lemma 33 ([233]). *Under the above setting, suppose we take the step size $\gamma = \frac{D_*}{M\sqrt{T}}$, then it holds that*

$$\mathbb{E}\left[F(\widetilde{\theta}) - \min_{\theta \in \Theta} F(\theta)\right] \leq \frac{D_* M}{\sqrt{T}},$$

where the constant $D_* = F(\theta_1) - \min_{\theta\in\Theta} F(\theta)$.

Smooth Non-Convex Optimization.

In this part, we do not assume the convexity of $F(\theta)$. Instead, we assume the objecitve $F(\theta)$ is continuously differentiable and S-smooth such that

$$\|F(\theta) - F(\theta')\| \le S\|\theta - \theta'\|, \qquad \forall \theta, \theta' \in \Theta.$$

Besides, assume one can obtain stochastic estimate $G(\theta, \xi)$ such that for any $\theta \in \Theta$,

- $\mathbb{E}[G(\theta, \xi)] = \nabla F(\theta)$;
- $\mathbb{E}\,\|G(\theta, \xi) - \nabla F(\theta)\|^2 \le \sigma^2$.

We generate iteration points using nearly the same procedure as in (Projected-SGD), except that we update iteration points using mini-batch gradient estimator:

$$\theta_{t+1} = \text{Proj}_\Theta(\theta_t - \gamma V(\theta_t, \xi_t^{1:m})), \quad t = 1, \ldots, T-1, \quad V(\theta_t, \xi_t^{1:m}) = \frac{1}{m}\sum_{i=1}^{m} G(\theta_t, \xi_t^i), \tag{Mini-Projected-SGD}$$

where $\xi_i^t, t = 1, \ldots, T-1, i = 1, \ldots, m$ are i.i.d. copies of ξ. We take the estimated optimal solution $\widetilde{\theta}$ as the one that is randomly selected from $\{\theta_1, \ldots, \theta_T\}$ with equal probability.

Lemma 34 (Corollary 3 in [132])**.** *Under the above setting, suppose we take the step size* $\gamma = \frac{1}{2S}$*, then it holds that*

$$\mathbb{E}\left\|\frac{1}{\gamma}\left[\widetilde{\theta} - \text{Proj}_\Theta\big(\widetilde{\theta} - \gamma\nabla F(\widetilde{\theta})\big)\right]\right\|^2 \le \frac{8SD_*}{T} + \frac{6\sigma^2}{m},$$

where the constant $D_* := F(\theta_1) - \min_{\theta\in\Theta} F(\theta)$.

E.2 Proofs of Technical Results in Section 6.2

Proof of Theorem 23. Based on the assumption, it holds that $\mathrm{d}\gamma(z, z') = \mathrm{d}\widehat{P}(z)\,\mathrm{d}\gamma_z(z')$ for some conditional optimal transport mapping γ_z. Then, Problem (Primal-ϕ-Reg) can be reformulated as

$$\sup_{\{\gamma_z\}_{z\in\operatorname{supp}\widehat{P}}} \left\{\mathbb{E}_{z\sim\widehat{P}}\mathbb{E}_{z'\sim\gamma_z}[f(z)] - \eta\mathbb{E}_{z\sim\widehat{P}}\mathbb{E}_{z'\sim\nu_z}\left[\phi\left(\frac{\mathrm{d}\gamma_z(z')}{\mathrm{d}\nu_z(z')}\right)\right]\right\}. \tag{E.1}$$

Since the optimization over $z \in \operatorname{supp}\widehat{P}$ is decomposable, it holds that (Primal-ϕ-Reg) equals

$$\mathbb{E}_{z\sim\widehat{P}}\left[\sup_{\gamma_z\in\mathcal{P}(\mathcal{Z})} \left\{\mathbb{E}_{z'\sim\gamma_z}[f(z')] - \eta\mathbb{E}_{z'\sim\nu_z}\left[\phi\left(\frac{\mathrm{d}\gamma_z(z')}{\mathrm{d}\nu_z(z')}\right)\right]\right\}\right]. \tag{E.2}$$

The inner supremum problem above is a phi-divergence regularized linear program. Based on the strong duality result (see, e.g., Lemma 35) that reformulates this subproblem, we arrive at the reformulation of Problem (Primal-ϕ-Reg).

Next, we show how to construct the worst-case distribution, which suffices to construct an optimal conditional transport mapping γ_z^* for $z \in \operatorname{supp}\widehat{P}$. By the change-of-measure technique with $\zeta(z') = \frac{\mathrm{d}\gamma_z(z')}{\mathrm{d}\nu_z(z')}$, the inner supremum of (E.2) becomes

$$\sup_{\zeta\in\mathcal{Z}_+^*} \left\{\mathbb{E}_{z'\sim\nu_z}\big[f(z')\zeta(z') - \eta\phi(\zeta(z'))\big] : \mathbb{E}_{\nu_z}[\zeta] = 1\right\}. \tag{E.3}$$

We now construct the Lagrangian function associated with (E.3) as

$$\mathsf{L}(\zeta, \mu) = \mathbb{E}_{z'\sim\nu_z}\big[(f(z') - \mu)\zeta(z') - \eta\phi(\zeta(z'))\big] + \mu.$$

Recall from [53, Proposition 3.3] that, if there exists (ζ_z^*, μ_z^*) such that

$$\zeta_z^* \in \mathcal{Z}_+^*, \quad \mathbb{E}_{\nu_z}[\zeta_z^*] = 1, \quad \zeta_z^* \in \operatorname*{arg\,max}_{\zeta\in\mathcal{Z}_+^*} \mathsf{L}(\zeta, \mu_z^*),$$

it holds that ζ_z^* solves (E.3). The proof is completed by substituting the expression of γ^* in terms of γ_z^* and then substituting the expression of γ_z^* in terms of ζ_z^*. $\square$

Lemma 35 ([279, Section 3.2])**.** *Given a probability reference measure $\gamma \in \mathcal{P}(\mathcal{Z})$ and regularization value $\eta > 0$, consider the ϕ-divergence regularized problem:*

$$\sup_{\gamma \in \mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{z\sim\gamma}[f(z)] - \eta \mathbb{E}_{z\sim\nu}\left[\phi\left(\frac{\mathrm{d}\gamma(z)}{\mathrm{d}\nu(z)}\right)\right]\right\}.$$

There exists an optimal solution to this primal problem, and also, it can be reformulated as the dual problem

$$\inf_{\mu\in\mathbb{R}} \left\{ \mu + \mathbb{E}_{z\sim\nu}\left[\eta\phi^*\left(\frac{f(z)-\mu}{\eta}\right)\right]\right\}.$$

Proof of Proposition 6. It is easy to verify that

$$\begin{aligned}\text{Optval(Primal-}\phi\text{-Reg)} &= \mathbb{E}_{z\sim\widehat{P}}\left[\sup_{\gamma_z\in\mathcal{P}(\mathcal{Z})}\left\{\mathbb{E}_{z'\sim\gamma_z}[f(z')] - \eta\mathbb{E}_{z'\sim\nu_z}\left[\phi\left(\frac{\mathrm{d}\gamma_z(z')}{\mathrm{d}\nu_z(z')}\right)\right]\right\}\right] \\ &\le \mathbb{E}_{z\sim\widehat{P}}\left[\sup_{\gamma_z\in\mathcal{P}(\mathcal{Z})}\left\{\mathbb{E}_{z'\sim\gamma_z}[f(z')]\right\}\right] = \mathbb{E}_{z\sim\widehat{P}}\left[\max_{z'\in\mathbb{B}_\rho(z)} f(z')\right].\end{aligned}$$

Now, it suffices to show the other direction. For fixed $\eta > 0$ and z, let $\mu_{z,\eta}^*$ be the minimizer to

$$\inf_{\mu}\left\{\mu + \mathbb{E}_{z'\sim\nu_z}\left[(\eta\phi)^*\big(f(z')-\mu\big)\right]\right\}.$$

(I) For the case where $\lim_{t\to\infty}\frac{\phi(t)}{t} < \infty$, by [22], the dual formulation (Dual-ϕ-Reg) implicitly imposes an extra constraint:

$$\mu_{z,\eta}^* \ge f(z') - \eta\lim_{t\to\infty}\frac{\phi(t)}{t}, \forall z' \in \operatorname{supp}\nu_z \implies \mu_{z,\eta}^* \ge \operatorname*{ess\,sup}_{\nu_z} f - \eta\lim_{t\to\infty}\frac{\phi(t)}{t}.$$

It follows that

$$\begin{aligned}\text{(Dual-}\phi\text{-Reg)} &= \mathbb{E}_{z\sim\widehat{P}}\left[\mu^*_{z,\eta} + \mathbb{E}_{z'\sim\nu_z}\left[\eta\phi^*\left(\frac{f(z')-\mu^*_{z,\eta}}{\eta}\right)\right]\right] \\ &\geq \operatorname*{ess\,sup}_{\nu_z} f - \eta\lim_{t\to\infty}\frac{\phi(t)}{t} + \eta\mathbb{E}_{z'\sim\nu_z}\left[\phi^*\left(\frac{f(z')-\mu^*_{z,\eta}}{\eta}\right)\right]\end{aligned}$$

By taking $\eta \to 0$ both sides, we find

$$\text{Optval(Primal-}\phi\text{-Reg)} \geq \operatorname*{ess\,sup}_{\nu_z} f.$$

(II) For the case where $\lim_{t\to\infty}\frac{\phi(t)}{t} = \infty$, it holds that $\phi^*(s) \in (-\infty, \infty)$ for any finite s. In this case, for fixed $\eta > 0$,

$$\text{(Dual-}\phi\text{-Reg)} = \mathbb{E}_{z\sim\widehat{P}}\left[\mu^*_{z,\eta} + \mathbb{E}_{z'\sim\nu_z}\left[\eta\phi^*\left(\frac{f(z')-\mu^*_{z,\eta}}{\eta}\right)\right]\right].$$

For sufficiently small η, assume on the contrary that $\mu^*_{z,\eta} < \operatorname{ess\,sup}_{\nu_z} f$, then the event $E_{z,\eta} := \{z' : f(z') > \mu^*_{z,\eta}\}$ satisfies $\nu_z(E_{z,\eta}) > 0$.

- For $z' \notin E_{z,\eta}$,

$$\lim_{\eta\to 0}\eta\phi^*\left(\frac{f(z')-\mu^*_{z,\eta}}{\eta}\right) = \lim_{\eta\to 0}\eta\phi^*(0) = 0.$$

- For $z' \in E_{z,\eta}$,

$$\lim_{\eta\to 0}\eta\phi^*\left(\frac{f(z')-\mu^*_{z,\eta}}{\eta}\right) = \lim_{t\to\infty}\frac{1}{t}\phi^*\left(t\big(f(z')-\mu^*_{z,\eta}\big)\right) \to \infty.$$

Then it follows that

$$\begin{aligned}&\mathbb{E}_{z'\sim\nu_z}\left[\eta\phi^*\left(\frac{f(z')-\mu^*_{z,\eta}}{\eta}\right)\right] \\ =&\mathbb{E}_{z'\sim\nu_z}\left[\eta\phi^*\left(\frac{f(z')-\mu^*_{z,\eta}}{\eta}\right)\mathbf{1}(E^c_{z,\eta})\right] + \mathbb{E}_{z'\sim\nu_z}\left[\eta\phi^*\left(\frac{f(z')-\mu^*_{z,\eta}}{\eta}\right)\mathbf{1}(E_{z,\eta})\right] \to \infty.\end{aligned}$$

In summary, under the case where $\mu^*_{z,\eta} < \operatorname{ess\,sup}_{\nu_z} f$, (Dual-$\phi$-Reg) $\to \infty$ as $\eta \to 0$, which is a contradiction. Therefore, $\mu^*_{z,\eta} \geq \operatorname{ess\,sup}_{\nu_z} f$, which follows that

$$\text{(Dual-}\phi\text{-Reg)} \geq \operatorname*{ess\,sup}_{\nu_z} f + \eta \mathbb{E}_{z' \sim \nu_z} \left[\phi^* \left(\frac{f(z') - \mu^*_{z,\eta}}{\eta} \right) \right]$$

By taking $\eta \to 0$ both sides, we obtain Optval(Primal-ϕ-Reg) $\geq \operatorname{ess\,sup}_{\nu_z} f$.

□

E.3 Proofs of Technical Results in Section 6.3

Proof of Proposition 7. We define **Case 1** as the scenarios where $\phi'(s) \to -\infty$ as $s \to 0+$, and **Case 2** as the scenarios where $\phi'(s) \to K$ as $s \to 0+$, with the constant $K > -\infty$ being lower bounded.

For any fixed Lagrangian multiplier μ, $\gamma^*(\mu) \in \mathbb{R}_+^m$ is the optimum solution to $\max_{\gamma \in \mathbb{R}_+^m} \mathcal{L}(\mu, \gamma)$ if and only if

$$f_i - \mu - \eta\phi'(m(\gamma^*(\mu))_i) \le 0, \forall i, \quad (\gamma^*(\mu))_i \cdot \Big(f_i - \mu - \eta\phi'(m(\gamma^*(\mu))_i)\Big) = 0.$$

Under **Case 1**, the above optimality condition simplifies into $f_i - \mu - \eta\phi'(m(\gamma^*(\mu))_i) = 0, \forall i$, which implies

$$(\gamma^*(\mu))_i = \frac{1}{m}(\phi')^{-1}\Big(\frac{f_i - \mu}{\eta}\Big).$$

Under **Case 2**, the above optimality condition simplifies into

$$(\gamma^*(\mu))_i = \begin{cases} 0, & \text{if } i \in \mathcal{N} \triangleq \Big\{i \in [m] : \ f_i \le \mu + \eta K\Big\}, \\ \dfrac{1}{m}(\phi')^{-1}\Big(\dfrac{f_i - \mu}{\eta}\Big), & \text{otherwise.} \end{cases}$$

Therefore, the remaining task of Algorithm (11) is to find the optimal Lagrangian multiplier μ such that

$$h(\mu) := \sum_{i \in [m]} (\gamma^*(\mu))_i - 1 = \frac{1}{m} \sum_{i \in [m] \setminus \mathcal{N}} (\phi')^{-1}\Big(\frac{f_i - \mu}{\eta}\Big) - 1 = 0.$$

Due to the strict convexity of ϕ, $h(\mu)$ is strictly decreasing in μ. Also, it can be verified that the optimal multiplier belongs to the interval $[\underline{\mu}, \overline{\mu}]$:

- By the increasing property of $(\phi')^{-1}$, it holds that $h(\underline{\lambda}) \ge 0$;
- Under **Case 1**, it holds that $h(\overline{\mu}) \le 0$. Under **Case 2**, it holds that $h(\overline{\mu}) \le -1$.

Hence, we only need to perform $\mathcal{O}(\log \frac{1}{\epsilon})$ iterations of bisection search to obtain a near-optimal multiplier with ϵ precision. At each iteration of bisection search, the worst-case computational cost is $O(m)$. To compute the index set $\mathcal{N}$ at each iteration, we need to enumerate all support points $\{f_1, \ldots, f_m\}$, whose computational cost is $\mathcal{O}(m)$. In summary, the overall cost is $\mathcal{O}(m \log \frac{1}{\epsilon})$. □

Proposition 18 (Error Bound on Function Approximation)**.** *Under Assumption 8*(II)*, it holds that* $0 \leq F(\theta) - F^{\ell}(\theta) \leq G_{idf} \cdot 2^{-\ell}, \forall \theta \in \Theta$.

Proof of Proposition 18. It is worth noting that

$$F(\theta) - F^{\ell}(\theta) = \mathbb{E}_{z\sim\widehat{P}}\mathbb{E}_{\{z_i'\}_{i\in[2^\ell]}\sim\nu_z}\left[R(\theta; z) - \widehat{R}\big(\theta; z, \{z_i'\}_{i\in[2^\ell]}\big)\right],$$

where

$$\begin{aligned} R(\theta; z) &= \inf_{\mu}\ \Big\{\mu + \mathbb{E}_{z\sim\nu_z}[(\eta\phi)^*(f_\theta(z') - \mu)]\Big\}, \\ \widehat{R}\big(\theta; z, \{z_i'\}_{i\in[2^\ell]}\big) &= \inf_{\mu}\ \Big\{\mu + \frac{1}{2^\ell}\sum_{i\in[2^\ell]}[(\eta\phi)^*(f_\theta(z_i') - \mu)]\Big\}. \end{aligned}$$

By Jensen's inequality, it holds for any fixed (z, θ) that

$$R(\theta; z) \geq \mathbb{E}_{\{z_i'\}_{i\in[2^\ell]}\sim\nu_z}[\widehat{R}\big(\theta; z, \{z_i'\}_{i\in[2^\ell]}\big)],$$

and therefore $F(\theta) - F^{\ell}(\theta) \geq 0$.

On the other hand, $R(\theta; z)$ denotes the optimal value of the standard ϕ-divergence DRO with reference distribution ν_z, and $\widehat{R}\big(\theta; z, \{z_i'\}_{i\in[2^\ell]}\big)$ denotes its sample estimate using 2^ℓ i.i.d. samples generated from ν_z. By [191, Proposition 1], it holds for any fixed z that

$$\mathbb{E}_{\{z_i'\}_{i\in[2^\ell]}\sim\nu_z}\left[R(\theta; z) - \widehat{R}\big(\theta; z, \{z_i'\}_{i\in[2^\ell]}\big)\right] \leq \frac{G_{\text{idf}}}{2^\ell}.$$

The proof is completed. □

To prove Proposition 8, we rely on the following technical lemma that has been revealed in literature.

Lemma 36 (Lemma 6 in [191]). *Let $\gamma \in \Delta^m$ with m being an even integer. Let $\mathcal{I}$ be a random subset of $[1:m]$ of size $m/2$. Then it holds that*

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}}\gamma_i - \frac{1}{2}\right]^2 \le \frac{1}{2m} D_{\chi^2}(\gamma, \frac{1}{m}\mathbf{1}).$$

Proof of Proposition 8. (I) By definition, SG and RT-MLMC estimators $V^{\text{SG}}(\theta)$ and $V^{\text{RT-MLMC}}(\theta)$ are the unbiased gradient estimators from some objective function $\widetilde{F}^\ell(\theta)$ such that $|\widetilde{F}^\ell(\theta) - F^\ell(\theta)| \le \epsilon$. Therefore, the bias can be bounded as

$$|\widetilde{F}^\ell(\theta) - F(\theta)| \le |\widetilde{F}^\ell(\theta) - F^\ell(\theta)| + |F^\ell(\theta) - F(\theta)| \le \epsilon + \frac{G_{\text{idf}}}{2^\ell},$$

where the last inequality follows from Proposition 18.

(II) By definition,

$$\left\|V^{\text{SG}}(\theta)\right\|^2 = \left\|\frac{1}{n_L^{\text{o}}}\sum_{i=1}^{n_L^{\text{o}}} g^L(\theta, \zeta_i^L)\right\|^2 \le \frac{1}{n_L^{\text{o}}}\sum_{i=1}^{n_L^{\text{o}}}\left\|g^L(\theta,\zeta_i^L)\right\|^2.$$

Since $\{\zeta_i^L\}_i$ are n_L^{o} i.i.d. copies of ζ^L, it holds that

$$\mathbb{E}\left\|V^{\text{SG}}(\theta)\right\|^2 \le \mathbb{E}\left\|g^L(\theta,\zeta^L)\right\|^2$$

To bound $\mathbb{E}\left\|g^L(\theta,\zeta^L)\right\|^2$, we define the following notations. Let $\widehat{\gamma}$ be the optimal solution to $\widehat{R}\big(\theta; z, \{z_i'\}_{i\in[1:2^L]}\big)$ defined in (6.9), and $\widetilde{\gamma}$ be the estimated solution used

by the estimator $g^L(\theta, \zeta^L)$. As a consequence,

$$\begin{aligned}\|g^\ell(\theta,\zeta^\ell)\| &= \left\| \sum_{i\in[1:2^\ell]} \widetilde{\gamma}_i \nabla_\theta f_\theta(z_i') \right\| \le L_f \sum_{i\in[1:2^\ell]} \widetilde{\gamma}_i \\ &\le L_f \sum_{i\in[1:2^\ell]} \left(\widehat{\gamma}_i + \|\widehat{\gamma} - \widetilde{\gamma}\|_\infty\right) = L_f \cdot \left[1 + \sqrt{\frac{2\epsilon}{\kappa\eta}}\right].\end{aligned}$$

Therefore,

$$\mathbb{E}\left\|g^L(\theta,\zeta^L)\right\|^2 \le L_f^2 \left[1 + \sqrt{\frac{2\epsilon}{\kappa\eta}}\right]^2 \le 2L_f^2 \left[1 + \frac{2\epsilon}{\kappa\eta}\right].$$

Following the similar argument as in bounding $\mathbb{E}\left\|V^{\text{SG}}(\theta)\right\|^2$, we find

$$\begin{aligned}\mathbb{E}\left\|V^{\text{RT-MLMC}}(\theta)\right\|^2 &\le \mathbb{E}_{\widehat{L}_1}\mathbb{E}_{\zeta^{\widehat{L}_1}} \left\| \frac{1}{\mathbb{P}(\widehat{L}=\widehat{L}_1)} G^{\widehat{L}_1}(\theta,\zeta^{\widehat{L}_1}) \right\|^2 \\ &= \sum_{\ell=0}^{L} \mathbb{P}(\widehat{L}=\ell)\mathbb{E}_{\zeta_i^\ell} \left\| \frac{1}{\mathbb{P}(\widehat{L}=\ell)} G^\ell(\theta,\zeta^\ell) \right\|^2 \\ &= \sum_{\ell=0}^{L} \frac{1}{\mathbb{P}(\widehat{L}=\ell)} \cdot \mathbb{E}_{\zeta^\ell} \left\|G^\ell(\theta,\zeta^\ell)\right\|^2.\end{aligned}$$

It suffices to bound $\mathbb{E}_{\zeta^\ell}\left\|G^\ell(\theta,\zeta^\ell)\right\|^2$ for fixed level $\ell = 0, 1, \ldots, L$. To simplify notation,

- Let $\gamma, \gamma', \gamma''$ be the estimated optimal solutions corresponding to the objectives $\widetilde{R}\big(\theta; z, \{z_i'\}_{i\in[1:2^\ell]}\big), \widetilde{R}\big(\theta; z, \{z_i'\}_{i\in[1:2^{\ell-1}]}\big)$, and $\widetilde{R}\big(\theta; z, \{z_i'\}_{i\in[2^{\ell-1}+1:2^\ell]}\big)$ defined in (6.10), respectively.
- Let $\bar{\gamma}, \bar{\gamma}', \bar{\gamma}''$ be the optimal solutions for $\widetilde{R}\big(\theta; z, \{z_i'\}_{i\in[1:2^\ell]}\big), \widetilde{R}\big(\theta; z, \{z_i'\}_{i\in[1:2^{\ell-1}]}\big)$ and $\widetilde{R}\big(\theta; z, \{z_i'\}_{i\in[2^{\ell-1}+1:2^\ell]}\big)$ defined in (6.9), respectively.

Then it holds that

$$\|G^\ell(\theta,\zeta^\ell)\| = \left\| \sum_{i\in[1:2^\ell]} \Big(\gamma_i - \frac{1}{2}\gamma_i' \cdot 1\{i\in[1:2^{\ell-1}]\} - \frac{1}{2}\gamma''_{i-2^{\ell-1}} \cdot 1\{i\in[2^{\ell-1}+1:2^\ell]\}\Big)\nabla_\theta f_\theta(z_i') \right\|$$
$$\le L_f \sum_{i\in[1:2^{\ell-1}]} |\gamma_i - \gamma_i'/2| + L_f \sum_{i\in[2^{\ell-1}+1:2^\ell]} |\gamma_i - \gamma''_{i-2^{\ell-1}}/2|.$$

Recall that for each $i\in[1:2^{\ell-1}]$, $\gamma_i = \frac{1}{m}(\phi')^{-1}(\frac{f(z_i)-\mu}{\eta})$ and $\gamma_i'/2 = \frac{1}{m}(\phi')^{-1}(\frac{f(z_i)-\mu'}{\eta})$ for constants $\mu,\mu'\in\mathbb{R}$. Since ϕ is strongly convex, $(\phi')^{-1}(\cdot)$ is a strictly increasing function, and $\gamma_i - \gamma_i'/2$ is always of a constant sign for all $i\in[1:2^{\ell-1}]$. Therefore,

$$\sum_{i\in[1:2^{\ell-1}]} |\gamma_i - \gamma_i'/2| = \left| \sum_{i\in[1:2^{\ell-1}]} \gamma_i - \frac{1}{2}\sum_{i\in[1:2^{\ell-1}]} \gamma_i' \right|$$
$$\le 2^{\ell-1}\|\gamma-\bar\gamma\|_\infty + 2^\ell\|\gamma'-\bar\gamma'\|_\infty + \left| \sum_{i\in[1:2^{\ell-1}]} \bar\gamma_i - \frac{1}{2}\sum_{i\in[1:2^{\ell-1}]} \bar\gamma_i' \right|$$
$$\le \sqrt{\frac{2\epsilon}{\kappa\eta}} + \left| \sum_{i\in[1:2^{\ell-1}]} \bar\gamma_i - \frac{1}{2} \right|,$$

where the first inequality is by triangular inequality, the second inequality is by Proposition 7(II) and the relation $\sum_{i\in[1:2^{\ell-1}]}\bar\gamma_i' = 1$. One can follow the similar procedure to bound $\sum_{i\in[2^{\ell-1}+1:2^\ell]}|\gamma_i - \gamma''_{i-2^{\ell-1}}/2|$. As a consequence,

$$\mathbb{E}_{\zeta^\ell}\|G^\ell(\theta,\zeta^\ell)\|^2 \le L_f^2\mathbb{E}\left[2\sqrt{\frac{2\epsilon}{\kappa\eta}} + \left| \sum_{i\in[1:2^{\ell-1}]} \bar\gamma_i - \frac{1}{2} \right| + \left| \sum_{i\in[2^{\ell-1}+1:2^\ell]} \bar\gamma_i - \frac{1}{2} \right| \right]^2$$
$$\le 3L_f^2 \cdot \left(\frac{8\epsilon}{\kappa\eta} + \mathbb{E}\left| \sum_{i\in[1:2^{\ell-1}]} \bar\gamma_i - \frac{1}{2} \right|^2 + \mathbb{E}\left| \sum_{i\in[2^{\ell-1}+1:2^\ell]} \bar\gamma_i - \frac{1}{2} \right|^2 \right)$$
$$\le \frac{24L_f^2}{\kappa\eta}\cdot\epsilon + \frac{3L_f^2 D_{\chi^2}(\bar\gamma, \frac{1}{2^\ell}\mathbf{1})}{2^\ell}$$
$$\le \frac{24L_f^2}{\kappa\eta}\cdot\epsilon + \frac{3L_f^2 C}{2^\ell}.$$

Finally,

$$\mathbb{E}\left\|V^{\text{RT-MLMC}}(\theta)\right\|^2 \leq \frac{96L_f^2}{\kappa\eta}\cdot 2^L\cdot\epsilon + 6(L+1)L_f^2C.$$

(III) Since the random vectors $g^L(\theta,\zeta_i^L)$ for $i=1,\ldots,n_L^{\text{o}}$ are i.i.d., it holds that

$$\mathbb{V}\text{ar}[V^{\text{SG}}(\theta)] = \mathbb{V}\text{ar}\left[\frac{1}{n_L^{\text{o}}}\sum_{i=1}^{n_L^{\text{o}}} g^L(\theta,\zeta_i^L)\right] = \frac{\mathbb{V}\text{ar}[g^L(\theta,\zeta^L)]}{n_L^{\text{o}}} \leq \frac{\mathbb{E}\|g^L(\theta,\zeta^L)\|^2}{n_L^{\text{o}}} \leq \frac{2L_f^2\left[1+(2\epsilon)/(\kappa\eta)\right]}{n_L^{\text{o}}}.$$

The same argument applies when bounding $\mathbb{V}\text{ar}[V^{\text{RT-MLMC}}(\theta)]$.

(IV) For fixed $i=1,\ldots,n_L^{\text{o}}$, computing $g^L(\theta,\zeta_i^L)$ requires generating 2^L samples and then solve the penalized ϕ-divergence DRO with 2^L support points with controlled optimality gap ϵ. According to Lemma 7, its complexity is $\mathcal{O}(2^L\log\frac{1}{\epsilon})$. Hence, generating the SG estimator $V^{\text{SG}}(\theta)$ has cost $\mathcal{O}(n_L^{\text{o}}\cdot 2^L\log\frac{1}{\epsilon})$.

For fixed $i=1,\ldots,n_L^{\text{o}}$ and $\ell=0,\ldots,L$, computing $G^\ell(\theta,\zeta_i^\ell)$, according to the definition in (6.12), requires computatonal cost $\mathcal{O}(2^{\ell+1}\log\frac{1}{\epsilon})$. Hence, generating the RT-MLMC estimator has expected computational cost

$$n_L^{\text{o}}\cdot\sum_{\ell=0}^{L}\mathbb{P}(\widehat{L}=\ell)\cdot\mathcal{O}(2^{\ell+1}\log\frac{1}{\epsilon}) = \mathcal{O}(n_L^{\text{o}}\cdot L\log\frac{1}{\epsilon}). \tag{E.4}$$

□

Proof of Theorem 24. We first show the generic result on SGD with biased gradient estimators. Denote by θ_* the optimal solution to $\min\ F(\theta)$, and $\widetilde{\theta}_*$ is the optimal solution to $\min\ \widetilde{F}(\theta)$, where SG and RT-MLMC estimators are unbiased gradient estimators of $\widetilde{F}(\cdot)$. Based on the triangle inequality, it holds that

$$\begin{aligned}\mathbb{E}\big[F(\widetilde{\theta}_{1:T})-F(\theta_*)\big] &\leq \mathbb{E}\big[F(\widetilde{\theta}_{1:T})-\widetilde{F}(\widetilde{\theta}_{1:T})\big]+\mathbb{E}\big[\widetilde{F}(\widetilde{\theta}_{1:T})-\widetilde{F}(\theta_*)\big]+\mathbb{E}\big[\widetilde{F}(\theta_*)-F(\theta_*)\big]\\ &\leq 2\|\widetilde{F}-F\|_\infty+\mathbb{E}\big[\widetilde{F}(\widetilde{\theta}_{1:T})-\widetilde{F}(\widetilde{\theta}_*)\big],\end{aligned}$$

where the last inequality is because of the sub-optimality of θ_* in terms of the objective $\widetilde{F}$. According to Proposition 8, it holds that

$$\|\widetilde{F} - F\|_\infty \leq \epsilon + \frac{G_{\text{idf}}}{2^L}.$$

According to Lemma 33, for SG or RT-MLMC estimator $V(\theta)$ satisfying $\mathbb{E}\|V(\theta)\|_2^2 \leq M^2$ and if we take step size $\gamma = \frac{\widetilde{D}_*}{M\sqrt{T}}$, it holds that

$$\mathbb{E}\big[\widetilde{F}(\widetilde{\theta}_{1:T}) - \widetilde{F}(\widetilde{\theta}_*)\big] \leq \frac{\widetilde{D}_* M}{\sqrt{T}},$$

where the constant

$$\widetilde{D}_* = \widetilde{F}(\theta_1) - \widetilde{F}(\widetilde{\theta}_*) \leq F(\theta_1) - F(\theta_*) + 2\|\widetilde{F} - F\|_\infty \leq F(\theta_1) - F(\theta_*) + 2\left[\epsilon + \frac{G_{\text{idf}}}{2^L}\right].$$

In summary, the error bound for $\widetilde{\theta}_{1:T}$ becomes

$$\mathbb{E}\big[F(\widetilde{\theta}_{1:T}) - F(\theta_*)\big] \leq 2\left[\epsilon + \frac{G_{\text{idf}}}{2^L}\right] + \frac{M\Big[F(\theta_1) - F(\theta_*) + 2\left[\epsilon + \frac{G_{\text{idf}}}{2^L}\right]\Big]}{\sqrt{T}}.$$

SG Estimator. For SG estimator $V^{\text{SG}}(\theta)$, by Lemma 8, it holds that $M = 2L_f^2[1 + 2\epsilon/(\kappa\eta)]$. To obtain the desired error bound $\mathbb{E}\big[F(\widetilde{\theta}_{1:T}) - F(\theta_*)\big] \leq \delta$, we specify hyper-parameters such that

$$2\left[\epsilon + \frac{G_{\text{idf}}}{2^L}\right] \leq \frac{\delta}{2}, \qquad \frac{M\Big[F(\theta_1) - F(\theta_*) + 2\left[\epsilon + \frac{G_{\text{idf}}}{2^L}\right]\Big]}{\sqrt{T}} \leq \frac{\delta}{2}.$$

We take $\epsilon = \frac{\delta}{8}$ and $L = \log \frac{8G_{\text{idf}}}{\delta}$ to make the relation on the left-hand-side holds. Then $M = \mathcal{O}(1)$. To make the other relation holds, it suffices to take

$$T \geq \frac{4M^2\Big[F(\theta_1) - F(\theta_*) + \frac{\delta}{2}\Big]^2}{\delta^2} = \mathcal{O}(1/\delta^2).$$

RT-MLMC Estimator. For SG estimator $V^{\text{RT-MLMC}}(\theta)$, by Lemma 8, it holds that $M = \frac{96L_f^2}{\kappa\eta} \cdot (2^L\epsilon) + 6(L+1)L_f^2 C$. To obtain the desired error bound $\mathbb{E}\big[F(\widetilde{\theta}_{1:T}) - F(\theta_*)\big] \leq \delta$, we specify hyper-parameters such that

$$2\left[\epsilon + \frac{G_{\text{idf}}}{2^L}\right] \leq \frac{\delta}{2}, \qquad \frac{M\left[F(\theta_1) - F(\theta_*) + 2\left[\epsilon + \frac{G_{\text{idf}}}{2^L}\right]\right]}{\sqrt{T}} \leq \frac{\delta}{2}.$$

Following the same argument as in the SG estimator part, we take $\epsilon = \frac{\delta}{8}$ and $L = \log \frac{8G_{\text{idf}}}{\delta}$. Then $M = \mathcal{O}(\log \frac{1}{\delta})$. To make the other relation holds, it suffices to take

$$T \geq \frac{4M^2\left[F(\theta_1) - F(\theta_*) + \frac{\delta}{2}\right]^2}{\delta^2} = \mathcal{O}((\log 1/\delta)^2/\delta^2).$$

□

In the following, we present a technical lemma that is helpful for the proof of Theorem 25. The proof of this technical lemma follows from [154, Lemma 3.1] and [155, Proposition 4.1].

Lemma 37. (I) *Under Assumption 7*(III)*, it holds that*

$$\left|F^\ell(\theta) - F(\theta)\right| \leq \eta e^{2B/\eta} \cdot 2^{-(\ell+1)}, \quad \forall \theta \in \Theta.$$

(II) *Under Assumptions 7*(III) *and 7*(II)*, it holds that*

$$\left\|\nabla F^\ell(\theta) - \nabla F(\theta)\right\|_2^2 \leq L_f^2 e^{4B/\eta} \cdot 2^{-\ell}, \quad \forall \theta \in \Theta.$$

(III) *Under Assumptions 7*(III) *and 7*(II)*, it holds that*

$$\mathbb{E}\left[\left\|G^\ell(\theta, \zeta^\ell)\right\|_2^2\right] \leq L_f^2 e^{4B/\eta} \cdot 2^{-\ell}, \quad \forall \theta \in \Theta.$$

(IV) *Under Assumptions 7*(III)*, 7*(II) *and 7*(IV)*, it holds that for any* $\ell \geq 0$*,* $F^\ell(\theta)$ *is*

$\overline{S}$*-smooth with*

$$\overline{S} := (S_f^2 + L_f^2/\eta)e^{B/\eta} + L_f^2/\eta e^{2B/\eta}. \tag{E.5}$$

Now we present the formal proof of Theorem 25.

Proof of Theorem 25. At the beginning, it is without the loss of generality to assume that there exists $\mathfrak{c}, \mathfrak{d}$ such that

$$\mathfrak{c}\|\cdot\|_2 \leq \|\cdot\|_\omega \leq \mathfrak{d}\|\cdot\|_2,$$

where $\|\cdot\|_\omega$ is the norm function used in defining the distance generating function for proximal mapping.

(I) We first specify the maximum level L such that $2L_f^2 e^{4B/\eta} \cdot 2^{-L} \leq \frac{1}{2}\epsilon^2$, i.e.,

$$L = \left\lceil \frac{1}{\log 2} \left[\log \frac{4L_f^2 \cdot e^{4B/\eta}}{\epsilon^2} \right] \right\rceil.$$

It suffices to specify hyper-parameters $n_L^{\mathrm{o}}, T, \gamma$ to make

$$2\mathbb{E}\left\| \frac{1}{\gamma}\left[\tilde{\theta} - \mathrm{Prox}_{\tilde{\theta}}\big(\gamma \nabla F^L(\tilde{\theta})\big)\right]\right\|_2^2 \leq \frac{1}{2}\epsilon^2.$$

Before applying Lemma 34 to derive upper bound on the left-hand-side term, it is worth noting that

- According to Lemma 37(IV), the objective $F^L(\theta)$ is $\mathfrak{c}^{-1}\mathfrak{d}\overline{S}$-smooth (with respect to $\|\cdot\|_\omega$) with the constant $\overline{S}$ defined in (E.5).
- According to Lemma 37(III), the term

$$\mathbb{E}\left\|v(\theta_t) - \nabla F^L(\theta_t)\right\|_\omega^2 \leq \frac{2\mathfrak{d}^2(L+1)L_f^2 e^{4B/\eta}}{n_L^{\mathrm{o}}}.$$

Therefore, when taking the step size $\gamma = \kappa/(2\overline{S})$, it holds that

$$2\mathbb{E}\left\|\frac{1}{\gamma}\left[\tilde{\theta} - \mathrm{Prox}_{\tilde{\theta}}\big(\gamma\nabla F^L(\tilde{\theta})\big)\right]\right\|_2^2 \leq \frac{16\mathfrak{c}^{-3}\mathfrak{d}\overline{S}\big(F^L(\theta_1) - \min_\theta\ F^L(\theta)\big)}{\kappa^2 T} + \frac{24\mathfrak{c}^{-2}\mathfrak{d}^2\cdot(L+1)L_f^2 e^{4B/\eta}}{\kappa^2 n_L^{\mathrm{o}}},$$

With the configuration of the following hyper-parameters, one can guarantee the RT-MLMC scheme finds ϵ-stationary point:

$$n_L^{\mathrm{o}} = \left\lceil \frac{96\mathfrak{d}^2(L+1)L_f^2 e^{4B/\eta}}{\kappa^2\mathfrak{c}^2\epsilon^2}\right\rceil, \qquad L = \left\lceil \frac{1}{\log 2}\left[\log\frac{4L_f^2\cdot e^{4B/\eta}}{\epsilon^2}\right]\right\rceil,$$
$$T = \left\lceil \frac{64\mathfrak{d}\overline{S}\big(F^L(\theta_1) - \min_\theta\ F^L(\theta)\big)}{\kappa^2\mathfrak{c}^3\epsilon^2}\right\rceil, \qquad \gamma = \kappa\mathfrak{c}/(2\mathfrak{d}\overline{S}).$$

(II) The proof in this part is a simple corollary from [155, Corollary 4.1]. With the configuration of the following hyper-parameters, one can guarantee the RT-MLMC scheme finds ϵ-stationary point:

$$n_L^{\mathrm{o}} = 1, \qquad L = \left\lceil \frac{1}{\log 2}\left[\log\frac{4L_f^2\cdot e^{4B/\eta}}{\epsilon^2}\right]\right\rceil,$$
$$T = \left\lceil \frac{128\big(F^L(\theta_1) - \min_\theta\ F^L(\theta)\big)\overline{S}M_2}{\epsilon^4}\right\rceil, \qquad \gamma = \sqrt{\frac{2\big(F^L(\theta_1) - \min_\theta\ F^L(\theta)\big)}{\overline{S}TM_2}},$$

where the constant $M_2 := 2(L+1)L_f^2 e^{4B/\eta}$.

□

E.4 Proofs of Technical Results in Section 6.4

We first show the following technical result, from which one can easily derive the main result in Section 6.4.

Proposition 19. *Under Assumption 9, it holds that* $\mathcal{E}_{\widehat{P}}(f;\rho,\eta) = \widetilde{\mathcal{E}}_{\widehat{P}}(f;\rho,\eta) + O(\rho^2)$, *where* $O(\cdot)$ *hides the multiplicative constant dependent on* $\mathbb{E}_{z\sim\widehat{P}}[S(z)]$.

Proof of Proposition 19. By definition,

$$\mathcal{E}_{\widehat{P}}(f;\rho,\eta) = \mathbb{E}_{z\sim\widehat{P}}\left[\sup_{\gamma_z\in\mathcal{P}(\mathcal{Z})}\left\{\mathbb{E}_{z'\sim\gamma_z}[f(z')-f(z)] - \eta\mathbb{E}_{z'\sim\nu_z}\left[\phi\left(\frac{\mathrm{d}\gamma_z(z')}{\mathrm{d}\nu_z(z')}\right)\right]\right\}\right].$$

For any $z' \in \operatorname{supp}\gamma_z \subseteq \mathbb{B}_\rho(z)$, it holds that

$$\begin{aligned}
&|f(z') - f(z) - \nabla f(z)^{\mathrm{T}}(z-z')| = |\nabla f(\tilde{z})^{\mathrm{T}}(z-z') - \nabla f(z)^{\mathrm{T}}(z-z')| \\
=&|(\nabla f(\tilde{z}) - \nabla f(z))^{\mathrm{T}}(z-z')| \le \|\nabla f(\tilde{z}) - \nabla f(z)\|_*\|z-z'\| \\
=&\|\tilde{z}-z\|\cdot\|z-z'\|\cdot S(z) \le S(x)\|z-z'\|^2 \le S(z)\rho^2,
\end{aligned}$$

where the second equality is by the mean value theorem and take $\tilde{z}$ to be some point on the line segment between z and z', and the second inequality is based on the fact that $z' \in \mathbb{B}_\rho(z)$.

Based on the relation above, it holds that

$$\left|\mathcal{E}_{\widehat{P}}(f;\rho,\eta) - \widetilde{\mathcal{E}}_{\widehat{P}}(f;\rho,\eta)\right| \le \mathbb{E}_{z\sim\widehat{P}}[S(z)]\rho^2,$$

where

$$
\begin{aligned}
&\widetilde{\mathcal{E}}_{\widehat{P}}(f;\rho,\eta)\\
=&\mathbb{E}_{z\sim\widehat{P}}\left[\sup_{\gamma_z\in\mathcal{P}(\mathcal{Z})}\left\{\mathbb{E}_{z'\sim\gamma_z}[\nabla f(z)^{\mathrm{T}}(z-z')]-\eta\mathbb{E}_{z'\sim\nu_z}\left[\phi\left(\frac{\mathrm{d}\gamma_z(z')}{\mathrm{d}\nu_z(z')}\right)\right]\right\}\right]\\
=&\mathbb{E}_{z\sim\widehat{P}}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\mathbb{E}_{z'\sim\nu_z}\left[\eta\phi^*\Big(\frac{\nabla f(z)^{\mathrm{T}}(z-z')-\mu}{\eta}\Big)\right]\right\}\right]\\
=&\mathbb{E}_{z\sim\widehat{P}}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\mathbb{E}_{b\sim\beta}\left[\eta\phi^*\Big(\frac{\rho\nabla f(z)^{\mathrm{T}}b-\mu}{\eta}\Big)\right]\right\}\right].
\end{aligned}
$$

By the change of variable technique that replaces μ by $\rho\mu$,

$$
\widetilde{\mathcal{E}}_{\widehat{P}}(f;\rho,\eta)=\rho\cdot\mathbb{E}_{z\sim\widehat{P}}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\frac{1}{\rho/\eta}\mathbb{E}_{b\sim\beta}\left[\phi^*\Big(\frac{\rho}{\eta}\cdot(\nabla f(z)^{\mathrm{T}}b-\mu)\Big)\right]\right\}\right].
$$

The proof is completed. □

Proof of Theorem 26(I). For all $z\in\operatorname{supp}\widehat{P}$, by strong duality theory of ϕ-divergence DRO [279],

$$
\inf_{\mu\in\mathbb{R}}\left\{\mu+\frac{1}{\rho/\eta}\mathbb{E}_{b\sim\beta}\left[\phi^*\Big(\frac{\rho}{\eta}\cdot(\nabla f(z)^{\mathrm{T}}b-\mu)\Big)\right]\right\}=\sup_{\beta'\in\mathcal{P}(\mathbb{B}_1(0))}\left\{\mathbb{E}_{b\sim\beta'}[\nabla f(z)^{\mathrm{T}}b]-\frac{1}{\rho/\eta}\mathbb{D}_\phi(\beta',\beta)\right\}
$$

Since the optimization problem on the right-hand-side (RHS) satisfies Slater's condition, the problem on the left-hand-side (LHS) must contain a non-empty and bounded set of optimal solutions [54, Theorem 2.165]. Subsequently, it holds from [280, Theorem 5.4] that, as $\rho/\eta\to C$,

$$
\inf_{\mu\in\mathbb{R}}\left\{\mu+\frac{1}{\rho/\eta}\mathbb{E}_{b\sim\beta}\left[\phi^*\Big(\frac{\rho}{\eta}\cdot(\nabla f(z)^{\mathrm{T}}b-\mu)\Big)\right]\right\}\to\inf_{\mu\in\mathbb{R}}\left\{\mu+\frac{1}{C}\mathbb{E}_{b\sim\beta}\left[\phi^*\Big(C\cdot(\nabla f(z)^{\mathrm{T}}b-\mu)\Big)\right]\right\}.
$$

Therefore,

$$
\widetilde{\mathcal{E}}_{\widehat{P}}(f;\rho,\eta)=\rho\cdot\mathbb{E}_{z\sim\widehat{P}}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\frac{1}{C}\mathbb{E}_{b\sim\beta}\left[\phi^*\Big(C\cdot(\nabla f(z)^{\mathrm{T}}b-\mu)\Big)\right]\right\}\right]+o(\rho)=\mathcal{R}_1(f;\rho,\eta)+o(\rho).
$$

By the relation above and Proposition 19, we obtain the desired result. □

Proof of Theorem 26(II). According to Proposition 19, it suffices to build the error bound between $\widetilde{\mathcal{E}}_{\widehat{P}}(f;\rho,\eta)$ and $\mathcal{R}_2(f;\rho,\eta)$. As $\rho/\eta \to \infty$, by repeating the proof argument as in Proposition 6, one can show that

$$\begin{aligned}&\mathbb{E}_{z\sim\widehat{P}}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\frac{1}{\rho/\eta}\mathbb{E}_{b\sim\beta}\left[\phi^*\Big(\frac{\rho}{\eta}\cdot(\nabla f(z)^{\mathrm{T}}b-\mu)\Big)\right]\right\}\right]\\ =&\mathbb{E}_{z\sim\widehat{P}}\left[\max_{b\in\mathbb{B}_1(0)}[\nabla f(z)^{\mathrm{T}}b]\right]+o(1)=\mathbb{E}_{z\sim\widehat{P}}\left[\|\nabla f(z)\|_*\right]+o(1).\end{aligned}$$

As such,

$$\left|\widetilde{\mathcal{E}}_{\widehat{P}}(f;\rho,\eta)-\mathcal{R}_2(f;\rho,\eta)\right|=o(\rho).$$

This completes the proof.

□

Proof of Theorem 26(III). Recall that

$$\begin{aligned}\widetilde{\mathcal{E}}_{\widehat{P}}(f;\rho,\eta)&=\rho\cdot\mathbb{E}_{z\sim\widehat{P}}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\frac{1}{\rho/\eta}\mathbb{E}_{b\sim\beta}\left[\phi^*\Big(\frac{\rho}{\eta}\cdot(\nabla f(z)^{\mathrm{T}}b-\mu)\Big)\right]\right\}\right]\\&=\rho\cdot\mathbb{E}_{z\sim\widehat{P}}\left[\sup_{\beta'\in\mathcal{P}(\mathbb{B}_1(0))}\left\{\mathbb{E}_{b\sim\beta'}[\nabla f(z)^{\mathrm{T}}b]-\frac{1}{\rho/\eta}\mathbb{D}_\phi(\beta',\beta)\right\}\right]\end{aligned}$$

and it suffices to analyze the approximation of $\sup_{\beta'\in\mathcal{P}(\mathbb{B}_1(0))}\left\{\mathbb{E}_{b\sim\beta'}[\nabla f(z)^{\mathrm{T}}b]-\frac{1}{\rho/\eta}\mathbb{D}_\phi(\beta',\beta)\right\}$ for $z\in\operatorname{supp}\widehat{P}$ when $\rho/\eta\to 0$.

Note that we can re-write

$$\sup_{\beta'\in\mathcal{P}(\mathbb{B}_1(0))}\left\{\mathbb{E}_{b\sim\beta'}[\nabla f(z)^{\mathrm{T}}b]-\frac{1}{\rho/\eta}\mathbb{D}_\phi(\beta',\beta)\right\}=\sup_{\Upsilon\geq 0:\ \mathbb{E}_\beta[\Upsilon]=1}\left\{\mathbb{E}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}}b\cdot\Upsilon(b)]-\frac{1}{\rho/\eta}\cdot\mathbb{E}_{b\sim\beta}[\phi(\Upsilon(b))]\right\},$$

where Υ is a non-negative random variable satisfying $\mathbb{E}_\beta[\Upsilon]=1$. We use the change-of-variable technique to define $\overline{\Delta}=(\rho/\eta)^{-1}\cdot(\Upsilon-1)$, then by the relation $\mathbb{E}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}}b]=0$,

the optimization above can be equivalently reformulated as

$$\frac{\rho}{\eta}\cdot\sup_{\overline{\Delta}\geq-(\rho/\eta)^{-1}:\ \mathbb{E}_\beta[\overline{\Delta}]=0}\Big\{\mathbb{E}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}}b\cdot\overline{\Delta}(b)]-(\rho/\eta)^{-2}\cdot\mathbb{E}_{b\sim\beta}[\phi(1+\rho/\eta\cdot\overline{\Delta}(b))]\Big\}.$$

We take a feasible solution $\overline{\Delta}(b)=a\cdot\nabla f(z)^{\mathrm{T}}b$ with some constant $a>0$ provided that $a\cdot\|\nabla f(z)\|_*\leq(\rho/\eta)^{-1}$. Then, it holds that

$$\begin{aligned}&\sup_{\overline{\Delta}\geq-(\rho/\eta)^{-1}:\ \mathbb{E}_\beta[\overline{\Delta}]=0}\Big\{\mathbb{E}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}}b\cdot\overline{\Delta}(b)]-(\rho/\eta)^{-2}\cdot\mathbb{E}_{b\sim\beta}[\phi(1+\rho/\eta\cdot\overline{\Delta}(b))]\Big\}\\ \geq&\sup_{a\geq0:\ a\cdot\|\nabla f(z)\|_*\leq(\rho/\eta)^{-1}}\Big\{a\cdot\mathbb{V}\mathrm{ar}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}}b]-(\rho/\eta)^{-2}\cdot\mathbb{E}_{b\sim\beta}[\phi(1+a\rho/\eta\cdot\nabla f(z)^{\mathrm{T}}b)]\Big\}\end{aligned}\tag{E.6}$$

Since $\phi(t)$ is two times continuously differentiable at $t=1$, as $\rho/\eta\to0$, the following convergence holds uniformly for any bounded $a\cdot\nabla f(z)^{\mathrm{T}}b$:

$$(\rho/\eta)^{-2}\cdot\phi(1+a\rho/\eta\cdot\nabla f(z)^{\mathrm{T}}b)\to a^2(\nabla f(z)^{\mathrm{T}}b)^2\phi''(1)/2.$$

Consequently, for any $\epsilon>0$, there exists $\delta_0>0$ such that as long as $\rho/\eta<\delta_0$, (E.6) can be lower bounded as the following:

$$\begin{aligned}\text{(E.6)}\geq&\sup_{a\geq0:\ a\cdot\|\nabla f(z)\|_*\leq(\rho/\eta)^{-1}}\Big\{a\cdot\mathbb{V}\mathrm{ar}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}}b]-(1+\epsilon)a^2\cdot\mathbb{V}\mathrm{ar}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}}b]\cdot\phi''(1)/2\Big\}\\ =&\frac{\mathbb{V}\mathrm{ar}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}}b]}{2(1+\epsilon)\phi''(1)}.\end{aligned}$$

Since ϵ can be arbitrarily small, it holds that

$$\widetilde{\mathcal{E}}_{\widehat{P}}(f;\rho,\eta)\geq\frac{\rho^2}{2\eta\cdot\phi''(1)}\cdot\mathbb{E}_{z\sim\widehat{P}}\Big[\mathbb{V}\mathrm{ar}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}}b]\Big]+o(\rho).$$

For the upper bound, by strong duality result,

$$
\begin{aligned}
&\sup_{\overline{\Delta}\geq -(\rho/\eta)^{-1}:\ \mathbb{E}_\beta[\overline{\Delta}]=0} \Big\{\mathbb{E}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}} b\cdot \overline{\Delta}(b)] - (\rho/\eta)^{-2}\cdot \mathbb{E}_{b\sim\beta}[\phi(1+\rho/\eta\cdot\overline{\Delta}(b))]\Big\}\\
=&\min_{\overline{\mu}}\ \Big\{\sup_{\overline{\Delta}\geq -(\rho/\eta)^{-1}} \mathbb{E}_{b\sim\beta}[(\nabla f(z)^{\mathrm{T}} b+\overline{\mu})\cdot \overline{\Delta}(b)] - (\rho/\eta)^{-2}\cdot \mathbb{E}_{b\sim\beta}[\phi(1+\rho/\eta\cdot\overline{\Delta}(b))]\Big\}\\
=&\min_{\overline{\mu}}\ \Big\{\mathbb{E}_{b\sim\beta}\Big[\sup_{\overline{\Delta}\geq -(\rho/\eta)^{-1}} (\nabla f(z)^{\mathrm{T}} b+\overline{\mu})\cdot \overline{\Delta} - (\rho/\eta)^{-2}\cdot \phi(1+\rho/\eta\cdot\overline{\Delta})\Big]\Big\}\\
\leq&\mathbb{E}_{b\sim\beta}\Big[\sup_{\overline{\Delta}\geq -(\rho/\eta)^{-1}} \nabla f(z)^{\mathrm{T}} b\cdot \overline{\Delta} - (\rho/\eta)^{-2}\cdot \phi(1+\rho/\eta\cdot\overline{\Delta})\Big]
\end{aligned}
$$

Since ϕ is convex with $\phi''(1)>0$, it holds that the family of continuous functions

$$
s_\delta(y) := \sup_{\overline{\Delta}\geq -(\rho/\eta)^{-1}} y\cdot\overline{\Delta} - (\rho/\eta)^{-2}\cdot\phi(1+\rho/\eta\cdot\overline{\Delta})
$$

converges uniformly on compact sets to

$$
s_0(y) := \sup_{\overline{\Delta}} y\cdot\overline{\Delta} - \frac{\overline{\Delta}^2\phi''(1)}{2} = \frac{y^2}{2\phi''(1)}.
$$

Consequently,

$$
\widetilde{\mathcal{E}}_{\widehat{P}}(f;\rho,\eta) \leq \frac{\rho^2}{\eta}\mathbb{E}_{z\sim\widehat{P}}\mathbb{E}_{b\sim\beta}\Big[\sup_{\overline{\Delta}\geq -(\rho/\eta)^{-1}} \nabla f(z)^{\mathrm{T}} b\cdot\overline{\Delta} - (\rho/\eta)^{-2}\cdot\phi(1+\rho/\eta\cdot\overline{\Delta})\Big].
$$

When we take $\rho/\eta\to 0$ both sides, the RHS becomes

$$
\frac{\rho^2}{\eta}\mathbb{E}_{z\sim\widehat{P}}\mathbb{E}_{b\sim\beta}\Big[\frac{(\nabla f(z)^{\mathrm{T}} b)^2}{2\phi''(1)}\Big] + o(\rho) = \frac{\rho^2}{2\eta\cdot\phi''(1)}\cdot\mathbb{E}_{z\sim\widehat{P}}\Big[\mathbb{V}\mathrm{ar}_{b\sim\beta}[\nabla f(z)^{\mathrm{T}} b]\Big] + o(\rho).
$$

Combining lower and upper bounds gives our desired result.

□

E.5 Proof of Technical Result in Section 6.5

Proof of Lemma 10. Let $\mathcal{C}(\mathcal{G}) = \Big\{(c_i^j(\cdot)) : \ i \in [n]\Big\}$ be a $(\epsilon, \|\cdot\|_\infty)$-cover of the set $\mathcal{G}_{|S}$. Define the operator

$$\widetilde{c}_i^j = \inf_{\mu \in \mathbb{R}} \left\{ \mu + \mathbb{E}_{b\sim\beta}\left[(\eta\phi)^*\Big(c_i^j(b) - \mu\Big)\right] \right\}.$$

We now claim that $\mathcal{C}(\mathcal{G}_{\text{adv}}) = \Big\{\widetilde{c}_i^j : \ i \in [n]\Big\}$ is a $(\epsilon, |\cdot|)$-cover of the set $\mathcal{G}_{\text{adv}|S}$. Indeed, for any $\theta \in \Theta$, there exists an index (by definition) $j(\theta)$ such that

$$\max_{i\in[n]} \max_{b\in\mathbb{B}_1(0)} \Big|\ell(g_\theta(x_i+b), y_i) - c_i^{j(\theta)}(x_i, y_i)\Big| = \max_{i\in[n]} \|\ell(g_\theta(x_i+\cdot), y_i) - c_i^{j(\theta)}(\cdot)\|_\infty \le \epsilon.$$

Define the functional $T : \ \mathbb{R}^{\mathbb{B}_1(0)} \to \mathbb{R}$ as

$$\begin{aligned} T(g) &= \inf_{\mu} \left\{ \mu + \mathbb{E}_{b\sim\beta}\left[(\eta\phi)^*\Big(g(b) - \mu\Big)\right] \right\} \\ &= \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{\mathbb{P}}[g] - \eta\mathbb{E}_{b\sim\mathbb{P}}\left[\phi\left(\frac{\mathrm{d}\mathbb{P}(b)}{\mathrm{d}\beta(b)}\right)\right]\right\}. \end{aligned}$$

Since ϕ is strictly convex, its directional derivative is well-defined, which is denoted as

$$\nabla T(g)V = \mathbb{E}_{b\sim\mathbb{P}_g^*}\Big[V(b)\Big],$$

where

$$\mathbb{P}_g^* = \operatorname*{arg\,max}_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{\mathbb{P}}[g] - \eta\mathbb{E}_{b\sim\mathbb{P}}\left[\phi\left(\frac{\mathrm{d}\mathbb{P}(b)}{\mathrm{d}\beta(b)}\right)\right]\right\}.$$

For fixed i, with slight abuse of notation, let $\tau_i(\cdot;t) = t\ell(g_\theta(x_i+\cdot), y_i) + (1-t)c_i^{j(\theta)}(\cdot)$. Therefore, we have that for $g \in \mathcal{G}_{\text{adv}}$,

$$\begin{aligned}
\max_{i\in[n]} \left| g(x_i, y_i) - \widetilde{c}_i^{j(\theta)} \right| &= \max_{i\in[n]} \left| T(\ell(g_\theta(x_i+\cdot), y_i)) - T(c_i^{j(\theta)}(\cdot)) \right| \\
&\leq \max_{i\in[n]} \sup_{t\in[0,1]} \left| \nabla T(\tau_i(\cdot;t))(\ell(g_\theta(x_i+\cdot), y_i) - c_i^{j(\theta)}(\cdot)) \right| \\
&= \max_{i\in[n]} \sup_{t\in[0,1]} \left| \mathbb{E}_{\varepsilon\sim\mathbb{P}^*_{\tau_i(\cdot;t)}} \left[\ell(g_\theta(x_i+\varepsilon), y_i) - c_i^{j(\theta)}(\varepsilon) \right] \right| \\
&\leq \max_{i\in[n]} \| \ell(g_\theta(x_i+\cdot), y_i) - c_i^{j(\theta)}(\cdot) \|_\infty \leq \epsilon.
\end{aligned}$$

The proof is completed. □

E.6 Proof of Strong Duality for ∞-Type Casual Optimal Transport DRO

Let us consider the ∞-type Casual Optimal transport DRO problem

$$\min_{\theta}\left\{\sup_{\mathbb{P},\gamma}\mathbb{E}_{(x,z)\sim\mathbb{P}}[\Psi(f_\theta(x),z)]:\begin{array}{l}((\widehat{X},\widehat{Z}),(X,Z))\sim\gamma\implies X\perp\widehat{Z}\mid\widehat{X},\\ \operatorname{ess\,sup}_\gamma\ \|(\widehat{X},\widehat{Z})-(X,Z)\|\le\rho,\mathrm{Proj}_{1\#}\gamma=\widehat{P},\mathrm{Proj}_{2\#}\gamma=\mathbb{P}\end{array}\right\},\tag{E.7}$$

where the norm $\|(\widehat{X},\widehat{Z})-(X,Z)\|\triangleq\|\widehat{X}-X\|+\infty\cdot\mathbf{1}\{\widehat{Z}\neq Z\}$, meaning we take into account only the distribution shift of the covariate and omit the random vector distribution shift. Here the transport mapping γ is said to be *casual* since $((\widehat{X},\widehat{Z}),(X,Z))\sim\gamma$ implies $X\perp\widehat{Z}\mid\widehat{X}$, and it additionally satisfies the ∞-type optimal transport constraint with transportation budget ρ. In the following, we derive the strong dual reformulation of (E.7). We expand its objective as

$$\begin{aligned}&\mathbb{E}_{(x,z)\sim\mathbb{P}}[\Psi(f_\theta(x),z)]=\mathbb{E}_{((\widehat{x},\widehat{z}),(x,z))\sim\gamma}[\Psi(f_\theta(x),z)]\\ =&\mathbb{E}_{\widehat{x}\sim\gamma(\widehat{x})}\mathbb{E}_{x\sim\gamma(x|\widehat{x})}\mathbb{E}_{\widehat{z}\sim\gamma(\widehat{z}|x,\widehat{x})}\mathbb{E}_{z\sim\gamma(z|\widehat{z},x,\widehat{x})}[\Psi(f_\theta(x),z)]\\ =&\mathbb{E}_{\widehat{x}\sim\gamma(\widehat{x})}\mathbb{E}_{x\sim\gamma(x|\widehat{x})}\mathbb{E}_{\widehat{z}\sim\gamma(\widehat{z}|\widehat{x})}\mathbb{E}_{z\sim\delta_{\widehat{z}}}[\Psi(f_\theta(x),z)]\\ =&\mathbb{E}_{\widehat{x}\sim\gamma(\widehat{x})}\mathbb{E}_{x\sim\gamma(x|\widehat{x})}\mathbb{E}_{\widehat{z}\sim\gamma(\widehat{z}|\widehat{x})}[\Psi(f_\theta(x),\widehat{z})]\end{aligned}$$

where the first and second equality is by the law of total probability, the third equality is because γ satisfies *casual* property and we impose infinity transportation cost for moving $\widehat{z}$

to other locations. Since $\gamma(\widehat{x}) = \widehat{\mathbb{P}}_{\widehat{x}}$ and $\gamma(\widehat{z} \mid \widehat{x}) = \widehat{P}_{\widehat{z}|\widehat{x}}$, we are able to reformulate (E.7) as

$$\min_{\theta} \left\{ \sup_{\mathbb{P},\gamma} \mathbb{E}_{\widehat{x}\sim\widehat{\mathbb{P}}_{\widehat{x}}} \mathbb{E}_{x\sim\gamma(x|\widehat{x})} \mathbb{E}_{\widehat{z}\sim\widehat{P}_{\widehat{z}|\widehat{x}}} [\Psi(f_\theta(x), \widehat{z})] : \begin{array}{l} \operatorname{ess\,sup}_\gamma \ \|\widehat{X} - X\| \le \rho, \\ \mathrm{Proj}_{1\#}\gamma = \widehat{P}, \mathrm{Proj}_{2\#}\gamma = \mathbb{P} \end{array} \right\} \tag{E.8}$$

$$= \min_{\theta} \left\{ \mathbb{E}_{\widehat{x}\sim\widehat{P}_{\widehat{X}}} \left[\sup_{x'\in\mathbb{B}_\rho(\widehat{x})} \mathbb{E}_{\widehat{z}\sim\widehat{P}_{\widehat{Z}|\widehat{X}=\widehat{x}}} [\Psi(f_\theta(x'), \widehat{z})] \right] \right\}, \tag{E.9}$$

where the last equality is by the ∞-Wasserstein DRO strong duality result, adopted from [123, Lemma EC.2], with loss $\mathbb{E}_{\widehat{z}\sim\widehat{P}_{\widehat{Z}|\widehat{X}=\widehat{x}}}[\Psi(f_\theta(x'), \widehat{z})]$.

Following similar procedure, one can express (6.26) using its primal reformulation that involves ϕ-divergence regularization:

$$\min_{\theta} \left\{ \sup_{\mathbb{P},\gamma} \mathbb{E}_{(x,z)\sim\mathbb{P}}[\Psi(f(x), z)] - \eta \mathbb{D}_\phi(\gamma, \gamma_0) : \begin{array}{l} ((\widehat{X}, \widehat{Z}), (X, Z)) \sim \gamma \implies X \perp \widehat{Z} \mid \widehat{X}, \\ \operatorname{ess\,sup}_\gamma \ \|(\widehat{X}, \widehat{Z}) - (X, Z)\| \le \rho, \mathrm{Proj}_{1\#}\gamma = \widehat{P}, \mathrm{Proj}_{2\#}\gamma = \mathbb{P} \end{array} \right\}, \tag{6.26-Primal}$$

where the reference measure γ_0 satisfies the *bicasual* property, $\gamma_0(x \mid \widehat{x}) \equiv \nu_{\widehat{x}}(x), \forall x$ and $\gamma_0(z \mid \widetilde{z}, \widehat{x}, x) \equiv \gamma_0(z \mid \widetilde{z}) \equiv \delta_{\widetilde{z}}(z), \forall z$. Namely, its joint distribution is decomposed as

$$\begin{aligned} \gamma_0((\widehat{x}, \widehat{z}), (x, z)) &= \gamma_0(\widehat{x})\gamma_0(x \mid \widehat{x})\gamma_0(\widehat{z} \mid \widehat{x}, x)\gamma_0(z \mid \widehat{z}, \widehat{x}, x) \\ &= \gamma_0(\widehat{x})\gamma_0(x \mid \widehat{x})\gamma_0(\widehat{z} \mid \widehat{x})\gamma_0(z \mid \widehat{z}) = \gamma_0(\widehat{x})\nu_{\widehat{x}}(x)\gamma_0(\widehat{z} \mid \widehat{x})\delta_{\widehat{z}}(z). \end{aligned}$$

E.7 Implementation Details for Loss in Section 6.2.2

For the loss $f(\cdot)$ displayed in Section 6.2.2, we take the loss $f(z) = (g(z) - 0)^2$, where $g : \mathbb{R} \to \mathbb{R}$ is a feed-forward neural network function. The structure of g is as follows. We first take a basis expansion to form $z' = (z, \sqrt{|z|}, z^2, \sin(z), \cos(z)) \in \mathbb{R}^5$. Then take

$$g(z) = W_4 \cdot \mathrm{Sigmoid}(W_3 \cdot \mathrm{Sp}(W_2 \cdot \mathrm{Sp}(W_1 z'))),$$

where $\mathrm{Sig}(\cdot)$ and $\mathrm{Sp}(\cdot)$ are the sigmoid and softplus activation functions, respectively. Weight matrics $W_1 \in \mathbb{R}^{512\times 5}, W_2 \in \mathbb{R}^{512\times 512}, W_2 \in \mathbb{R}^{10\times 512}, W_4 \in \mathbb{R}^{10\times 1}$, and the entries of W_2, W_3, W_4 follow i.i.d. from $\mathcal{N}(0,1)$ whereas that of W_1 follow $\mathcal{N}(0, 0.25)$. This example can be viewed as adversarial robust supervised learning for using a neural network to fit a constant function at the origin.

REFERENCES

[1] Adachi, S., Iwata, S., Nakatsukasa, Y., and Takeda, A. (2017). Solving the trust-region subproblem by a generalized eigenvalue problem. *SIAM Journal on Optimization*, 27(1):269–291.

[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] Ajalloeian, A. and Stich, S. U. (2020). On the convergence of sgd with biased gradients. *arXiv preprint arXiv:2008.00051*.

[5] Alizadeh, F. (1995). Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM journal on Optimization*, 5(1):13–51.

[6] 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.

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

[8] Arcones, M. A. and Gine, E. (1992). On the bootstrap of u and v statistics. *The Annals of Statistics*, pages 655–674.

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

[10] Attias, I., Kontorovich, A., and Mansour, Y. (2019). Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, pages 162–183. PMLR.

[11] Awasthi, P., Frank, N., and Mohri, M. (2020). Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pages 431–441. PMLR.

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

[13] Azizian, W., Iutzeler, F., and Malick, J. (2023). Regularization for wasserstein distributionally robust optimization. *ESAIM: Control, Optimisation and Calculus of Variations*, 29:33.

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

[15] 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.

[16] Bajwa, W. U. and Mixon, D. G. (2012). Group model selection using marginal correlations: The good, the bad and the ugly. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing*, pages 494–501.

[17] Balagopalan, A., Novikova, J., Mcdermott, M. B., Nestor, B., Naumann, T., and Ghassemi, M. (2020). Cross-language aphasia detection using optimal transport domain adaptation. In *Machine Learning for Health Workshop*, pages 202–219. PMLR.

[18] 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.

[19] Bartl, D. and Mendelson, S. (2022). Structure preservation via the Wasserstein distance. *arXiv preprint arXiv:2209.07058*.

[20] Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249.

[21] Barvinok, A. I. (1995). Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13:189–202.

[22] Bayraksan, G. and Love, D. K. (2015). Data-driven stochastic programming using phi-divergences. In *The Operations Research Revolution*, pages 1–19. INFORMS.

[23] Beale, E. M. L. and Tomlin, J. A. (1970). Special facilities in a general mathematical programming system for non-convex problems using ordered sets of variables. *Operations Research*, 69(447-454):99.

[24] 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.

[25] Ben-Tal, A. and Nemirovski, A. (2021). Lectures on modern convex optimization 2020. *SIAM, Philadelphia*.

[26] Ben-Tal, A. and Teboulle, M. (1987). Penalty functions and duality in stochastic programming via φ-divergence functionals. *Mathematics of Operations Research*, 12(2):224–240.

[27] Berk, L. and Bertsimas, D. (2019). Certifiably optimal sparse principal component analysis. *Mathematical Programming Computation*, 11:381–420.

[28] Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

[29] Bertsekas, D. P. (1997). Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334.

[30] Bertsimas, D., Cory-Wright, R., and Pauphilet, J. (2022). Solving large-scale sparse PCA to certifiable (near) optimality. *Journal of Machine Learning Research*, 23:13–1.

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

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

[33] Bertsimas, D., Pauphilet, J., and Van Parys, B. (2021). Sparse classification: a scalable discrete optimization perspective. *Machine Learning*, 110:3177–3209.

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

[35] Bertsimas, D. and Tsitsiklis, J. (1993). Simulated annealing. *Statistical science*, 8(1):10–15.

[36] Binkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying MMD GANs. In *International Conference on Learning Representations*.

[37] Birkhoff, G. (1946). Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman, Ser. A*, 5:147–154.

[38] Blackwell, D. and Ryll-Nardzewski, C. (1963). Non-existence of everywhere proper conditional distributions. *The Annals of Mathematical Statistics*, 34(1):223–225.

[39] Blair, C. (1985). Problem complexity and method efficiency in optimization (a. s. nemirovsky and d. b. yudin). *SIAM Review*, 27(2):264–265.

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

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

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

[47] 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.

[48] Blanchet, J. and Shapiro, A. (2023). Statistical limit theorems in distributionally robust optimization. In *2023 Winter Simulation Conference (WSC)*, pages 31–45. IEEE.

[49] Blanchet, J. H. and Glynn, P. W. (2015). Unbiased monte carlo for optimization and functions of expectations via multi-level randomization. In *2015 Winter Simulation Conference (WSC)*, pages 3656–3667.

[50] Boedihardjo, M. (2025). Sharp bounds for max-sliced wasserstein distances. *Foundations of Computational Mathematics*.

[51] Bonami, P., Biegler, L. T., Conn, A. R., Cornuéjols, G., Grossmann, I. E., Laird, C. D., Lee, J., Lodi, A., Margot, F., Sawaya, N., et al. (2008). An algorithmic framework for convex mixed integer nonlinear programs. *Discrete optimization*, 5(2):186–204.

[52] Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

[53] Bonnans, J. F. and Shapiro, A. (2000). Stability and sensitivity analysis. *Perturbation Analysis of Optimization Problems*, pages 260–400.

[54] Bonnans, J. F. and Shapiro, A. (2013). *Perturbation analysis of optimization problems*. Springer Science & Business Media.

[55] Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45.

[56] 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.

[57] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

[58] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[59] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

[60] 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.

[61] Butler, R. W. (2007). *Saddlepoint approximations with applications*, volume 22. Cambridge University Press.

[62] Caponnetto, A., Micchelli, C. A., Pontil, M., and Ying, Y. (2008). Universal multi-task kernels. *Journal of Machine Learning Research*, 9(52):1615–1646.

[63] Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee.

[64] Carriere, M., Cuturi, M., and Oudot, S. (2017). Sliced Wasserstein kernel for persistence diagrams. In *International conference on machine learning*, pages 664–673. PMLR.

[65] Chan, S. O., Papailliopoulos, D., and Rubinstein, A. (2016). On the approximability of sparse PCA. In *Conference on Learning Theory*, pages 623–646.

[66] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3).

[67] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

[68] Chen, J. and Luss, R. (2018). Stochastic gradient descent with biased but consistent gradient estimators. *arXiv preprint arXiv:1807.11880*.

[69] Chen, R. and Paschalidis, I. C. (2019). Selecting optimal decisions via distributionally robust nearest-neighbor regression. In *Advances in Neural Information Processing Systems*.

[70] Chen, X., He, N., Hu, Y., and Ye, Z. (2024a). Efficient algorithms for a class of stochastic hidden convex optimization and its applications in network revenue management. *Operations Research*.

[71] Chen, X., Hu, Y., and Zhao, M. (2024b). Landscape of policy optimization for finite horizon mdps with general state and action. *arXiv preprint arXiv:2409.17138*.

[72] 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.

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

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

[75] Cheng, X. and Cloninger, A. (2022). Classification logit two-sample testing by neural networks for differentiating near manifold densities. *IEEE Transactions on Information Theory*, 68(10):6631–6662.

[76] Cheng, X. and Xie, Y. (2021a). Kernel mmd two-sample tests for manifold data. *arXiv preprint arXiv: 2105.03425*.

[77] Cheng, X. and Xie, Y. (2021b). Neural tangent kernel maximum mean discrepancy. In *Advances in Neural Information Processing Systems*.

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

[79] 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.

[80] Chwialkowski, K. P., Ramdas, A., Sejdinovic, D., and Gretton, A. (2015). Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, volume 28.

[81] Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, T. (2018). Deep learning for classical japanese literature. In *NeurIPS*. Kuzushiji-MNIST.

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

[83] 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.

[84] Conn, A. R., Gould, N. I., and Toint, P. L. (2000). *Trust region methods*. SIAM.

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

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

[87] Courty, N., Flamary, R., and Tuia, D. (2014b). Domain adaptation with regularized optimal transport. In [86], pages 274–289.

[88] 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.

[89] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience.

[90] Cressie, N. and Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 46(3):440–464.

[91] Cudeck, R. (2000). Exploratory factor analysis. In *Handbook of applied multivariate statistics and mathematical modeling*, pages 265–296. Elsevier.

[92] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transportation distances. In *Advances in neural information processing systems*.

[93] del Barrio, E., Cuesta-Albertos, J. A., Matrain, 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.

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

[95] 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.

[96] Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE signal processing magazine*, 29(6):141–142.

[97] Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. (2019). Max-sliced Wasserstein distance and its use

for gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10648–10656.

[98] Deshpande, I., Zhang, Z., and Schwing, A. G. (2018). Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3483–3491.

[99] Dey, S. S., Kocuk, B., and Santana, A. (2019). Convexifications of rank-one-based substructures in qcqps and applications to the pooling problem. *Journal of Global Optimization*, 77(2):227–272.

[100] Dey, S. S., Mazumder, R., and Wang, G. (2022a). Using ℓ_1-relaxation and integer programming to obtain dual bounds for sparse PCA. *Operations Research*, 70(3):1914–1932.

[101] Dey, S. S., Molinaro, M., and Wang, G. (2022b). Solving sparse principal component analysis with global support. *Mathematical Programming*, pages 1–39.

[102] Deza, M. M., Laurent, M., and Weismantel, R. (1997). *Geometry of cuts and metrics*, volume 2. Springer.

[103] Diamond, S. and Boyd, S. (2016). Cvxpy: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.

[104] Diao, S. and Sen, S. (2020). Distribution-free algorithms for learning enabled optimization with non-parametric estimation. *Management Science*, 66(3):1025–1044.

[105] 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.

[106] Dominici*, D. E. (2003). The inverse of the cumulative standard normal probability function. *Integral Transforms and Special Functions*, 14(4):281–292.

[107] Duchi, J. and Namkoong, H. (2019). Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55.

[108] 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).

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

[110] 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.

[111] 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.

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

[113] Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

[114] Fletcher, R. and Leyffer, S. (1994). Solving mixed integer nonlinear programs by outer approximation. *Mathematical programming*, 66:327–349.

[115] 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*, pages 1737–1746.

[116] Fournier, N. and Guillin, A. (2014). On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738.

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

[118] Fukumizu, K., Gretton, A., Lanckriet, G., Schölkopf, B., and Sriperumbudur, B. K. (2009). Kernel choice and classifiability for rkhs embeddings of probability distributions. *Advances in neural information processing systems*, 22.

[119] Gally, T. and Pfetsch, M. E. (2016). Computing restricted isometry constants via mixed-integer semidefinite programming. *Preprint Available at Optimization Online*.

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

[121] Gao, R., Arora, R., and Huang, Y. (2022a). Data-driven multistage distributionally robust optimization with nested distance: Time consistency and tractable dynamic reformulations. *Available at Optimization Online*.

[122] Gao, R., Chen, X., and Kleywegt, A. J. (2022b). Wasserstein distributionally robust optimization and variation regularization. In [123].

[123] Gao, R., Chen, X., and Kleywegt, A. J. (2022c). Wasserstein distributionally robust optimization and variation regularization. *Operations Research*.

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

[125] Gao, R. and Kleywegt, A. (2023b). Distributionally robust stochastic optimization with Wasserstein distance. In [124], pages 603–655.

[126] Gao, R., Xie, L., Xie, Y., and Xu, H. (2018). Robust hypothesis testing using Wasserstein uncertainty sets. *Advances in Neural Information Processing Systems*, 31.

[127] Genevay, A. (2019). *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE).

[128] 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.

[129] 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.

[130] 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.

[131] Genevay, A., Peyré, G., and Cuturi, M. (2018b). Learning generative models with sinkhorn divergences. In [130], pages 1608–1617.

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

[133] Ghosh, S., Squillante, M., and Wollega, E. (2018). Efficient stochastic gradient descent for learning with distributionally robust optimization. *arXiv preprint arXiv:1805.08728*.

[134] Gin, E. and Nickl, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, USA.

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

[136] Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

[137] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples.

[138] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012a). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.

[139] 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.

[140] Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213.

[141] 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.

[142] Hara, S., Morimura, T., Takahashi, T., Yanagisawa, H., and Suzuki, T. (2015). A consistent method for graph based anomaly localization. In *Artificial intelligence and statistics*, pages 333–341.

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

[144] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

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

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

[147] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

[148] Hildreth, C. (1957). A quadratic programming procedure. *Naval Research Logistics Quarterly*, 4(1):79–85.

[149] Hotelling, H. (1931). The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2:360–378.

[150] 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.

[151] Hu, B., Seiler, P., and Lessard, L. (2017). Analysis of biased stochastic gradient descent using sequential semidefinite programs. *arXiv preprint arXiv:1711.00987*.

[152] Hu, J., Liu, X., Wen, Z., and Yuan, Y. (2019). A brief introduction to manifold optimization. *arXiv preprint arXiv:1906.05450*.

[153] Hu, X., Prashanth, L., György, A., and Szepesvari, C. (2016). (bandit) convex optimization with biased noisy gradient oracles. In *Artificial Intelligence and Statistics*, pages 819–828. PMLR.

[154] 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.

[155] Hu, Y., Chen, X., and He, N. (2021a). On the bias-variance-cost tradeoff of stochastic optimization. *Advances in Neural Information Processing Systems*, 34:22119–22131.

[156] Hu, Y., Chen, X., and He, N. (2021b). On the bias-variance-cost tradeoff of stochastic optimization. In [155], pages 22119–22131.

[157] Hu, Y., Wang, J., Chen, X., and He, N. (2024). Multi-level monte-carlo gradient methods for stochastic optimization with biased oracles. *arXiv preprint arXiv:2408.11084*.

[158] Hu, Y., Wang, J., Xie, Y., Krause, A., and Kuhn, D. (2023). Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems*, 36.

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

[160] Hu, Y., Zhang, S., Chen, X., and He, N. (2020c). Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. In [159], pages 2759–2770.

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

[162] Huang, M., Ma, S., and Lai, L. (2021a). Projection robust Wasserstein barycenters. In *International Conference on Machine Learning*, pages 4456–4465. PMLR.

[163] 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.

[164] Huber, P. J. (1965). A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758.

[165] Idé, T., Lozano, A. C., Abe, N., and Liu, Y. (2009). Proximity-based anomaly detection using sparse structure learning. In *Proceedings of the 2009 SIAM international conference on data mining*, pages 97–108.

[166] Idé, T., Papadimitriou, S., and Vlachos, M. (2007). Computing correlation anomaly scores using stochastic nearest neighbors. In *Seventh IEEE international conference on data mining*, pages 523–528.

[167] Ji, K., Yang, J., and Liang, Y. (2022). Theoretical convergence of multi-step model-agnostic meta-learning. *Journal of machine learning research*, 23(29):1–41.

[168] Jiang, B. and Liu, Y.-F. (2024). A riemannian exponential augmented lagrangian method for computing the projection robust Wasserstein distance. *Advances in Neural Information Processing Systems*, 36.

[169] 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*.

[170] Jitkrittum, W., Szabó, Z., Chwialkowski, K. P., and Gretton, A. (2016). Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems*, 29.

[171] Jolliffe, I. (1986). *Principal Component Analysis*. Springer Verlag.

[172] 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*.

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

[174] Khim, J. and Loh, P.-L. (2018). Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*.

[175] Kirchler, M., Khorasani, S., Kloft, M., and Lippert, C. (2020). Two-sample testing using deep learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1398. PMLR.

[176] Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on optimization*, 12(2):479–502.

[177] Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. (2019). Generalized sliced Wasserstein distances. *Advances in neural information processing systems*, 32.

[178] Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. (2018). Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*.

[179] Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, Toronto, Canada.

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

[181] Kübler, J. M., Jitkrittum, W., Schölkopf, B., and Muandet, K. (2022a). A witness two-sample test. In *International Conference on Artificial Intelligence and Statistics*, pages 1403–1419.

[182] Kübler, J. M., Stimper, V., Buchholz, S., Muandet, K., and Schölkopf, B. (2022b). Automl two-sample test. *Advances in Neural Information Processing Systems*, 35:15929–15941.

[183] Kuhn, D., Shafieezadeh-Abadeh, S., and Wiesemann, W. (2024). Distributionally robust optimization. *arXiv preprint arXiv:2411.02549*.

[184] Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

[185] Kurakin, A., Goodfellow, I., and Bengio, S. (2017). Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

[186] Lam, H. (2016). Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275.

[187] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998a). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[188] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998b). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[189] Ledoux, M. (1999). Concentration of measure and logarithmic sobolev inequalities. *Séminaire de probabilités de Strasbourg*, 33:120–216.

[190] Lei, J. (2020). Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1).

[191] 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.

[192] Li, J., Lin, S., Blanchet, J., and Nguyen, V. A. (2022). Tikhonov regularization is optimal transport robust under martingale constraints. *Advances in Neural Information Processing Systems*, 35:17677–17689.

[193] Li, Y., Fampa, M., Lee, J., Qiu, F., Xie, W., and Yao, R. (2024). D-optimal data fusion: Exact and approximation algorithms. *INFORMS Journal on Computing*, 36(1):97–120.

[194] Li, Y. and Xie, W. (2020). Exact and approximation algorithms for sparse PCA. *arXiv preprint arXiv:2008.12438*.

[195] Li, Y. and Xie, W. (2022). On the exactness of dantzig-wolfe relaxation for rank constrained optimization problems. *arXiv preprint arXiv:2210.16191*.

[196] Li, Y. and Xie, W. (2023). Beyond symmetry: Best submatrix selection for the sparse truncated SVD. *Mathematical Programming*, pages 1–50.

[197] Li, Y. and Xie, W. (2024a). Best principal submatrix selection for the maximum entropy sampling problem: scalable algorithms and performance guarantees. *Operations Research*, 72(2):493–513.

[198] Li, Y. and Xie, W. (2024b). On the partial convexification for low-rank spectral optimization: Rank bounds and algorithms. *arXiv preprint arXiv:2305.07638, Forthcoming at Integer Programming and Combinatorial Optimization*.

[199] 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.

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

[201] 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.

[202] 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 *International Conference on Machine Learning*, pages 6316–6326.

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

[204] 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.

[205] Lofberg, J. (2004). Yalmip: A toolbox for modeling and optimization in matlab. In *2004 IEEE international conference on robotics and automation*, pages 284–289.

[206] Lopez-Paz, D. and Oquab, M. (2018). Revisiting classifier two-sample tests. In *International Conference on Learning Representations*.

[207] Love, D. and Bayraksan, G. (2015). Phi-divergence constrained ambiguous stochastic programs for data-driven optimization. *Technical report, Department of Integrated Systems Engineering, The Ohio State University, Columbus, Ohio*.

[208] 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*.

[209] 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.

[210] Luo, Z.-Q., Sidiropoulos, N. D., Tseng, P., and Zhang, S. (2007). Approximation bounds for quadratic optimization with homogeneous quadratic constraints. *SIAM Journal on optimization*, 18(1):1–28.

[211] Ma, S. (2013). Alternating direction method of multipliers for sparse principal component analysis. *Journal of the Operations Research Society of China*, 1:253–274.

[212] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

[213] Magdon-Ismail, M. (2017). Np-hardness and inapproximability of sparse PCA. *Information Processing Letters*, 126:35–38.

[214] Magesh, A., Sun, Z., Veeravalli, V. V., and Zou, S. (2023). Robust hypothesis testing with moment constrained uncertainty sets. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[215] Mairal, J. and Vert, J.-P. (2018). Machine learning with kernels. mines paristech, paris, france.

[216] Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

[217] McDiarmid, C. (1989). *On the method of bounded differences*, pages 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press.

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

[219] Micchelli, C. A. and Pontil, M. A. (2005). On learning vector-valued functions. *Neural Computation*, 17(1):177–204.

[220] Micchelli, C. A., Xu, Y., and Zhang, H. (2006). Universal kernels. *Journal of Machine Learning Research*, 7(12).

[221] Mika, S., Schölkopf, B., Smola, A., Müller, K.-R., Scholz, M., and Rätsch, G. (1998). Kernel pca and de-noising in feature spaces. *Advances in neural information processing systems*, 11.

[222] Miller, A. (2002). *Subset selection in regression*. chapman and hall/CRC.

[223] 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.

[224] Mitsuzawa, K., Kanagawa, M., Bortoli, S., Grossi, M., and Papotti, P. (2023). Variable selection in maximum mean discrepancy for interpretable distribution comparison. *arXiv preprint arXiv:2311.01537*.

[225] Moghaddam, B., Weiss, Y., and Avidan, S. (2005). Spectral bounds for sparse PCA: Exact and greedy algorithms. *Advances in neural information processing systems*, 18.

[226] 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.

[227] Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141.

[228] Mueller, J. and Jaakkola, T. (2015). Principal differences analysis: Interpretable characterization of differences between distributions. In *Advances in Neural Information Processing Systems*, volume 28.

[229] Mustafa, W., Lei, Y., and Kloft, M. (2022). On the generalization analysis of adversarial learning. In *International Conference on Machine Learning*, pages 16174–16196. PMLR.

[230] Nadjahi, K., Durmus, A., Simsekli, U., and Badeau, R. (2019). Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *Advances in Neural Information Processing Systems*, 32.

[231] 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.

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

[233] 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.

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

[235] Nguyen, K. and Ho, N. (2023). Energy-based sliced Wasserstein distance. *Advances in Neural Information Processing Systems*, 36.

[236] 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.

[237] 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*.

[238] Nietert, S., Goldfeld, Z., Sadhu, R., and Kato, K. (2022). Statistical, robustness, and computational guarantees for sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 35:28179–28193.

[239] Niles-Weed, J. and Rigollet, P. (2022). Estimation of wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688.

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

[241] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A.

(2016a). The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE.

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

[243] Pataki, G. (1998). On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research*, 23(2):339–358.

[244] 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.

[245] Paty, F.-P. and Cuturi, M. (2019). Subspace robust wasserstein distances.

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

[247] 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.

[248] 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.

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

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

[251] Pólik, I. and Terlaky, T. (2007). A survey of the S-lemma. *SIAM review*, 49(3):371–418.

[252] Poor, H. and Hadjiliadis, O. (2008). *Quickest detection*. Cambridge University Press.

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

[254] Pratt, J. W. and Gibbons, J. D. (1981). *Kolmogorov-Smirnov Two-Sample Tests*. Springer New York.

[255] 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*.

[256] Ramdas, A., García Trillos, N., and Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47.

[257] 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*, pages 3571–3577.

[258] Rockafellar, R. T. (1970). *Convex analysis*. Princeton Mathematical Series. Princeton University Press.

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

[260] Rozsa, A., Gunther, M., and Boult, T. E. (2018). Towards robust deep neural networks with bang. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 803–811.

[261] Rubinstein, R. Y. and Marcus, R. (1985). Efficiency of multivariate control variates in monte carlo simulation. *Operations Research*, 33(3):661–677.

[262] Sadana, U., Chenreddy, A., Delage, E., Forel, A., Frejinger, E., and Vidal, T. (2023). A survey of contextual optimization methods for decision making under uncertainty. *arXiv preprint arXiv:2306.10374*.

[263] Samadi, S., Tantipongpipat, U., Morgenstern, J. H., Singh, M., and Vempala, S. (2018). The price of fair pca: One extra dimension. *Advances in neural information processing systems*, 31.

[264] Savage, D., Zhang, X., Yu, X., Chou, P., and Wang, Q. (2014). Anomaly detection in online social networks. *Social networks*, 39:62–70.

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

[266] Schober, P. and Vetter, T. (2019). Two-sample unpaired t tests in medical research. *Anesthesia and analgesia*, 129:911.

[267] 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.

[268] Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319.

[269] Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

[270] Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2021). Mmd aggregated two-sample test. *arXiv preprint arXiv:2110.15073*.

[271] Schrab, A., Kim, I., Guedj, B., and Gretton, A. (2022). Efficient aggregated kernel tests using incomplete u-statistics. *Advances in Neural Information Processing Systems*, 35:18793–18807.

[272] 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*.

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

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

[275] Shafieezadeh Abadeh, S., Mohajerin Esfahani, P. M., and Kuhn, D. (2015). Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, volume 28.

[276] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

[277] Shalev-Shwartz, S. and Singer, Y. (2006). Efficient learning of label ranking by soft projections onto polyhedra. *Journal of Machine Learning Research*.

[278] Shapiro, A. (2001). On duality theory of conic linear problems. In *Semi-infinite programming*, pages 135–165. Springer.

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

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

[281] Shapiro, A., Zhou, E., and Lin, Y. (2023a). Bayesian distributionally robust optimization. In [282], pages 1279–1304.

[282] Shapiro, A., Zhou, E., and Lin, Y. (2023b). Bayesian distributionally robust optimization. *SIAM Journal on Optimization*, 33(2):1279–1304.

[283] Sidiropoulos, N. D., Davidson, T. N., and Luo, Z.-Q. (2006). Transmit beamforming for physical-layer multicasting. *IEEE transactions on signal processing*, 54(6):2239–2251.

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

[285] 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.

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

[287] Sinha, A., Namkoong, H., Volpi, R., and Duchi, J. (2020). Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*.

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

[289] Song, J., He, N., Ding, L., and Zhao, C. (2023). Provably convergent policy optimization via metric-aware trust region methods. *Transactions on Machine Learning Research*.

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

[291] Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and

Gretton, A. (2016). Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*.

[292] Szepesvári, C. and Littman, M. L. (1999). A unified analysis of value-function-based reinforcement-learning algorithms. *Neural computation*, 11(8):2017–2060.

[293] Taguchi, G. and Rajesh, J. (2000). New trends in multivariate diagnosis. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 233–248.

[294] TinyImageNet (2014). TinyImageNet Visual Recognition Challenge.

[295] Todd, M. J. (2001). Semidefinite optimization. *Acta Numerica*, 10:515–560.

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

[297] Van Der Maaten, L. (2014). Accelerating t-sne using tree-based algorithms. *The journal of machine learning research*, 15(1):3221–3245.

[298] Van Parys, B. P., Goulart, P. J., and Kuhn, D. (2015). Generalized gauss inequalities via semidefinite programming. *Mathematical Programming*, 156(1-2):271–302.

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

[300] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

[301] 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.

[302] Wang, J., Boedihardjo, M., and Xie, Y. (2024a). Statistical and computational guarantees of kernel max-sliced wasserstein distances. *arXiv preprint arXiv:2405.15441*.

[303] Wang, J., Chen, M., Zhao, T., Liao, W., and Xie, Y. (2023a). A manifold two-sample test study: integral probability metric with neural networks. *Information and Inference: A Journal of the IMA*, 12(3):1867–1897.

[304] Wang, J., Dey, S. S., and Xie, Y. (2023b). Variable selection for kernel two-sample tests. *arXiv preprint arXiv:2302.07415*.

[305] Wang, J., Gao, R., and Xie, Y. (2021a). Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*.

[306] Wang, J., Gao, R., and Xie, Y. (2021b). Two-sample test using projected Wasserstein distance. In *2021 IEEE International Symposium on Information Theory*, pages 3320–3325.

[307] 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.

[308] Wang, J., Gao, R., and Xie, Y. (2022b). Two-sample test with kernel projected Wasserstein distance. In [307].

[309] Wang, J., Gao, R., and Xie, Y. (2024b). Non-convex robust hypothesis testing using Sinkhorn uncertainty sets. *arXiv preprint arXiv:2403.14822*.

[310] Wang, J., Gao, R., and Xie, Y. (2024c). Regularization for adversarial robust learning. *arXiv preprint arXiv:2408.09672*.

[311] Wang, J., Gao, R., and Zha, H. (2024d). Reliable off-policy evaluation for reinforcement learning. In [312], pages 699–716.

[312] Wang, J., Gao, R., and Zha, H. (2024e). Reliable off-policy evaluation for reinforcement learning. *Operations Research*, 72(2):699–716.

[313] Wang, J., Moore, R., Xie, Y., and Kamaleswaran, R. (2022c). Improving sepsis prediction model generalization with optimal transport. In *Machine Learning for Health*, pages 474–488. PMLR.

[314] Wang, J. and Xie, Y. (2022). A data-driven approach to robust hypothesis testing using sinkhorn uncertainty sets. *arXiv preprint arXiv:2202.04258*.

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

[316] Wen, Z. and Yin, W. (2012). A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434.

[317] Wendland, H. (2004). *Scattered data approximation*, volume 17. Cambridge university press.

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

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

[320] Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

[321] Xiao, J., Fan, Y., Sun, R., and Luo, Z.-Q. (2022). Adversarial rademacher complexity of deep neural networks. *arXiv preprint arXiv:2211.14966*.

[322] Xie, L. and Xie, Y. (2021). Sequential change detection by optimal weighted ℓ_2 divergence. *IEEE Journal on Selected Areas in Information Theory*, 2(2):747–761.

[323] 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).

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

[325] Xie, Y., Luo, Y., and Huo, X. (2022). An accelerated stochastic algorithm for solving the optimal transport problem. *arXiv preprint arXiv:2203.00813*.

[326] Xu, C., Lee, J., Cheng, X., and Xie, Y. (2024). Flow-based distributionally robust optimization. *IEEE Journal on Selected Areas in Information Theory*.

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

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

[329] Yang, J., Zhang, L., Chen, N., Gao, R., and Hu, M. (2022). Decision-making with side information: A causal transport robust approach. *Optimization Online*.

[330] Yin, D., Kannan, R., and Bartlett, P. (2019). Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pages 7085–7094. PMLR.

[331] 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.

[332] 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.

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

[334] Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142.

[335] Zhang, Q., Zhou, Y., Prater-Bennette, A., Shen, L., and Zou, S. (2024). Large-scale non-convex stochastic constrained distributionally robust optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8217–8225.

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

[337] 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.

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

VITA

Jie Wang was born in Jieshou, Anhui Province, China in January 1999. He received the B.Sc. in Pure Mathematics in 2020 at The Chinese University of Hong Kong, Shenzhen. Afterwards, his interest was drawn towards industrial engineering and operations research, and he joined the Ph.D. program in the School of Industrial and Systems Engineering at Georgia Institute of Technology, under supervision of Prof. Yao Xie. With a fulfilling period and many stimulating experiences, he has completed his Ph.D. studies and now he is ready for new adventures. These will take him to The Chinese University of Hong Kong, Shenzhen, where he will be an assistant professor in the School of Artificial Intelligence and the School of Data Science.