Quantifying the Empirical Wasserstein Distance to a Set of Measures: Beating the Curse of Dimensionality

Nian Si
Department of Management Science and Engineering
Stanford University
Huang Engineering Center, 475 Via Ortega, Stanford, California 94305, United States
niansi@stanford.edu

Jose Blanchet
Department of Management Science and Engineering
Stanford University
Huang Engineering Center, 475 Via Ortega, Stanford, California 94305, United States
jose.blanchet@stanford.edu

Soumyadip Ghosh
Mathematical Sciences
IBM Research
Yorktown Heights, NY 10598, USA
ghoshs@us.ibm.com

Mark S. Squillante
Mathematical Sciences
IBM Research
Yorktown Heights, NY 10598, USA
mss@us.ibm.com

Abstract

We consider the problem of estimating the Wasserstein distance between the empirical measure and a set of probability measures whose expectations over a class of functions (hypothesis class) are constrained. If this class is sufficiently rich to characterize a particular distribution (e.g., all Lipschitz functions), then our formulation recovers the Wasserstein distance to such a distribution. We establish a strong duality result that generalizes the celebrated Kantorovich-Rubinstein duality. We also show that our formulation can be used to beat the curse of dimensionality, which is well known to affect the rates of statistical convergence of the empirical Wasserstein distance. In particular, examples of infinite-dimensional hypothesis classes are presented, informed by a complex correlation structure, for which it is shown that the empirical Wasserstein distance to such classes converges to zero at the standard parametric rate. Our formulation provides insights that help clarify why, despite the curse of dimensionality, the Wasserstein distance enjoys favorable empirical performance across a wide range of statistical applications.

1 Introduction

In this paper we consider the problem of projecting the empirical measure, under the Wasserstein distance, to a set of probability measures that are constrained to satisfy a family of expectations over a class of functions. We call this class of functions the "hypothesis class", examples of which include moment constraints or expectations of functions other than polynomials.

The Wasserstein distance has generated a great deal of attention in recent years across a broad spectrum of areas, ranging from artificial intelligence, learning and statistics to areas such as image analysis, economics and operations research [1, 18, 9, 12, 15]. However, despite its versatility and

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

modelling power, classical results on the rates of statistical convergence of the Wasserstein distance metric show that these rates scale poorly as a function of the dimension of the space [8]. This may suggest that comparing distributions based on the Wasserstein distance is a strategy that is bound to suffer from the so-called curse of dimensionality. Nevertheless, such theoretical performance in terms of rates of statistical convergence seems to be incompatible with the popularity of the Wasserstein distance based on the empirical performance observed in the previously mentioned application areas.

Our goal in this paper is to shed light on some of the fundamental reasons that explain the empirical performance of the Wasserstein distance as an effective way to compare distributions, guided by the following intuition. The Wasserstein distance (using, say, the Euclidean metric in $\mathbb{R}^d$) has substantial power to "separate" two distributions based on a wide and detailed range of characteristics. Meanwhile, some users of Wasserstein distances may be interested in only a subset of these characteristics (maybe a large subset, but just a subset, nonetheless). Hence, in the end, these users may be interested in only testing if an empirical sample is compatible with a subset of characteristics. Since this subset of characteristics of interest are likely to change from user to user or from task to task, the power of the Wasserstein distance to discriminate widely makes is particularly convenient for multiple users or tasks with different preferences because of this type of versatility. In practice, however, when testing if the data is compatible with the characteristics required for a particular user or task, such a user typically exploits the Wasserstein distance to obtain key insights and a deeper understanding while, in the end, making final decisions with a criterion that may ignore a lack of fit of certain aspects.

To be more precise, consider as a canonical example the process of using the Wasserstein distance in the Wasserstein GAN application [1]. The general goal is to fine tune a neural network to generate synthetic data that is similar in some sense to a target data set. The network is trained in order to minimize the Wasserstein distance. However, if the generative models eventually produce the desirable features (e.g., faces that appear to be realistic), we may choose to ignore imperfections in, for example, the background of the picture. Hence, "faces" are what we choose to emphasize in the training process and the rest of the data characteristics are not given as much importance.

The idea of choosing a hypothesis class corresponds precisely to modeling the set of characteristics that are important. The hypothesis class partitions the set of distributions into equivalence classes, where two distributions are equivalent if the expectations coincide over the hypothesis class. Formally, we posit that many users of the Wasserstein distance are actually testing if the data belongs to a certain equivalence class. To provide a solid statistical footing for such scenarios, this then involves computing the distance between the empirical measure and the target equivalence class, evaluating a corresponding asymptotic quantile statistic, and rejecting the hypothesis of membership in the target equivalence class for large values of the statistic relative to the desired confidence quantile.

More formally, suppose that P_n denotes the empirical measure of independent and identically distributed (i.i.d.) samples $\{X_i\}_{i=1}^n \subseteq \mathbb{R}^d$ generated from a distribution P_*. Let us write $\mathcal{W}(P, P_n)$ to denote the Wasserstein distance [24] between P_n and a given (Borel) probability measure P. (We recall the formal definition of the Wasserstein distance in Section 2.1.)

Let $\mathcal{B}$ be a given hypothesis class of interest. To avoid technicalities, let us focus in this introductory discussion on a given subset of the space of continuous and bounded functions with certain characteristics. Next, for $f \in \mathcal{B}$, we write $\mathbb{E}_P[f(X)]$ to denote the expectation of $f(X)$ under the measure P; so, for example, $\mathbb{E}_{P_n}[f(X)] = n^{-1}\sum_{i=1}^n f(X_i)$.

Our goal then in this paper is to study

$$R_n = \inf_P \{\mathcal{W}(P, P_n) : \mathbb{E}_P[f(X)] = \mathbb{E}_{P_*}[f(X)] \text{ for all } f \in \mathcal{B}\}. \tag{1}$$

The main contributions of this paper are as follows. First, we provide a duality result that shows

$$R_n = \sup_{f \in \mathcal{LB}} \{\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_{P_n}[f^c(X)]\},$$

where f^c is a suitable transformation (to be described precisely in Theorem 1) and $\mathcal{LB}$ is the linear span generated by $\mathcal{B}$. If $\mathcal{B}$ is the class of all 1-Lipschitz functions and the cost function is the corresponding metric, then it turns out for $f \in \mathcal{B}$ that $f^c = f$ and we recover the celebrated Kanotorovich-Rubinstein duality.

The second contribution of this paper is to study the rate of statistical convergence for R_n. Note that, if $\mathcal{B}$ is the class of 1-Lipschitz functions, then R_n will typically converge to zero at the rate

$O_p\left(n^{-1/(d\vee 2)}\right)$ [8], where d is the dimension of the underlying space. If $\mathcal{B}$ is finite dimensional, then convergence of R_n occurs at a parametric rate [3]. However, we also establish more general conditions that accommodate infinite dimensional hypothesis classes $\mathcal{B}$ and for which a parametric rate of convergence is also achievable, thus beating the curse of dimensionality. Examples of infinite-dimensional hypothesis classes, informed by a complex correlation structure, are considered.

Moreover, we are able to explicitly characterize the asymptotically limiting distribution of $n \times R_n$ as $n \to \infty$, which is the maximum of a Gaussian process, indexed by functions in $\mathcal{LB}$. The existence of this asymptotic distribution is critically important from the standpoint of using these results for hypothesis testing, as we have discussed above, either by explicit evaluation of quantiles or by means of subsampling as considered in [17].

There is a rapidly growing research literature discussing the statistical properties of the Wasserstein distance and how to beat the curse of dimensionality. Weed and Bach [25] claim that the Wasserstein distance enjoys a faster convergence rate if the true measure has support on a lower-dimensional manifold. Weed and Berthet [26] produce a new density estimator that converges faster if the true measure has sufficiently smooth density. Tameling et al. [20] recover the parametric rates of convergence, but under the assumption that the underlying measures are atomic. Genevay et al. [11] study Wasserstein distance with entropy regularization (Sinkhorn Divergences), but their convergence rate is exponential in the regularization power ε. In connection to our study, Blanchet et al. [2, 3] focus on finite hypothesis classes and prove that the canonical rate of statistical convergence can be obtained. We study cases in which the hypothesis class may form an infinite-dimensional vector space encoding complex information about the joint distribution, for which we are able to show, for the first time, that it is not only possible to also obtain a canonical rate of statistical convergence in these types of complex formulations, but to further obtain a characterization of the limiting distribution.

Our formulation is also related to distributionally robust optimization (DRO) with the Wasserstein distance metric [4, 16, 27, 3, 10, 5]. In this literature, estimators are obtained as the solution of a $\min$-$\max$ game in which the optimizer seeks to minimize a loss, while an adversary chooses a probability distribution inside a so-called "uncertainty set" defined around the empirical measure. The Wasserstein distance is used to describe the uncertainty set and R_n is used to describe the radius of the uncertainty set (also called the size of uncertainty). One criterion for choosing the size of uncertainty is to minimize the size of a natural confidence region for the parameter of interest; refer to [3]. Under this criterion, it is shown that the optimal size of uncertainty coincides with a quantile of R_n (which, in this literature, is known as the "Robust Wasserstein Profile" function).

The paper is organized as follows. In Section 2, we provide the necessary definitions and setup to state our duality result in compact spaces, which is presented in Section 2.2. Then, in Section 3, we discuss the statistical guarantee that R_n satisfies, where we present a central limit theorem for R_n. Further, in Section 4, we extend our duality result and our statistical guarantee to non-compact spaces. Finally, Section 5 illustrates the use of our results in the context of a hypothesis testing example.

Notation. Let $\mathcal{C}^k(\Omega)$ represent the space of all k-th continuous differentiable functions defined on the domain Ω, where $\mathcal{C}(\Omega)$ denotes the space of continuous functions and $\mathcal{C}_b(\Omega)$ the space of bounded continuous functions. Denote by $\mathcal{P}(\Omega)$ the space of all Borel probability measures on the underlying space Ω. Let $L_1(\mu)$ be the space of all integrable functions with respect to measure μ. Denote by $\mathbb{Z}_+$ the set of all positive integers and by $\|\cdot\|_F$ the Frobenius norm of a matrix. Let $\Rightarrow$ denote the weak convergence in a given probability space, and $\mathcal{N}(\mu, \sigma^2)$ a Gaussian distribution with mean μ and variance σ^2. For a vector $x \in \mathbb{R}^d$, we use $x^{(i)}, i = 1, 2, \ldots, d$, to denote the i-th entry of x.

2 Main Duality Result

The goal of this section is to present our new strong duality result, also providing the necessary definitions to do so. Recall that this result extends the existing optimal transport duality theory in a geometric sense by closing the gap between the renowned Kantorovich-Rubinstein duality result [24] at one extreme and the recent strong duality result in [3] at the other extreme. In doing so, our new strong duality result helps to reduce the computational burdens encountered in practice by establishing an equivalence with a problem that is easier and more computationally efficient to solve.

We start by reviewing the definition of the Wasserstein distance and the elements required to pose the dual problem. We then state our new strong duality result together with some examples of applying

the result, which further illustrate some of the benefits of our duality result. This is then followed by an extension of our strong duality result beyond the Wasserstein distance in (1) to the max-sliced Wasserstein distance in [7].

2.1 Wasserstein Distance

For a given closed set $\Omega \subseteq \mathbb{R}^d$, we endow Ω with a metric, denoted by $\varrho(\cdot)$, which may be naturally defined in terms of a norm such as $\varrho(x,y) = \|y - x\|$. Let $c : \Omega \times \Omega \to [0,\infty)$ be a continuous function with respect to $\varrho(\cdot)$. Then the optimal transport cost between $P, Q \in \mathcal{P}(\Omega)$ is defined as

$$\begin{aligned}\mathcal{D}_c(P,Q) \quad = \quad & \min_{\pi \in \mathcal{P}(\Omega\times\Omega)} \left\{ \left(\int c(x,w)\pi(\mathrm{d}x,\mathrm{d}w) \right) \right. \\ & \left. : \int_{w\in\mathbb{R}^d} \pi(\mathrm{d}x,\mathrm{d}w) = P(\mathrm{d}x), \int_{x\in\mathbb{R}^d} \pi(\mathrm{d}x,\mathrm{d}w) = Q(\mathrm{d}w) \} \right\}.\end{aligned}$$

If $c(\cdot) = \varrho(\cdot)$, then $\mathcal{W}_1(P,Q) = \mathcal{D}_\varrho(P,Q)$ is the Wasserstein distance generated by such a metric [24]. However, we may also be interested in cases where $c(\cdot) = \varrho^r(\cdot)$ for $r > 1$ in order to study the Wasserstein distance of order r, which is defined as $\mathcal{W}_r(P,Q) = \mathcal{D}_{\varrho^r}^{1/r}(P,Q)$.

2.2 Strong Duality

The hypothesis class $\mathcal{B}(\Omega)$ is assumed to be given throughout our discussion which follows where we further assume that $\mathcal{B}(\Omega) \subseteq \mathcal{C}(\Omega) \cap L_1(P_*)$ for a targeting probability measure P_*. We may also assume, without loss of generality, that $\mathbf{1} \in \mathcal{B}(\Omega)$ (i.e., constant functions belong to the hypothesis class). Let $\mathcal{LB}(\Omega)$ denote the linear span generated by $\mathcal{B}(\Omega)$, namely

$$\mathcal{LB}(\Omega) = \left\{ f(\cdot) = \sum_{i=1}^m \lambda_i f_i(\cdot) : \{f_i(\cdot)\}_{i=1}^m \subset \mathcal{B}(\Omega), \lambda \in \mathbb{R}^m, \text{ and } m \in \mathbb{Z}_+ \right\}.$$

We formally state our assumptions as follows.

Assumption 1. *1. The function class satisfies $\mathcal{B}(\Omega) \subseteq \mathcal{C}(\Omega) \cap L_1(P_*)$.*

2. The cost function $c(\cdot,\cdot)$ is a non-negative continuous function with $c(x,x) = 0$, for $x \in \Omega$.

Given a probability measure $P_0 \in \mathcal{P}(\Omega)$ (which eventually will be taken as an empirical measure), we are interested in studying the robust Wasserstein profile function

$$R_0 = \inf_{P\in\mathcal{P}(\Omega)} \{\mathcal{D}_c(P,P_0) : \mathbb{E}_P[f(X)] = \mathbb{E}_{P_*}[f(X)], \text{ for all } f \in \mathcal{B}(\Omega)\}. \tag{2}$$

Observe that writing $\mathcal{B}(\Omega)$ or $\mathcal{LB}(\Omega)$ in the definition of R_0 leads to an equivalent formulation due to the linearity of the constraints defining R_0. We now state our main duality result.

Theorem 1. *Suppose Assumption 1 is enforced and $\mathcal{B}(\Omega) \subset L_1(P_0)$. We then have the weak duality*

$$R_0 \geq \sup_{f\in\mathcal{LB}(\Omega)} \{\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_{P_0}[f^c(X)]\},$$

where f^c is the c-transform of f, which is defined by

$$f^c(x) = \sup_{z\in\Omega} \{f(z) - c(z,x)\}.$$

Furthermore, if Ω is compact, we have the strong duality

$$R_0 = \sup_{f\in\mathcal{LB}(\Omega)} \{\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_{P_0}[f^c(X)]\}. \tag{3}$$

The key to the proof is first writing R_0 in a Lagrangian form and then applying Sion's minimax theorem [19]. The technical details and complete proof are provided in Appendix A.1.

Remark 1. *Notice that, for the strong duality, we require the sample space to be compact. For the non-compact space, the strong duality does not hold in general and should be treated on a case-by-case basis. We will discuss such strong duality results for some examples in Section 4.*

Remark 2. *Note that the dual formulation* (3) *shares some similarities with the Integral Probability Metric (IPM),which is defined as*

$$\mathrm{IPM}_{\mathcal{F}}(P, P_0) = \sup_{f \in \mathcal{F}} \left| \int f \mathrm{d}P - \int f \mathrm{d}P_0 \right|,$$

for a function class $\mathcal{F}$. The similarities are not surprising since the dual formulations of Wasserstein distances have deep connections with IPM. However, it is important to note that our primary intention is not to define a new metric. Rather we seek to provide a thorough analysis of the Wasserstein distance, which has been the focus of a great deal of attention in the statistical learning research literature. In particular, we add a new modeling feature, which is the hypothesis class or the actor critic class. This induces a class of dual functions; and we note that our expression for the strong duality (generalizing the celebrated Kantorovich-Rubinstein duality) uses the combination of both the function f and its c-transform f^c in contrast with IPM.

Problem (2) is an infinite-dimensional optimization problem that cannot be solved directly. Our main duality results (Theorem 1) enable us to compute R_0 using function approximators for functions in $\mathcal{LB}(\Omega)$, such as wavelet basis expansions. We will discuss computing R_0 in Section 5.

For now, let us consider a few examples that apply our results to illustrate some of the benefits which they provide. In order to connect these examples with our future statistical development, recall that $\{X_i\}_{i=1}^n \subseteq \Omega$ are i.i.d. samples from a data-generating distribution $P_* \in \mathcal{P}(\Omega)$ and that $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the corresponding empirical measure. We next apply our strong duality result where P_0 is replaced by P_n and the corresponding R_n is defined as

$$R_n = \inf_{P \in \mathcal{P}(\Omega)} \{\mathcal{D}_c(P, P_n) : \mathbb{E}_P[f(X)] = \mathbb{E}_{P_*}[f(X)], \text{ for all } f \in \mathcal{B}(\Omega)\}.$$

Example 1. *Suppose $\mathcal{LB}(\Omega)$ is sufficiently rich to uniquely determine any distributions and assume that $c = \varrho$. Then, we might assume that $\mathcal{LB}(\Omega)$ is the space of all Lipschitz functions, which also determines any distribution. Let $\mathrm{Lip}_1(\Omega)$ be the space of all 1-Lipschitz functions. Hence, by our weak duality result, we have*

$$\sup_{f \in \mathrm{Lip}_1(\Omega)} \{\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_{P_n}[f^c(X)]\} \leq R_n.$$

On the other hand, since $f(x) \leq \sup_{z \in \Omega}\{f(z) - c(z, x)\}$, we also have

$$R_n \leq \sup_{f \in \mathcal{LB}(\Omega)} \{\mathbb{E}_{P_*}[f^c(X)] - \mathbb{E}_{P_n}[f^c(X)]\}.$$

Finally, it is well known (see, e.g., [24]) that $f^c(x)$ is a 1-Lipschitz function, and therefore

$$\sup_{f \in \mathcal{LB}(\Omega)} \{\mathbb{E}_{P_*}[f^c(X)] - \mathbb{E}_{P_n}[f^c(X)]\} \leq \sup_{f^c \in \mathrm{Lip}_1(\Omega)} \{\mathbb{E}_{P_*}[f^c(X)] - \mathbb{E}_{P_n}[f^c(X)]\}.$$

Consequently, if $\mathcal{LB}(\Omega)$ determines any distribution, then our result recovers the renowned Kantorovich-Rubinstein duality result [24, Theorem 5.10]:

$$R_n = \sup_{f \in \mathrm{Lip}_1(\Omega)} \{\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_{P_n}[f(X)]\} = \mathcal{W}_1(P_*, P_n).$$

It is important to keep in mind that, if P_* has bounded moments, then $R_n = O\left(n^{-1/(d \vee 2)}\right)$ as $n \to \infty$ (see, e.g., [8, Theorem 1]).

Example 2. *Suppose that $\mathcal{B}(\Omega)$ is finite dimensional, such as $\mathcal{B}(\Omega) = \{f_i(x)\}_{i=1}^K$. Then, we have*

$$R_n = \sup_{\lambda \in \mathbb{R}^K} \left\{ \mathbb{E}_{P_*}\left[\sum_{i=1}^K \lambda_i f_i(X)\right] - \mathbb{E}_{P_n}\left[\sup_{z \in \Omega}\left\{\sum_{i=1}^K \lambda_i f_i(z) - c(z, X)\right\}\right] \right\},$$

which recovers the duality result obtained in [3]. Note that [3] also provides a typical rate $R_n = O_p\left(n^{-1}\right)$ as $n \to \infty$ under some regularity conditions.

Example 3. *Fix linearly independent unit vectors $\theta_1, \ldots, \theta_K \in \mathbb{R}^d$, $K \leq d$, and let a function class $\mathcal{F}_{\mathcal{B}} \subseteq \mathcal{C}_b(\mathbb{R})$ collect some bounded continuous functions in $\mathbb{R}$. We consider the function class $\mathcal{B}(\Omega) = \cup_{i=1}^K \mathcal{B}_i(\Omega)$, where $\mathcal{B}_i(\Omega) = \left\{f(\theta_i^\top \cdot)|_\Omega : f \in \mathcal{F}_{\mathcal{B}}\right\}$, in which case*

$$\mathcal{LB}(\Omega) = \left\{ f(\cdot) = \sum_{i=1}^K \lambda_i f_i(\theta_i^\top \cdot)|_\Omega : \{f_i(\cdot)\}_{i=1}^K \subset \mathcal{F}_{\mathcal{B}}, \lambda \in \mathbb{R}^K \right\}.$$

This example is particularly interesting because it is infinite dimensional if $\mathcal{F}_{\mathcal{B}}$ is infinite dimensional. The hypothesis class carries a substantial amount of information about the dependence structure of P_* and yet, as we shall see, for this hypothesis class and the cost function $c(x,y) = \|x-y\|_2^2$, we also conclude that $R_n = O_p\left(n^{-1/2}\right)$ for $\Omega = \mathbb{R}^d$ (Theorem 4 below) and $R_n = O_p\left(n^{-1}\right)$ under suitable regularity (Theorem 2 below).

At first glance, Example 3 is similar to the max-sliced Wasserstein distance [7]. Recall that the max-sliced Wasserstein distance is defined as

$$\max\text{-}\mathcal{W}_r\left(P,Q\right) = \left[\max_{\theta:\|\theta\|_2=1} \mathcal{W}_r\left(\theta_\sharp P, \theta_\sharp Q\right)^r\right]^{1/r},$$

where $\theta_\sharp P(\theta_\sharp Q)$ is the push-forward measure from $\mathcal{P}(\Omega)$ to $\mathcal{P}\left(\theta^\top\Omega\right)$ such that, for any Borel set A in $\theta^\top\Omega$,

$$\left(\theta_\sharp P\right)(A) = P\left(\left\{x\in\Omega : \theta^\top x \in A\right\}\right). \tag{4}$$

Proposition 1 provides a strong duality result for $\max\text{-}\mathcal{W}_r\left(P,Q\right)$.

Proposition 1. *Consider $\Omega = \mathbb{R}^d, r = 2$, and $\varrho(x,y) = |x-y|$, for $x,y\in\mathbb{R}$. Denote by S^{d-1} a unit sphere in $\mathbb{R}^d$, i.e., $S^{d-1} = \left\{x\in\mathbb{R}^d : \|x\|_2 = 1\right\}$. Then, for $\Theta\subset S^{d-1}$, we have the strong duality*

$$\max_{\theta\in\Theta} \mathcal{W}_2\left(\theta_\sharp P, \theta_\sharp Q\right)^2 = \sup_{f\in\mathcal{B}_{\max}(\mathbb{R}^d,\Theta)} \left\{\mathbb{E}_P\left[f\left(X\right)\right] - \mathbb{E}_Q\left[f^c(X)\right]\right\}, \tag{5}$$

where the cost function $c(x,y) = \|x-y\|_2^2$ and

$$\mathcal{B}_{\max}\left(\mathbb{R}^d,\Theta\right) = \left\{f\left(\theta^\top\cdot\right) : f\in\mathcal{C}_b(\mathbb{R}), \theta\in\Theta\right\}.$$

In particular, for the max-sliced distance, we have the strong duality

$$\left(\max\text{-}\mathcal{W}_2\left(P,Q\right)\right)^2 = \sup_{f\in\mathcal{B}_{\max}(\mathbb{R}^d,S^{d-1})} \left\{\mathbb{E}_P\left[f\left(X\right)\right] - \mathbb{E}_Q\left[f^c(X)\right]\right\}.$$

The proof of Proposition 1 is provided in Appendix A.2. The key difference between the dualities (1) and (5) is that $\mathcal{B}_{\max}\left(\mathbb{R}^d,\Theta\right)$ is not a vector space in general. Therefore, even if $\Theta = \{\theta_i\}_{i=1}^K$, $\mathcal{LB}(\mathbb{R}^d)$ could be much larger than $\mathcal{B}_{\max}\left(\mathbb{R}^d,\Theta\right)$.

3 Statistical Convergence

In this section, we present a formal central limit theorem result on the rate of statistical convergence for R_n in the case of infinite dimensional constraints, which also extends and improves the corresponding results in [3] for the finite dimensional case. This further extends conventional results on the rate of statistical convergence for Wasserstein distances between an empirical distribution and the true (unknown) distribution. Such central limit theorem results on the rate of statistical convergence for R_n provide a critically important understanding that can inform and guide algorithms, computation, and experiments.

Following the setting in Example 3, we consider a convex compact domain Ω and let $\mathcal{B}_i(\Omega)$ be any subclass of the function class $\left\{f(\theta_i^\top\cdot)|_\Omega : f\in\mathcal{C}^2(\mathbb{R})\right\}$. As an analog of $\mathcal{LB}(\Omega)$, we define $\mathcal{LB}_i(\Omega)$ to be

$$\mathcal{LB}_i(\Omega) = \left\{f(\cdot) = \sum_{j=1}^m \lambda_j f_j(\cdot) : \{f_j(\cdot)\}_{j=1}^m \subset \mathcal{B}_i(\Omega), \lambda\in\mathbb{R}^m, \text{ and } m\in\mathbb{Z}_+\right\}.$$

Notice that any function in $\mathcal{LB}_i(\Omega)$ can be written as $f(\theta_i^\top x)$. We assume that the function classes $\mathcal{LB}_i(\Omega)$ satisfy the following condition.

Assumption 2. *For any $f\left(\theta_i^\top\cdot\right)\in\mathcal{LB}_i(\Omega)$, $i = 1,2,\ldots,K$, the ratio bound*

$$\frac{\sup_{x\in\Omega}\left|f''(\theta_i^\top x)\right|}{\sqrt{\int_\Omega f'(\theta_i^\top x)^2\mathrm{d}x}} \le M$$

holds for a universal constant $M\in(0,+\infty)$, where we use the convention $0/0 = 0$.

As in Example 3, we consider the function space $\mathcal{B}(\Omega) = \cup_{i=1}^{K} \mathcal{B}_i(\Omega)$. We make further assumptions on the domain Ω, the data-generating probability measure P_*, and the linear projection vectors $\theta_1, \theta_2, \ldots, \theta_K$ in Assumption 3 as follows.

Assumption 3. *1. The sample space Ω is a convex and compact subset of $\mathbb{R}^d$.*

2. The data-generating probability measure P_* has a non-zero density f_{P_*} with respect to Lebesgue measure in $\mathbb{R}^d$. The density has a uniform non-zero lower bound, i.e., $f_{P_*}(x) \geq \underline{b} > 0$ for $x \in \Omega$.

3. The vectors $\theta_1, \ldots, \theta_K$ are linearly independent with $\|\theta_i\|_2 = 1$ for $i = 1, 2, \ldots, K$.

Theorem 2. *Suppose Assumptions 1, 2 and 3 are enforced. For the cost function $c(x, y) = \|x - y\|_2^2$, we then have the central limit theorem result*

$$nR_n \Rightarrow \sup_{f \in \mathcal{LB}(\Omega)} \left\{ -2H^f - \mathbb{E}_{P_*} \left[\|\nabla_X f(X)\|_2^2 \right] \right\},$$

where $\nabla_x f(x)$ is the gradient of $f(\cdot)$ evaluated at x and H^f is a Gaussian process indexed by f with

$$H^f \sim \mathcal{N}(0, \operatorname{var}(f(X))) \text{ and } \operatorname{cov}(H^{f_1}, H^{f_2}) = \operatorname{cov}(f_1(X), f_2(X)).$$

Sketch of Proof. Define $\mathcal{UB}(\Omega) = \left\{ f(\cdot) \in \mathcal{LB}(\Omega) : \mathbb{E}_{P_*} \left[\|\nabla_x f(X)\|_2^2 \right] = 1, f(0) = 0 \right\}$. By Theorem 1, we have nR_n is equal to

$$\sup_{\lambda \in \mathbb{R}} \sup_{f \in \mathcal{UB}(\Omega)} \left\{ -2\lambda H_n^f - \frac{1}{n} \sum_{i=1}^{n} \left[\sup_{X_i + \Delta/\sqrt{n} \in \Omega} \left\{ 2\lambda \sqrt{n} \left(f\left(X_i + \Delta/\sqrt{n}\right) - f(X_i) \right) - \|\Delta\|_2^2 \right\} \right] \right\},$$

where $H_n^f = n^{-1/2} \left(\sum_{i=1}^{n} f(X_i) - E_P[f(X)] \right)$. Then, by the uniform convergence theory of the P-Donsker class and the P-Glivenko-Cantelli class (see [21, Chapter 19]), we obtain for any $b > 0$

$$\sup_{|\lambda| \leq b} \sup_{f \in \mathcal{UB}(\Omega)} \left\{ \lambda H_n^f \right\} \Rightarrow \sup_{|\lambda| \leq b} \sup_{f \in \mathcal{UB}(\Omega)} \left\{ \lambda H^f \right\}, \text{ and} \tag{6}$$

$$\sup_{|\lambda| \leq b} \sup_{f \in \mathcal{UB}(\Omega)} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\sup_{X_i + \Delta/\sqrt{n} \in \Omega} \left\{ 2\lambda \sqrt{n} \left(f(X_i + \Delta/\sqrt{n}) - f(X_i) \right) - \|\Delta\|_2^2 \right\} \right] - \lambda^2 \right| \to 0. \tag{7}$$

Furthermore, we show that λ is bounded with high probability when n is large. Upon combining (6) and (7) with the boundedness of λ, we obtain the desired central limit theorem. □

Theorem 2 demonstrates a parametric rate of convergence, in contrast with the standard $O\left(n^{-1/(d \vee 2)}\right)$ convergence rate of Wasserstein distances (see, e.g., [8, Theorem 1]). The technical details and complete proof are presented in Appendix A.3.

4 Extension to Non-Compact Spaces

Our previous discussions and results on strong duality and statistical convergence have been limited to the case of compact domains. We now turn to consider results on strong duality and statistical convergence for the case when the sample space Ω is not compact.

We start by considering our results on strong duality in the case of non-compact domains, and then considering our results on the rate of statistical convergence in the case of non-compact domains, both following along the lines of Example 3 above.

Theorem 3. *Consider $\Omega = \mathbb{R}^d$ and a continuous cost function $c(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ with $c(x, x) = 0$. Assume that $\mathbb{E}_{P_*}[c(X, y)] < \infty$ for any $y \in \mathbb{R}^d$, and that the set $\left\{ x \in \mathbb{R}^d : c(x, x_0) \leq a \right\}$ is compact for any $a > 0$. Following the setting in Example 3, for linearly independent unit vectors θ_1,...,θ_K and $\mathcal{F}_{\mathcal{B}} = \mathcal{C}_b(\mathbb{R})$, we have the strong duality*

$$R_n = \sup_{f \in \mathcal{LB}(\mathbb{R}^d)} \left\{ \mathbb{E}_{P_*}[f(X)] - \mathbb{E}_{P_n}[f^c(X)] \right\}.$$

Sketch of proof. Since we have the weak duality proven in Theorem 1, we only need to show

$$\mathcal{D} := \sup_{f \in \mathcal{LB}(\mathbb{R}^d)} \left\{ \mathbb{E}_{P_*}\left[f(X)\right] - \mathbb{E}_{P_n}\left[f^c(X)\right] \right\} \geq R_n.$$

Our strategy for this proof is to pick a series of large compact sets, so that we can approximate the solution to the primal problem by restricting the functions $c(\cdot,\cdot)$ and f on the compact set.

We then apply strong duality for the compact problem and subsequently show that the dual optimal value $\mathcal{D}$ can be approximated by the dual optimal value of the compact problem, when we apply the truncation to the cost function $c_a(x,y) = \min\{a, c(x,y)\}$. Finally, the optimal value with the cost function $c_a(x,y)$ converges to the optimal value with the cost function $c(x,y)$. □

The detailed proof of Theorem 3 is provided in Appendix A.4. An important element which distinguishes the proof of the results from standard strong duality in optimal transport is that the usual technique to construct improving dual functions is not applicable since $f^c \notin \mathcal{LB}(\mathbb{R}^d)$ in general.

We next study the rate of statistical convergence within the context of Example 3.

Theorem 4. *Assume $\Omega = \mathbb{R}^d$ and the cost function $c(x,y) = \|x-y\|_2^2$ with $\mathbb{E}_{P_*}[\|X\|_2^{4+\epsilon}] < \infty$ for some $\epsilon > 0$. Let $M(P_*) = \max\left\{1, \mathbb{E}_{P_*}[\|X\|_2^{4+\epsilon}]\right\}$. Following the setting in Example 3, for linearly independent unit vectors θ_1,...,θ_K and any $\mathcal{F}_\mathcal{B} \subset \mathcal{C}_b(\mathbb{R})$, there exists a universal constant C such that $\mathbb{E}[R_n] \leq C\rho^* K (M(P_*))^2 n^{-1/2}$, where ρ^* denotes the spectral radius of the matrix $C_K = [\theta_1, \theta_2, \ldots, \theta_K]^\top$.*

The key to the proof is to perform the transformation $Y_K = C_K X$ and to apply the standard convergence result in [8, Theorem 1]. The technical details and complete proof are provided in Appendix A.5.

Remark 3. *The convergence rate $O_p(1/\sqrt{n})$ in Theorem 4 is slower than the rate $O_p(1/n)$ in Theorem 2. We emphasize that the rate $O_p(1/\sqrt{n})$ is also tight in situations where the support is non-compact. It is consistent with the observation in the one-dimensional Wasserstein distance of order 2 [6, Corollary 5.10].*

5 Numerical Experiments

We provide experimental results on testing the hypothesis that a set of n samples $\{X_1, X_2, \ldots, X_n\} \subset \mathbb{R}^d$ is compatible with a candidate distribution P_* for a set of user-desired characteristics, specifically the test described in Example 3. The projection directions $\{\theta_1, \ldots, \theta_K\}$ could be viewed as the characteristics of interest to the user (as discussed in the introduction). Theorem 4 shows that, if the hypothesis is true, then the robust Wasserstein profile function $R_n = O_p(n^{-1/2})$. We implement the test by first estimating this distribution of R_n in its dual form (3). The hypothesis test can then be conducted in a standard manner by constructing the test statistic R_n for the given empirical distribution $P_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ and checking whether it is within the desired confidence level.

The key step in estimating R_n is to solve for $f^c(x)$. Let $\Omega = \mathbb{R}^d$, $C_K = [\theta_1, \theta_2, \ldots, \theta_K]^\top$ and $\Gamma_K = C_K C_K^\top$. We then have $f^c(x) = \sup_{z \in \mathbb{R}^K}\left(\sum_{j=1}^K f_j(\theta_j^\top x + z^{(j)}) - z^\top \Gamma_K^{-1} z\right)$, referring to Appendix B.2 for the technical details. Therefore, the inner supremum is a K-dimensional optimization problem instead of a d-dimensional problem.

We use Marr wavelet basis functions [13] to approximate the function class $\mathcal{B}(\Omega)$. In particular, we use a finite collection $\{b_l\}_{l=1}^L$ of Marr wavelet bases, where we provide the explicit expressions in Appendix B.1. Hence, the R_n is approximated by :

$$\hat{R}_n = \sup_{w_{jl}} \left\{ \mathbb{E}_{P_*}\left[\sum_{j=1}^K \sum_{l=1}^L w_{jl} b_l(\theta_j^\top X)\right] - \frac{1}{n}\sum_{i=1}^n \sup_{z_i} \sum_{j=i}^K \sum_{l=1}^L \left[w_{jl} b_l(\theta_j^\top x_i + z_i^{(j)}) - z_i^\top \Gamma_k^{-1} z_i\right] \right\}.$$

Stochastic approximation (SA, also known as SGD) iterations are used to obtain the optimal solution of $\hat{R}_n$. Specifically, each SA iteration estimates expectations $\mathbb{E}_{P_*}$ using a mini-batch sample from P_* of size 50. During each iteration, the n inner supremum problems are solved by Newton iterations with 150 restarts (see Appendix B.5 for the details).

To reject the hypothesis that the given set is from P_*, we use the 95% quantile of the distribution of $\hat{R}_n$ obtained when the empirical sets are indeed from P_* as a threshold. We construct an estimate of this quantile from the empirical distribution of $\hat{R}_n$ obtained by from 50 instances of n sized samples P_n generated from P_*. The P_* distribution is an equal mixture of four standard Gaussians with $d = 20$. Our n-sized test set P_n^{alt} is from an alternate distribution P_*^{alt} that is also a mixture of standard Gaussians but with different centering points. The test statistic $\hat{R}_n^{\text{alt}}$ computed for P_n^{alt} against P_* is thus tested against the 95% quantile of $\hat{R}_n$ to decide on the hypothesis that P_n^{alt} is from P_*. Three ($K = 3$) projection directions θ_j are carefully chosen to be linearly independent and such that they can reveal the modes of P_*, the user-preferred characteristics of interest. We set $n = 25$ and choose $L \sim 30$ basis functions. Each computational run to estimate $\hat{R}_n(\hat{R}_n^{\text{alt}})$ for a given $P_n(P_n^{\text{alt}})$ takes on average 10 minutes to compute on a dual AMD EPYC 7301 16-Core Processor machine with 64GB of memory utilizing 50 subprocesses to solve the inner supremum problems in parallel.

We report the results in Figure 1. On the left we plot the projection of the two distributions P_* (blue shade) and P_*^{alt} (red shade) on the three θ_j directions. Notice that each plot reveals that P_* has at least three modes along the projection direction θ_j; while, on the other hand, these directions θ_j reveal only one mode each for P_*^{alt}. The right plot of Figure 1 shows the distributions of $\hat{R}_n$ (blue histogram) and $\hat{R}_n^{\text{alt}}$ (red histogram) estimated by computing for $\hat{R}_n$ and $\hat{R}_n^{\text{alt}}$ repeatedly for 50 times each. The black dashed line marks the estimated 95% quantile of $\hat{R}_n$. In this case, we control the type I error as 5% and obtain a type II error as 32%. This shows that the method, based on our theoretical results, can efficiently distinguish between two $d = 20$ distributions in terms of the user-preferred characteristics while providing good accuracy even for relatively small values of n.

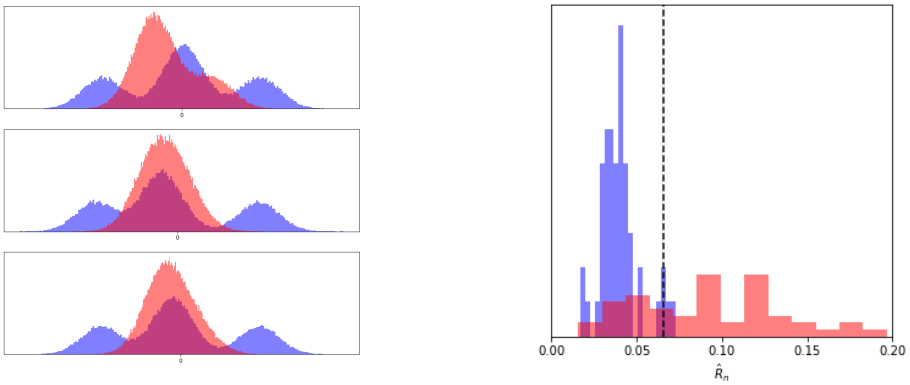

Figure 1: Left: projections of P^* (blue shade) and P_*^{alt} (red shade) along the three θ_j directions; Right: histograms of 50 samples of $\hat{R}_n$ (blue histogram) and $\hat{R}_n^{\text{alt}}$ (red histogram) with the 95% quantile of $\hat{R}_n$ marked as a dashed black line.

6 Discussion

Motivated by the intuition that decision makers may only be concerned with some characteristics instead of all the details of the entire distribution, we consider the problem of projecting the empirical measure under the Wasserstein distance to a set of probability measures that are constrained to satisfy a family of expectations over a class of functions. In particular, we study theoretical aspects of the robust Wasserstein profile functions R_n. We believe this work provides important insights into the empirical success of the Wasserstein distance despite the curse of dimensionality. Interesting future directions include studying statistical convergence for general function classes, developing efficient algorithms to compute R_n, and applying our methods in practice leveraging our theoretical insights.

Acknowledgement

Material in this paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397. Additional support is gratefully acknowledged from NSF grants 1915967, 1820942 and 1838576.

Broader Impact

This is a theoretical contribution that, nevertheless, has the potential of impacting a wide range of application domains in business, engineering and science. In particular, all of those in which the Wasserstein distance has been extensively used as a statistical inference tool (e.g. image analysis and computer vision, signal processing, operations research, and so on). Because our paper provides a step towards breaking the curse of dimensionality in statistical rates of convergence, we believe that we have the potential of enabling more applications to multiple hypothesis testing (e.g., certifying Wasserstein GANs). In turn, we plan to improve human resource development by including some of the main findings in this paper in Ph.D. courses.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[2] Jose Blanchet, Lin Chen, and Xun Yu Zhou. Distributionally robust mean-variance portfolio selection with Wasserstein distances. *arXiv preprint arXiv:1802.04885*, 2018.

[3] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

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

[5] Jose Blanchet, Karthyek Murthy, and Nian Si. Confidence regions in Wasserstein distributionally robust estimation. *arXiv preprint arXiv:1906.01614*, 2019.

[6] Sergey Bobkov and Michel Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society, 2019.

[7] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-Sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.

[8] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[9] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems (NIPS) 28*, 2015.

[10] Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[11] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

[12] Jun-Ya Gotoh, Michael Jong Kim, and Andrew EB Lim. Calibration of distributionally robust empirical optimization models. *arXiv preprint arXiv:1711.06565*, 2017.

[13] A. Graps. An introduction to wavelets. *IEEE Computational Science and Engineering*, 2(2):50–61, 1995.

[14] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[15] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, July 2017.

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

[17] Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling*. Springer Science & Business Media, 1999.

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

[19] Maurice Sion et al. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

[20] Carla Tameling, Max Sommerfeld, and Axel Munk. Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29(5):2744–2781, 2019.

[21] Aad W Van der Vaart. *Asymptotic statistics*. Cambridge University press, 2000.

[22] A.W. Van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics Springer Series in Statistics*. Springer, 1996.

[23] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[24] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[25] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

[26] Jonathan Weed and Quentin Berthet. Estimation of smooth densities in Wasserstein distance. *arXiv preprint arXiv:1902.01778*, 2019.

[27] Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262 – 267, 2018.

Appendix A Proofs of Main Results

Appendix A.1 Proof of Theorem 1

Let $\mathcal{M}(\Omega), \mathcal{M}_+(\Omega)$ and $\mathcal{M}_+^a(\Omega)$ denote the set of all signed measures, positive Borel measures and positive Borel measures with total mass less than or equal to a on the underlying space Ω, respectively, where $\mathcal{M}(\Omega), \mathcal{M}_+(\Omega)$ and $\mathcal{M}_+^a(\Omega)$ are equipped with the weak topology. Recall that, in the weak topology, $\pi_n \Rightarrow \pi$ if and only if, for every continuous and bounded function $h : \Omega \to \mathbb{R}$, we have that $\int h \mathrm{d}\pi_n \to \int h \mathrm{d}\pi$ as $n \to \infty$.

We will prove some of our duality results in this section, where we divide the proof into two steps:

1. Prove the weak duality in general spaces.
2. Prove the strong duality in compact spaces.

Let us first rewrite R_0 as a linear programming problem:

$$\begin{aligned} R_0 = & \inf_{\pi \in \mathcal{M}_+(\Omega\times\Omega), P\in\mathcal{M}_+(\Omega)} \int_{\Omega\times\Omega} c(x,y)\pi(\mathrm{d}x,\mathrm{d}y), \qquad \text{(A.1)} \\ s.t. \quad & \pi(A\times\Omega) = P_0(A), \pi(\Omega\times A) = P(A) \text{ for every measurable set } A, \\ & \mathbb{E}_P[f(X)] = \mathbb{E}_{P_*}[f(X)] \text{ for all } f \in \mathcal{B}(\Omega). \end{aligned}$$

Appendix A.1.1 Weak Duality

We consider the set $\mathcal{JL}(c) = \{(\alpha,\beta) \in L_1(P_*) \times L_1(P_0) : \alpha(x) + \beta(y) \le c(x,y)\}$. Notice that

$$R_0 = \inf_{\pi\in\mathcal{M}_+(\Omega\times\Omega), P\in\mathcal{M}_+(\Omega)} \sup_{(\alpha,\beta)\in L_1(P_*)\times L_1(P_0), f\in\mathcal{LB}(\Omega)} \left\{ \int_{\Omega\times\Omega} c(x,y)\pi(\mathrm{d}x,\mathrm{d}y) + \right.$$

$$\left. (\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_P[f(X)]) + \mathbb{E}_P[\alpha(X)] + \int_\Omega \beta(y)P_0(\mathrm{d}y) - \int_{\Omega\times\Omega} (\alpha(x)+\beta(y))\pi(\mathrm{d}x,\mathrm{d}y) \right\}.$$

Letting $\alpha(x) = f(x)$, we obtain

$$\begin{aligned} R_0 \ge & \inf_{\pi\in\mathcal{M}_+(\Omega\times\Omega), P\in\mathcal{M}_+(\Omega)} \sup_{\beta\in\mathcal{C}_b(\Omega), f\in\mathcal{LB}(\Omega)} \mathbb{E}_{P_*}[f(X)] \\ & + \int_{\Omega\times\Omega} c(x,y)\mathrm{d}\pi(x,y) + \left[\int_\Omega \beta(y)P_0(\mathrm{d}y) - \int_{\Omega\times\Omega} (f(x)+\beta(y))\pi(\mathrm{d}x,\mathrm{d}y) \right] \\ \ge & \inf_{\pi\in\mathcal{M}_+(\Omega\times\Omega), P\in\mathcal{M}_+(\Omega)} \sup_{(f,\beta)\in\mathcal{JL}(c), f\in\mathcal{LB}(\Omega)} \mathbb{E}_{P_*}[f(X)] \\ & + \left(\int_{\Omega\times\Omega} c(x,y)\pi(\mathrm{d}x,\mathrm{d}y) - \int_{\Omega\times\Omega} (f(x)+\beta(y))\pi(\mathrm{d}x,\mathrm{d}y) \right) + \int_\Omega \beta(y)P_0(\mathrm{d}y) \\ \ge & \inf_{\pi\in\mathcal{M}_+(\Omega\times\Omega), P\in\mathcal{M}_+(\Omega)} \sup_{(f,\beta)\in\mathcal{JL}(c), f\in\mathcal{LB}(\Omega)} \mathbb{E}_{P_*}[f(X)] + \int_\Omega \beta(y)P_0(\mathrm{d}y) \\ = & \sup_{(f,\beta)\in\mathcal{JL}(c), f\in\mathcal{LB}(\Omega)} \mathbb{E}_{P_*}[f(X)] + \int_\Omega \beta(y)P_0(\mathrm{d}y). \end{aligned}$$

It is readily verified that $f(x) - f^c(y) \le c(x,y)$ and $f^c(y)$ is lower semicontinuous (see, e.g., [24, Remark 5.5]), and thus measurable. Since $f^c(x) \ge f(x)$, we have $\mathbb{E}_{P_0}[f^c(X)] > -\infty$. Moreover, if $\mathbb{E}_{P_0}[f^c(X)] = +\infty$, then $\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_{P_0}[f^c(X)] = -\infty$. By choosing $f = \beta = 0$, we obtain

$$\sup_{(f,\beta)\in\mathcal{JL}(c), f\in\mathcal{LB}(\Omega)} \mathbb{E}_{P_*}[f(X)] + \int_\Omega \beta(y)P_0(\mathrm{d}y) \ge 0.$$

If $f^c(\cdot) \notin L_1(P_0)$, then $f^c(\cdot)$ cannot be the optimizer, and therefore we have

$$\sup_{(f,\beta)\in\mathcal{JL}(c),f\in\mathcal{LB}(\Omega)} \left(\mathbb{E}_{P_*}[f(X)] + \int_\Omega \beta(y)P_0(\mathrm{d}y)\right) = \sup_{f\in\mathcal{LB}(\Omega)} \left(\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_{P_0}[f^c(X)]\right).$$

This completes the weak duality proof.

Appendix A.1.2 Strong Duality in Compact Spaces

We assume Ω is a compact space and consider the set $\mathcal{J}(c) = \{(\alpha,\beta)\in\mathcal{C}(\Omega)\times\mathcal{C}(\Omega) : \alpha(x)+\beta(y)\le c(x,y)\}$. Notice that, in compact spaces, $\mathcal{C}(\Omega) = \mathcal{C}_b(\Omega)$ and thus $\mathcal{C}(\Omega)\in L_1(P)$ for any probability measure $P\in\mathcal{P}(\Omega)$. Sion's minimax Theorem [19], which will be useful for our proof, can be expressed as follows.

Theorem A1 (Sion's minimax Theorem)**.** *Consider two convex spaces M and N, one of which is compact, and let $g: M\times N\to R$ be such that, for each $y\in N$, $g(\cdot,y)$ is lower semicontinous and convex and, for each $x\in N$, $g(x,\cdot)$ is upper semicontinous and concave. Then,*

$$\inf_{x\in M}\sup_{y\in N} g(x,y) = \sup_{y\in N}\inf_{x\in M} g(x,y).$$

We now apply Sion's minimax Theorem. First define

$$\begin{aligned} g((\pi,P),(\alpha,\beta)) \;=\; & \int_{\Omega\times\Omega} c(x,y)\pi(\mathrm{d}x,\mathrm{d}y) + \left(\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_P[f(X)]\right) \\ & + \mathbb{E}_P[\alpha(X)] + \int_\Omega \beta(y)P_0(\mathrm{d}y) - \int_{\Omega\times\Omega} \left(\alpha(x)+\beta(y)\right)\pi(\mathrm{d}x,\mathrm{d}y). \end{aligned}$$

Next, for each $a\ge 1$, note that

$$\begin{aligned} R_0 = \inf_{(\pi,P)\in\mathcal{M}_+^a(\Omega\times\Omega)\times\mathcal{M}_+^a(\Omega)} \sup_{(\alpha,\beta,f)\in\mathcal{C}(\Omega)\times\mathcal{C}(\Omega)\times\mathcal{B}(\Omega)} \Bigg\{ & \int_{\Omega\times\Omega} c(x,y)\pi(\mathrm{d}x,\mathrm{d}y) + \\ & \left(\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_P[f(X)]\right) + \mathbb{E}_P[\alpha(X)] + \int_\Omega \beta(y)P_0(\mathrm{d}y) - \int_{\Omega\times\Omega} \left(\alpha(x)+\beta(y)\right)\pi(\mathrm{d}x,\mathrm{d}y) \Bigg\}. \end{aligned}$$

Let spaces $M = \mathcal{M}_+^a(\Omega\times\Omega)\times\mathcal{M}_+^a(\Omega)$ and $N = \mathcal{C}(\Omega)\times\mathcal{C}(\Omega)\times\mathcal{B}(\Omega)$ be equipped with the product of weak topology and product of uniform topology, respectively. Both M and N are convex spaces and M is compact by Prohorov's Theorem (see, e.g., [21, Theorem 2.4]), since Ω is compact.

Clearly, the function $g((\pi,P),(\alpha,\beta,f))$ is linear in (π,P) and (α,β,f) so that $g(\cdot)$ is convex-concave as required by Sion's Theorem. We claim that $g(\cdot,(\alpha,\beta))$ is continous under the weak topology. For any $(\pi_n,P_n)\Rightarrow(\pi,P)$, we have for any continous and bounded functions ϕ_1 and ϕ_2

$$\int\phi_1\pi_n + \int\phi_2 P_n \to \int\phi_1\pi + \int\phi_2 P.$$

Since $c(\cdot),\alpha(\cdot),\beta(\cdot)$ and $g(\cdot)$ are continuous functions on a compact space, they are all bounded and therefore, by the definition of weak convergence, we immediately obtain that

$$g((\pi_n,P_n),(\alpha,\beta)) \to g((\pi,P),(\alpha,\beta)).$$

On the other hand, for any $(\alpha_n,\beta_n,f_n)\to(\alpha,\beta,f)$ uniformly, we have

$$g((\pi,P),(\alpha_n,\beta_n,f_n)) \to g((\pi,P),(\alpha,\beta,f)),$$

given that $\pi(\Omega\times\Omega)<\infty$ and $P(\Omega\times\Omega)<\infty$, by the bounded convergence theorem.

Hence, we now can apply Sion's duality:

$$\begin{aligned}
R_0 &= \sup_{(\alpha,\beta,f)\in\mathcal{C}(\Omega)\times C(\Omega)\times\mathcal{LB}(\Omega)} \inf_{(\pi,P)\in\mathcal{M}_+^a(\Omega\times\Omega)\times\mathcal{M}_+^a(\Omega)} \left\{ \int_{\Omega\times\Omega} c(x,y)\pi(\mathrm{d}x,\mathrm{d}y) + \right. \\
&\quad \left(\mathbb{E}_{P_*}\left[f(X)\right] - \int_\Omega f(x)P(\mathrm{d}x)\right) + \int_\Omega \alpha(x)P(\mathrm{d}x) + \int_\Omega \beta(y)P_0(\mathrm{d}y) \\
&\quad \left. - \int_{\Omega\times\Omega} \left(\alpha(x)+\beta(y)\right)\pi(\mathrm{d}x,\mathrm{d}y) \right\} \\
&= \sup_{(\alpha,\beta,f)\in\mathcal{C}(\Omega)\times\mathcal{C}(\Omega)\times\mathcal{LB}(\Omega)} \inf_{(\pi,P)\in\mathcal{M}_+^a(\Omega\times\Omega)\times\mathcal{M}_+^a(\Omega)} \left\{ \left(\mathbb{E}_{\mathbb{P}_*}\left[f(X)\right] - \mathbb{E}_P\left[f(X)\right]\right) + \right. \\
&\quad \left. \mathbb{E}_P\left[\alpha(X)\right] + \int_\Omega \beta(y)P_0(\mathrm{d}y) + \int_{\Omega\times\Omega} \left(c(x,y)-\alpha(x)-\beta(y)\right)\pi(\mathrm{d}x,\mathrm{d}y) \right\}.
\end{aligned}$$

By choosing $\alpha = \beta = f = 0$, we conclude that $R_0 \geq 0$.

We first claim that there is no incentive for the sup player to select functions $\alpha(\cdot)$ and $\beta(\cdot)$ such that $\alpha(x)+\beta(y) > c(x,y)$ for some $(x,y)\in\Omega\times\Omega$. In that case, let $\pi(\{x,y\}) = a$, $\pi\left(\{x,y\}^c\right) = 0$ and $P = P_*$. With $a \geq 1$ being chosen arbitrarily, this implies that

$$R_0 \leq \int_\Omega f(x)P(\mathrm{d}x) + \int_\Omega \beta(y)P_0(\mathrm{d}y) - \left(\alpha(x)+\beta(y)-c(x,y)\right)a < 0.$$

Therefore, we conclude that

$$\begin{aligned}
R_0 &= \sup_{(\alpha,\beta,f)\in\mathcal{J}(c)\times\mathcal{LB}(\Omega)} \inf_{(\pi,P)\in\mathcal{M}_+^a(\Omega\times\Omega)\times\mathcal{M}_+^a(\Omega)} \left\{ \left(\mathbb{E}_{P_*}\left[f(X)\right] - \int_\Omega f(x)P(\mathrm{d}x)\right) + \right. \\
&\quad \left. \int_\Omega \alpha(x)P(\mathrm{d}x) + \int_\Omega \beta(y)P_0(\mathrm{d}y) + \int_{\Omega\times\Omega} \left(c(x,y)-\alpha(x)-\beta(y)\right)\pi(\mathrm{d}x,\mathrm{d}y) \right\}.
\end{aligned}$$

We then claim that $\alpha(x) < f(x)$ for some $x\in\Omega$ is impossible. By choosing $P\left(\{x\}\right) = a$, we have $P\left(\{x\}^c\right) = 0$, $\pi\left(\Omega\times\Omega\right) = 0$ and $\beta(\cdot) = 0$. By choosing sufficiently large a, we obtain

$$R_0 \leq \mathbb{E}_{P_*}\left[f(X)\right] - a\left(f(x)-\alpha(x)\right) < 0.$$

Therefore, we conclude that

$$\begin{aligned}
R_0 = \sup_{(\alpha,\beta,f)\in\mathcal{J}(c)\times\mathcal{LB}(\Omega),f\leq\alpha} \inf_{(\pi,P)\in\mathcal{M}_+^a(\Omega\times\Omega)\times\mathcal{M}_+^a(\Omega)} &\left\{ \left(\mathbb{E}_{P_*}\left[f(X)\right] - \int_\Omega f(x)P(\mathrm{d}x)\right) + \right. \\
&\left. \int_\Omega \alpha(x)P(\mathrm{d}x) + \int_\Omega \beta(y)P_0(\mathrm{d}y) + \int_{\Omega\times\Omega} \left(c(x,y)-\alpha(x)-\beta(y)\right)\pi(\mathrm{d}x,\mathrm{d}y) \right\}.
\end{aligned}$$

For the inner infimum, we can always choose $\pi\left(\Omega\times\Omega\right) = 0$ and $P\left(\Omega\right) = 0$. Notice that $\alpha^c(x)$ is a continuous function in the compact space and thus

$$\begin{aligned}
R_0 &\leq \sup_{(\alpha,\beta,f)\in\mathcal{J}(c)\times\mathcal{LB}(\Omega),f\leq\alpha} \mathbb{E}_{P_*}\left[f(X)\right] + \int_\Omega \beta(y)P_0(\mathrm{d}y) \\
&= \sup_{(\alpha,\beta,f)\in\mathcal{J}(c)\times\mathcal{LB}(\Omega),f\leq\alpha} \mathbb{E}_{P_*}\left[f(X)\right] - \int_\Omega \alpha^c\left(y\right)P_0(\mathrm{d}y) \\
&\leq \sup_{f\in\mathcal{LB}(\Omega)} \mathbb{E}_{P_*}\left[f(X)\right] - \int_\Omega f^c\left(y\right)P_0(\mathrm{d}y),
\end{aligned}$$

where the last inequality follows from

$$\sup_{x'\in\Omega}\left\{\alpha(x') - c(x',x)\right\} \geq \sup_{x'\in\Omega}\left\{f(x') - c(x',x)\right\}, \quad \text{if } \alpha(\cdot)\geq f(\cdot).$$

Appendix A.2 Proof of Proposition 1

By the Kantorovich-Rubinstein duality [24, Theorem 5.10], we have

$$\begin{aligned}
&\max_{\theta\in\Theta} \mathcal{W}_2\left(\theta_\sharp P, \theta_\sharp Q\right)^2 \\
= &\max_{\theta\in\Theta} \max_{f\in\mathcal{C}_b(\mathbb{R})} \left(\mathbb{E}_P\left[f\left(\theta^\top X\right)\right] - \mathbb{E}_Q\left[\sup_{t\in\mathbb{R}} f(t) - \left(\theta^\top X - t\right)^2\right]\right) \\
= &\max_{\theta\in\Theta, f\in\mathcal{C}_b(\mathbb{R})} \left(\mathbb{E}_P\left[f\left(\theta^\top X\right)\right] - \mathbb{E}_Q\left[\sup_{y\in\mathbb{R}^d} f(\theta^\top y) - \left(\theta^\top X - \theta^\top y\right)^2\right]\right).
\end{aligned}$$

On the other hand, for any θ with $\|\theta\|_2 = 1$ and $f \in \mathcal{C}_b(\mathbb{R})$, we obtain

$$\begin{aligned}
\sup_{y\in\mathbb{R}^d}\left(f(\theta^\top y) - \|x-y\|_2^2\right) &= \sup_{y\in\mathbb{R}^d}\left(f(\theta^\top y) - \inf_{y':\theta^\top y'=\theta^\top y}\|x-y'\|_2^2\right) \\
&= \sup_{y\in\mathbb{R}^d}\left(f(\theta^\top y) - \left(\theta^\top x - \theta^\top y\right)^2\right).
\end{aligned}$$

Therefore, we have

$$\max_{\theta\in\Theta} \mathcal{W}_2\left(\theta_\sharp P, \theta_\sharp Q\right)^2 = \sup_{f\in\mathcal{B}_{\max}(\mathbb{R}^d,\Theta)}\left(\mathbb{E}_P\left[f\left(X\right)\right] - \mathbb{E}_Q\left[\sup_{y\in\mathbb{R}^d}\left(f(y) - \|x-y\|_2^2\right)\right]\right),$$

which completes the proof.

Appendix A.3 Proof of Theorem 2

Define

$$\mathcal{UB}_i(\Omega) = \left\{f(\theta_i^\top\cdot) \in \mathcal{LB}_i(\Omega) : \mathbb{E}_{P_*}\left[f'(\theta_i^\top X)^2\right] = 1, f(0) = 0\right\},$$

and accordingly

$$\mathcal{UB}(\Omega) = \left\{f(\cdot) \in \mathcal{LB}(\Omega) : \mathbb{E}_{P_*}\left[\|\nabla_X f(X)\|_2^2\right] = 1, f(0) = 0\right\}.$$

Since P_* has a non-zero density, we have $\mathcal{LB}(\Omega) = \{\lambda f(\cdot) + c \,|\, \lambda, c \in \mathbb{R}, f(\cdot) \in \mathcal{UB}(\Omega)\}$.

By Theorem 1, we obtain

$$\begin{aligned}
R_n &= \sup_{\lambda\in\mathbb{R}}\sup_{f\in\mathcal{UB}(\Omega)}\left\{E_P\left[\lambda f(X)\right] - \frac{1}{n}\sum_{i=1}^n\left[\sup_{\Delta+X_i\in\Omega}\left\{\lambda f(X_i+\Delta) - \|\Delta\|_2^2\right\}\right]\right\} \\
&= \sup_{\lambda\in\mathbb{R}}\sup_{f\in\mathcal{UB}(\Omega)}\left\{E_P\left[\lambda f(X)\right] - \frac{1}{n}\sum_{i=1}^n \lambda f(X_i)\right. \\
&- \left.\frac{1}{n}\sum_{i=1}^n\left[\sup_{\Delta+X_i\in\Omega}\left\{\lambda f(X_i+\Delta) - \lambda f(X_i) - \|\Delta\|_2^2\right\}\right]\right\}.
\end{aligned}$$

Let $H_n^f = n^{-1/2}\left(\sum_{i=1}^n f(X_i) - E_P\left[f(X)\right]\right)$. By rescaling the variables λ and Δ, we have

$$nR_n =$$

$$\sup_{\lambda\in\mathbb{R}}\sup_{f\in\mathcal{UB}(\Omega)}\left\{-2\lambda H_n^f - \frac{1}{n}\sum_{i=1}^n\left[\sup_{X_i+\Delta/\sqrt{n}\in\Omega}\left\{2\lambda\sqrt{n}\left(f(X_i+\Delta/\sqrt{n}) - f(X_i)\right) - \|\Delta\|_2^2\right\}\right]\right\}.$$

The following sequence of results will then be useful in proving Theorem 2. To simplify the notation, we denote

$$M_n\left(\lambda, f\right) = \frac{1}{n}\sum_{i=1}^n\left[\sup_{X_i+\Delta/\sqrt{n}\in\Omega}\left\{2\lambda\sqrt{n}\left(f(X_i+\Delta/\sqrt{n}) - f(X_i)\right) - \|\Delta\|_2^2\right\}\right].$$

Henceforth, we refer to a function class as a Donsker class or a Glivenko-Cantelli class if the function class is a P-Donsker class or P-Glivenko-Cantelli class for all Borel measure P supported on the sample domain Ω.

Proposition A1. *There exists $M_B < \infty$ such that*

$$\mathcal{UB}(\Omega) \subset \mathcal{F}^{M_B} := \left\{ \sum_{i=1}^{K} \xi_i f_i\left(\theta_i^\top \cdot\right) : |\xi_i| \leq M_B \text{ and } f_i\left(\theta_i^\top \cdot\right) \in \mathcal{UB}_i(\Omega) \right\},$$

and thus $\mathcal{UB}(\Omega)$ is a Donsker class.

Proposition A2. *The function class $\mathcal{UB}'(\Omega) = \left\{ \|\nabla_x f(\cdot)\|_2^2 : f \in \mathcal{UB}(\Omega) \right\}$ is a Glivenko-Cantelli class. Furthermore, $\mathcal{UB}'_\epsilon(\Omega) = \left\{ \|\nabla_x f(\cdot)\|_2^2 \mathbb{I}\{B_\epsilon(\cdot) \subset \Omega\} : f \in \mathcal{UB}(\Omega) \right\}$ are also Glivenko-Cantelli classes, where $B_\epsilon(x)$ is a closed ball around x with radius ϵ, i.e., $B_\epsilon(x) = \{y \in \mathbb{R}^d : \|x - y\| \leq \epsilon\}$.*

Proposition A3. *For every $\epsilon > 0$, there exists $n_0 > 0$ and $b \in (0, \infty)$ such that*

$$\mathbb{P}\left(\sup_{|\lambda| > b} \sup_{f \in \mathcal{UB}(\Omega)} \left\{ -2\lambda H_n^f - M_n(\lambda, f) \right\} > 0 \right) \leq \epsilon,$$

for all $n \geq n_0$.

Proposition A4. *We have*

$$\sup_{|\lambda| \leq b} \sup_{f \in \mathcal{UB}(\Omega)} \left(\left| M_n(\lambda, f) - \lambda^2 \right| \right) \to 0,$$

almost surely.

Based on Propositions A1 – A4, we are now ready to present the proof of Theorem 2.

Proof of Theorem 2. By the uniform convergence property of Donsker classes and Glivenko-Cantelli classes, we have

$$\sup_{|\lambda| \leq b} \sup_{f \in \mathcal{UB}(\Omega)} \left\{ -2\lambda H_n^f - M_n(\lambda, f) \right\} \Rightarrow \sup_{|\lambda| \leq b} \sup_{f \in \mathcal{UB}(\Omega)} \left\{ -2\lambda H^f - \lambda^2 \right\},$$

where H^f is a Gaussian process with $H^f \sim \mathcal{N}(0, \operatorname{var}(f(X)))$ and $\operatorname{cov}(H^{f_1}, H^{f_2}) = \operatorname{cov}(f_1(X), f_2(X))$. Furthermore, Proposition A3 implies

$$\mathbb{P}\left(\sup_{\lambda \in \mathbb{R}} \sup_{f \in \mathcal{UB}(\Omega)} \left\{ -2\lambda H_n^f - M_n(\lambda, f) \right\} = \sup_{|\lambda| \leq b} \sup_{f \in \mathcal{UB}(\Omega)} \left\{ -2\lambda H_n^f - M_n(\lambda, f) \right\} \right) \geq 1 - \epsilon.$$

Therefore, by Slutsky's Theorem, we obtain

$$\begin{aligned} \sup_{\lambda \in \mathbb{R}} \sup_{f \in \mathcal{UB}(\Omega)} \left\{ -2\lambda H_n^f - M_n(\lambda, f) \right\} \quad &\Rightarrow \quad \sup_{\lambda \in \mathbb{R}} \sup_{f \in \mathcal{UB}(\Omega)} \left\{ -2\lambda H^f - \lambda^2 \right\} \\ &= \quad \sup_{f \in \mathcal{LB}(\Omega)} \left\{ -2H^f - \mathbb{E}_{P_*}\left[\|\nabla_X f(X)\|_2^2 \right] \right\}. \end{aligned}$$

□

Appendix A.3.1 Proofs of Propositions A1 – A4

Lemma A1. *$\mathcal{UB}_i(\Omega)$ is a Donsker class.*

Proof. By Assumptions 2 and 3, we have for any $f(\theta_i^\top \cdot) \in \mathcal{UB}_i(\Omega)$,

$$\sup_{x \in \theta_i^\top \Omega} |f''(x)| \leq M \sqrt{\int_\Omega f'(\theta_i^\top x)^2 \mathrm{d}x} \leq \frac{M}{\sqrt{\underline{b}}} \sqrt{\mathbb{E}_{P_*}\left[f'(\theta_i^\top X)^2 \right]} = \underline{b}^{-1/2} M,$$

where $\theta_i^\top \Omega$ is defined as $\theta_i^\top \Omega = \{\theta_i^\top x : x \in \Omega\}$. Notice that $|f'(x)| = \left| f'(0) + \int_0^x f''(s) \mathrm{d}s \right|$ and

$$|f'(0)| + M|x| \geq |f'(x)| \geq |f'(0)| - \underline{b}^{-1/2} M |x|. \tag{A.2}$$

Recall that $\mathbb{E}_{P_*}\left[f'(\theta_i^\top X)^2\right] = 1$, and thus

$$1 = \mathbb{E}_{P_*}\left[f'(\theta_i^\top X)^2\right] \geq \mathbb{E}_{P_*}\left[\left(|f'(0)| - \underline{b}^{-1/2} M \left|\theta_i^\top X\right|\right)^2\right].$$

Since $\mathbb{E}_{P_*}\left[f'\left(\theta_i^\top X\right)^2\right]$ is bounded, we have $|f'(0)|$ is bounded from above. Furthermore, by (A.2), we conclude that $|f'(x)|$ is bounded from above in $\theta_i^\top \Omega$, which means $f(x)$ is a Lipschitz function with a bounded Lipschitz constant. By Example 19.9 in [21], we have the space of one-dimensional bounded Lipschitz functions with $f(0) = 0$ is a Donsker class for all Borel probability measures supported in $\theta_i^\top \Omega$. Recalling $(\theta_i)_\sharp P$ is the push-forward measure from $\mathcal{P}(\Omega)$ to $\mathcal{P}\left(\theta_i^\top \Omega\right)$ defined in (4), we then obtain $\mathbb{E}_{(\theta_i)_\sharp P}\left[f\left(X\right)\right] = \mathbb{E}_P\left[f\left(\theta_i^\top X\right)\right]$. We conclude $\mathcal{UB}_i(\Omega)$ is also a Donsker class for all Borel probability measures supported in Ω. □

Lemma A2. *For two positive semidefinite matrices $A, B \in \mathbb{R}_{K\times K}$, define the Hadamard product $A \circ B$ as*

$$(A \circ B)_{ij} = A_{ij} B_{ij}.$$

Then, if $B_{ii} = 1, \forall i$, we have $\sigma_K(A \circ B) \geq \sigma_K(A)$, where $\sigma_K(\cdot)$ denotes the smallest eigenvalue of the matrix.

Proof. Denote the k-th leading principle minor of a matrix A as $[A]_{1:k,1:k}$. Let $\mu = \sigma_K(A)$. Then, $A - \mu I$ is positive semidefinite. By Oppenheim's inequality (see, e.g., [14, Theorem 7.8.16]), we have $\det\left([A-\mu I]_{1:k,1:k} \circ [B]_{1:k,1:k}\right) \geq \det[A-\mu I]_{1:k,1:k} \geq 0$. Hence, $(A-\mu I)\circ B = A \circ B - \mu I$ is also positive semi-definite and $\sigma_K(A \circ B) \geq \mu$. □

We are now ready to present the proof of Proposition A1.

Proof of Proposition A1.

$$\mathbb{E}_{P_*}\left[\left\|\nabla_x f(x)\right\|_2^2\right] = \mathbb{E}_{P_*}\left[\left\|\sum_{i=1}^K \lambda_i f_i'(\theta_i^\top X)\theta_i\right\|_2^2\right] = \xi^T (A \circ B)\,\xi,$$

where

$$B_{ij} = \mathbb{E}\left[f_i'(\theta_i^\top X) f_j'(\theta_j^\top X)\right] \text{ and } A_{ij} = \theta_i^\top \theta_j.$$

Since, by construction, A and B are both positive semidefinite matrices, then by Lemma A2 we have

$$1 = \xi^T (A \circ B)\,\xi \geq \left(\xi^\top \xi\right)\sigma_K(A).$$

Since $\theta_1, \ldots, \theta_K$ are linearly independent, $\sigma_K(A) > 0$. Letting $M_B = \sqrt{1/\sigma_K(A)}$, we then have $|\xi_i| \leq \sqrt{1/\sigma_K(A)}$ and $\mathcal{F}^{M_B}$ is a Donsker class by Theorem 2.10.3 in [22]. Therefore, by Theorem 2.10.1 in [22], we have $\mathcal{UB}(\Omega)$ is a Donsker class. □

Lemma A3. *For any function $f \in \mathcal{UB}(\Omega)$, define the Hessian matrix of $H^f(x)$ as $H_{ij}^f(x) = \frac{\partial}{\partial x_i \partial x_j} f(x)$. Then, there exist universal constants M_1 and M_2 such that $\sup_{f\in\mathcal{UB}(\Omega)} \sup_{x\in\Omega} \left\|\nabla_x f(x)\right\|_2 \leq M_1$ and $\sup_{f\in\mathcal{UB}(\Omega)} \sup_{x\in\Omega} \left\|H^f(x)\right\|_F \leq M_2$.*

Proof. For $f(x) = \sum_{i=1}^K \lambda_i f_i(\theta_i^\top x)$, we have

$$\left\|\nabla_x f(x)\right\|_2 \leq \sum_{i=1}^K \left\|\lambda_i f_i'(\theta_i^\top x)\theta_i\right\|_2 \leq K M_B \sup_{1\leq i\leq K} \sup_{x\in\theta_i^\top\Omega} \left|f_i'(x)\right|,$$

where, by Lemma A1, $\sup_{1\leq i\leq K} \sup_{x\in\theta_i^\top\Omega} |f_i'(x)| < \infty$. Furthermore, we obtain

$$\left|H_{ij}^f(x)\right| \leq \sum_{k=1}^K \left|\lambda_k f_k''(\theta_k^\top x)\theta_k^{(i)}\theta_k^{(j)}\right| \leq \underline{b}^{-1/2} M K M_B.$$

Let $M_1 = K M_B \sup_{1\leq i\leq K} \sup_{x\in\theta_i^\top\Omega} |f_i'(x)|$, $M_2 = d^2 \underline{b}^{-1/2} M K M_B$, and thus $\sup_f \sup_x \left\|\nabla_x f(x)\right\|_2 \leq M_1$ and $\left\|H^f(x)\right\|_F \leq M_2$ for any $f \in \mathcal{UB}(\Omega)$ and $x \in \Omega$. □

We are now ready to present the proof of Proposition A2.

Proof of Proposition A2. Consider

$$\begin{aligned}\frac{\partial \left\|\nabla_x f(x)\right\|_2^2}{\partial x_k} &= \frac{\partial \left\|\sum_{i=1}^K \lambda_i f_i'(\theta_i^\top x)\theta_i\right\|_2^2}{\partial x_k} \\ &= \sum_{i=1}^K\sum_{j=1}^K \lambda_i\lambda_j\theta_i^\top\theta_j\left(f_i''(\theta_i^\top x)f_j'\left(\theta_j^\top x\right)\theta_i^{(k)} + f_i'(\theta_i^\top x)f_j''\left(\theta_j^\top x\right)\theta_j^{(k)}\right) \\ &\leq 2K^2M_B^2\left(\sup_{1\leq i\leq K}\sup_{x\in\theta_i^\top\Omega}|f_i'(x)|\right)\underline{b}^{-1}M.\end{aligned}$$

Therefore, the partial derivative $\frac{\partial\|\nabla_x f(x)\|_2^2}{\partial x_k}$ is bounded. By Example 19.9 and Theorem 19.4 in [21], we have $\mathcal{UB}'(\Omega)$ is a Glivenko-Cantelli class. Furthermore, since the bracketing number of $\mathcal{UB}'_\epsilon(\Omega)$ is bounded by that of $\mathcal{UB}'(\Omega)$, we have $\mathcal{UB}'_\epsilon(\Omega)$ are also Glivenko-Cantelli classes. □

We are now ready to present the proof of Proposition A3.

Proof of Proposition A3. By Taylor expansion, we have

$$\left|f(X_i+\Delta/\sqrt{n}) - f(X_i) - \frac{1}{\sqrt{n}}\left(\nabla_X f(X_i)\right)^\top\Delta\right| \leq M_2\left\|\Delta\right\|_2^2/n.$$

By substituting $\Delta_i = c_1\nabla_X f(X_i)$, we obtain

$$\begin{aligned}&\sup_{X_i+\Delta/\sqrt{n}\in\Omega}\left\{2\lambda\sqrt{n}\left(f(X_i+\Delta/\sqrt{n}) - f(X_i)\right) - \left\|\Delta\right\|_2^2\right\} \\ \geq\ &\left(2\left|\lambda\right|\left(\left(\nabla_X f(X_i)\right)^\top\Delta_i - M_2\left\|\Delta_i\right\|_2^2/\sqrt{n}\right) - \left\|\Delta_i\right\|_2^2\right)\mathbb{I}\{X_i+\Delta_i/\sqrt{n}\in\Omega\} \\ =\ &\left(2\left|\lambda\right|c_1 - c_1^2 - 2\left|\lambda\right|c_1^2M_2/\sqrt{n}\right)\left\|\nabla_X f(X_i)\right\|_2^2\mathbb{I}\{X_i+c_1\nabla_X f(X_i)/\sqrt{n}\in\Omega\} \\ \geq\ &\left(2\left|\lambda\right|c_1 - c_1^2 - 2\left|\lambda\right|c_1^2M_2/\sqrt{n}\right)\left\|\nabla_X f(X_i)\right\|_2^2\mathbb{I}\{B_{c_1M_1/\sqrt{n}}(X_i)\in\Omega\}.\end{aligned}$$

We then derive

$$\begin{aligned}&\sup_{|\lambda|>b}\sup_{f\in\mathcal{UB}(\Omega)}\left\{-2\lambda H_n^f - M_n(\lambda,f)\right\} \\ \leq\ &\sup_{|\lambda|>b}\sup_{f\in\mathcal{UB}(\Omega)}\left\{-2\lambda H_n^f\right. \\ &\left.-\frac{1}{n}\sum_{i=1}^n\left(2\left|\lambda\right|c_1 - c_1^2 - 2\left|\lambda\right|c_1^2M_2/\sqrt{n}\right)\left\|\nabla_X f(X_i)\right\|_2^2\mathbb{I}\{B_{c_1M_1/\sqrt{n}}(X_i)\in\Omega\}\right\} \\ \leq\ &\sup_{|\lambda|>b}\sup_{f\in\mathcal{UB}(\Omega)}\left\{-2\lambda H_n^f\right. \\ &\left.-\left|\lambda\right|\frac{1}{n}\sum_{i=1}^n\left(2c_1 - \frac{c_1^2}{b} - 2c_1^2M_2/\sqrt{n}\right)\left\|\nabla_X f(X_i)\right\|_2^2\mathbb{I}\{B_{c_1M_1/\sqrt{n}}(X_i)\in\Omega\}\right\} \\ \leq\ &\sup_{|\lambda|>b}\left|\lambda\right|\left(2\sup_{f\in\mathcal{UB}(\Omega)}\left|H_n^f\right|\right. \\ &\left.-\left(2c_1 - \frac{c_1^2}{b} - 2c_1^2M_2/\sqrt{n}\right)\inf_{f\in\mathcal{UB}(\Omega)}\left(\frac{1}{n}\sum_{i=1}^n\left\|\nabla_X f(X_i)\right\|_2^2\right)\mathbb{I}\{B_{c_1M_1/\sqrt{n}}(X_i)\in\Omega\}\right),\end{aligned}$$

where $B_\varepsilon(X_i) = \{y\in\mathbb{R}^d : \|y-X_i\|_2\leq\varepsilon\}$. Since $\mathcal{UB}(\Omega)$ is a Donsker class, we have

$$\sup_{f\in\mathcal{UB}(\Omega)}\left|H_n^f\right| \Rightarrow \sup_{f\in\mathcal{UB}(\Omega)}\left|H^f\right|,$$

where $\sup_{f\in\mathcal{UB}(\Omega)}\left|H^f\right|<\infty$ almost surely. Hence, there exist n_1 and b' such that, for $n\geq n_1$, $\mathbb{P}\left(\sup_{f\in\mathcal{UB}(\Omega)}\left|H_n^f\right|>b'\right)<\epsilon/2$.

Since $P_*(\Omega^\circ)=1$, we can choose $\varepsilon'>0$ such that

$$\mathbb{E}_{P_*}\left[\left\|\nabla_X f(X_i)\right\|_2^2\mathbb{I}\{B_{\varepsilon'}(X_i)\in\Omega\}\right]>\frac{3}{4}.$$

By Lemma A1, there exists $n_2>n_1$ such that, for $n\geq n_2$,

$$\mathbb{P}\left(\inf_{f\in\mathcal{UB}(\Omega)}\left(\frac{1}{n}\sum_{i=1}^n\left\|\nabla_X f(X_i)\right\|_2^2\right)\mathbb{I}\{B_{\varepsilon'}(X_i)\in\Omega\}\leq 1/2\right)<\epsilon/2.$$

Letting $c_1=4b'$, $b=2c_1$ and for $n>n_0=\max\{(4c_1M_2)^2,(c_1M_1/\varepsilon')^2,n_2\}$, we then have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{|\lambda|>b}\sup_{f\in\mathcal{UB}(\Omega)}\left\{-2\lambda H_n^f-M_n(\lambda,f)\right\}>0\right)\\
\leq\ &\mathbb{P}\left(\inf_{f\in\mathcal{UB}(\Omega)}\left(\frac{1}{n}\sum_{i=1}^n\left\|\nabla_X f(X_i)\right\|_2^2\right)\mathbb{I}\{B_{\varepsilon'}(X_i)\in\Omega\}\leq 1/2\right)\\
+\ &\mathbb{P}\left(\sup_{f\in\mathcal{UB}(\Omega)}\left|H_n^f\right|>b'\right)\\
<\ &\epsilon.
\end{aligned}$$

□

We are now ready to present the proof of Proposition A4.

Proof of Proposition A4. First note that, for $|\lambda|\leq b$, we have

$$\begin{aligned}
&2\lambda\sqrt{n}\left(f(X_i+\Delta/\sqrt{n})-f(X_i)\right)-\left\|\Delta\right\|_2^2\\
=\ &2\lambda\int_0^1\left(\nabla_X f(X_i+n^{-1/2}\Delta u)\right)^\top\Delta\mathrm{d}u-\left\|\Delta\right\|_2^2\\
\leq\ &2b\int_0^1\left\|\nabla_X f(X_i+n^{-1/2}\Delta u)\right\|_2\left\|\Delta\right\|_2\mathrm{d}u-\left\|\Delta\right\|_2^2\\
\leq\ &2bM_1\left\|\Delta\right\|_2-\left\|\Delta\right\|_2^2.
\end{aligned}$$

Therefore, we only need to consider $\left\|\Delta\right\|_2\leq 2bM_1$. Recalling the Taylor expansion

$$\sup_{\|\Delta\|\leq 2bM_1}\left|f(X_i+\Delta/\sqrt{n})-f(X_i)-n^{-1/2}\left(\nabla_X f(X_i)\right)^\top\Delta\right|\leq\frac{M_2}{n}\left\|\Delta\right\|_2^2<\frac{M_2}{n}\left(2bM_1\right)^2,$$

we then obtain

$$\begin{aligned}
&\sup_{|\lambda|\leq b}\sup_{f\in\mathcal{UB}(\Omega)}\left(\sup_{X_i+\Delta/\sqrt{n}\in\Omega,\|\Delta\|\leq 2bM_1}M_n(\lambda,f)\right.\\
&\left.-\frac{1}{n}\sum_{i=1}^n\sup_{X_i+\Delta/\sqrt{n}\in\Omega,\|\Delta\|\leq 2bM_1}\left(2\lambda\left(\nabla_X f(X_i)\right)^\top\Delta-\left\|\Delta\right\|_2^2\right)\right)\\
\leq\ &M_2\left(2bM_1\right)^2/\sqrt{n}\to 0.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\lambda^2\left\|\nabla_X f(X_i)\right\|_2^2\mathbb{I}\{B_{\lambda M_1/\sqrt{n}}(X_i)\in\Omega\}\\
\leq\ &\sup_{X_i+\Delta/\sqrt{n}\in\Omega,\|\Delta\|\leq 2bM_1}\left(2\lambda\left(\nabla_X f(X_i)\right)^\top\Delta-\left\|\Delta\right\|_2^2\right)\leq\lambda^2\left\|\nabla_X f(X_i)\right\|_2^2.
\end{aligned}$$

Since $P_*(\Omega^\circ) = 1$, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\left(\lambda^2\left\|\nabla_x f(X_i)\right\|_2^2 \mathbb{I}\{B_{\lambda M_1/\sqrt{n}}(X_i)\in\Omega\} - \lambda^2\left\|\nabla_x f(X_i)\right\|_2^2\right)\to 0,$$

almost surely. Therefore, we conclude

$$\sup_{|\lambda|\le b}\sup_{f\in\mathcal{UB}(\Omega)}\left|M_n(\lambda,f) - \frac{1}{n}\sum_{i=1}^{n}\lambda^2\left\|\nabla_x f(X_i)\right\|_2^2\right|\to 0,$$

almost surely. By Lemma A2, we have the desired results. □

Appendix A.4 Proof of Theorem 3

We apply a line of arguments similar to steps 2 and 3 in the proof of Theorem 1.3 in [23]. To simplify the notation, we define $\mathcal{V} := R_n$ and

$$\mathcal{D} := \sup_{f\in\mathcal{LB}(\mathbb{R}^d)}\left\{\mathbb{E}_{P_*}\left[f(X)\right] - \mathbb{E}_{P_n}\left[f^c(X)\right]\right\}.$$

Since $\mathcal{V}\ge\mathcal{D}$ is proved in Appendix A.1.1, we only need to show $\mathcal{D}\ge\mathcal{V}$. The strategy of this proof is to pick a series of large compact sets, so that we can approximate the solution to the primal problem by restricting the functions $c(\cdot,\cdot)$ and f on the compact set.

We next apply strong duality for the compact problem and then show that the dual optimal value $\mathcal{D}$ can be approximated by the dual optimal value of the compact problem, when we apply the truncation to the cost function $c_a(x,y) = \min\{a, c(x,y)\}$. Finally, we show that the optimal value with the cost function $c_a(x,y)$ converges to the optimal value with the cost function $c(x,y)$.

Appendix A.4.1 Primal Approximation

For any $\epsilon > 0$, we pick a large compact set $\mathcal{K}$ such that

$$P_*(\mathcal{K}) > 1-\epsilon,\, P_n(\mathcal{K}) = 1 \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_*}\left[c(X_i,X)\mathbb{I}\{X\in\mathcal{K}^c\}\right] < \epsilon.$$

Define the measure $P_{\mathcal{K}}$ supported on $\mathcal{K}$ with $P_{\mathcal{K}}(A) = P_*(A)/P_*(\mathcal{K})$ for any Borel measurable set $A\subset\mathcal{K}$. Then, consider the primal problem restricted in space $\mathcal{K}$:

$$\mathcal{V}_{\mathcal{K}} = \inf_{P\in\mathcal{P}(\mathcal{K})}\left\{\mathcal{D}_c(P_n,P) : \mathbb{E}_{P_*}\left[f(X)\right] = \mathbb{E}_{P_{\mathcal{K}}}\left[f(X)\right] \text{ for all } f\in\mathcal{B}(\mathbb{R}^d)\big|_{\mathcal{K}}\right\}, \tag{A.3}$$

where $\mathcal{B}(\mathcal{X})|_{\mathcal{K}}$ is the restriction of $\mathcal{B}(\mathbb{R}^d)$ on set $\mathcal{K}$. Notice that, for any feasible solution $P\in\mathcal{P}(\mathcal{K})$ of problem (A.3), we can construct

$$P'(A) = P(A\cap\mathcal{K})P_*(\mathcal{K}) + P_*(A\cap\mathcal{K}^c),$$

which is a feasible solution of problem (1). Let $\pi_{\mathcal{K}}$ be the coupling between P_n and P. Then, we can define a coupling between P_n and P' as π^ϵ:

$$\pi^\epsilon(\{X_i\},A) = \pi_{\mathcal{K}}(\{X_i\},A\cap\mathcal{K})P_*(\mathcal{K}) + \frac{1}{n}P_*(A\cap\mathcal{K}^c)$$

for $i = 1,2,\ldots,n$ and any Borel measurable set $A\subset\mathcal{X}$. Then, for every feasible P, we have

$$\mathcal{V}\le\mathcal{D}_c(P_n,P')\le P_*(\mathcal{K})\mathcal{D}_c(P_n,P) + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_*}\left[c(X_i,X)\mathbb{I}\{X\in\mathcal{K}^c\}\right].$$

Therefore, we obtain $\mathcal{V}\le\mathcal{V}_{\mathcal{K}}+\epsilon$.

Appendix A.4.2 Dual Approximation

We first find unit vectors $\theta_{K+1}, \theta_{K+2}, \ldots, \theta_d$ such that $\theta_1, \theta_2, \ldots, \theta_d$ are linearly independent and thus are a basis of $\mathbb{R}^d$. Define compact sets $\mathcal{K}_m$ as

$$\mathcal{K}_m = \bigcap_{i=1}^{d} \left\{ x \in \mathbb{R}^d \mid \theta_i^\top x \in [-m, m] \right\}.$$

It is easy to see that $\mathcal{K}_m$ is a nonempty compact set for $m > 0$, given $\theta_1, \theta_2, \ldots, \theta_d$ are linearly independent. We pick m sufficiently large such that $P_n(\mathcal{K}_m) = 1$. Define the dual problem

$$\mathcal{D}_m = \sup_{f \in \mathcal{LB}(\mathbb{R}^d)|_{\mathcal{K}_m}} \left\{ \mathbb{E}_{P_{\mathcal{K}_m}} \left[f(X) \right] - \mathbb{E}_{P_n} \left[f^c(X) \right] \right\}. \tag{A.4}$$

Then, for any $f(x) = \sum_{i=1}^K f_i\left(\theta_i^\top x\right) \in \mathcal{LB}(\mathbb{R}^d)\big|_{\mathcal{K}_m}$, we define $\bar{f}(x)$ as the extension of $f(x)$ to $\mathbb{R}^d$: for any $z \in \mathbb{R}^d$, let x^* be the unique solution of the linear equation system

$$\theta_i^\top x = \max\left\{ \min\left\{ \left(\theta_i^\top z\right), m \right\}, -m \right\},\ i = 1, 2, \ldots, d;$$

then, $x^* \in \mathcal{K}_m$ and let $\bar{f}(z) = f(x^*) = \sum_{i=1}^K f_i\left(\theta_i^\top x^*\right)$. Therefore, $\bar{f}(z) \in \mathcal{LB}(\mathbb{R}^d)$.

We consider the truncated cost function $c_a(\cdot) = \min\{a, c(\cdot)\}$ for $0 < a < \infty$. Let f be an ϵ-optimizer of problem (A.4) with the cost function $c_a(\cdot)$. Since $\mathcal{D}_m \geq 0$, there exist $x_0 \in \mathcal{K}_m$ and $y_0 \in \{X_i\}_{i=1}^n$ such that (assuming $0 < \epsilon < 1$)

$$f(x_0) - f^{c_a}(y_0) \geq -1.$$

Without loss of generality, we assume $f_i(\theta_i^\top x_0) \geq -1/K$ for $i = 1, 2, \ldots, K$ and $f^c(y_0) \leq 1$. Then, we obtain

$$\begin{aligned}
f(x) &\leq f^{c_a}(y_0) + c_a(x, x_0) \leq a + 1, \text{ for } x \in \mathcal{K}_m, \\
f^{c_a}(x) &\geq f(x_0) - c_a(x, x_0) \geq -a - 1, \text{ and} \\
f^{c_a}(x) &= \sup_{y \in \mathcal{K}_m} f(y) - c_a(x, y) \leq a + 1, \text{ for } x \in \{X_i\}_{i=1}^n.
\end{aligned}$$

By construction, we have $\bar{f}(x) \leq a + 1$ for any $x \in \mathbb{R}^d$. Since $\left\{ x \in \mathbb{R}^d : c(x, x_0) \leq a \right\}$ is compact, we are able to pick a sufficiently large $\mathcal{K}_m$ such that

$$\inf_{x \in \mathcal{K}_m^c, y \in \{X_i\}_{i=1}^n} c_a(x, y) = a.$$

Therefore, we obtain $\bar{f}^{c_a}(x) = f^{c_a}(x)$ for $x \in \{X_i\}_{i=1}^n$.

Then, for $z \in \mathbb{R}$ and any $j = 1, 2, \ldots, K$, let x' be the unique solution of the linear system

$$\begin{aligned}
\theta_j^\top x &= z; \\
\theta_i^\top x &= \theta_i^\top x_0,\ i = 1, 2, \ldots j-1, j+1, \ldots, d.
\end{aligned}$$

Since $\sum_{i=1}^K \bar{f}_i(\theta_i^\top x') = \bar{f}(x') \leq a + 1$, we have

$$\bar{f}_j(z) \leq a + 1 - \sum_{i=1, i \neq j}^{K} \bar{f}_i(\theta_i^\top x') \leq a + 2. \tag{A.5}$$

Furthermore, we claim $g(x) = \sum_{i=1}^K \max\left\{ \bar{f}_i(\theta_i^\top x), -K(a+2) \right\}$ is a valid ϵ-optimizer with $g^{c_a}(x) = \bar{f}^{c_a}(x)$ for $x \in \{X_i\}_{i=1}^n$. This is because for any $y_0 \in \{X_i\}_{i=1}^n$, if $\bar{f}_i\left(\theta_i^\top x\right) \leq -K(a+2)$, we have

$$\bar{f}(x) - c(x, y_0) \leq -K(a+2) + (K-1)(a+2) = -(a+2) < \bar{f}^{c_a}(y_0).$$

Therefore, $\mathbb{E}_{P_{\mathcal{K}_m}}\left[g(x)|_{\mathcal{K}_m} \right] \geq \mathbb{E}_{P_{\mathcal{K}_m}}\left[f(x) \right]$ with bounds $a + 1 \geq g(x) \geq -K^2(a+2)$. Finally, by picking sufficiently large $\mathcal{K}_m$ with $P_*(\mathcal{K}_m) > 1 - \epsilon$, we obtain

$$\begin{aligned}
& \mathbb{E}_{P_*}\left[g(X) \right] - \mathbb{E}_{P_n}\left[g^{c_a}(X) \right] \\
\geq\ & (1 - \epsilon)\, \mathbb{E}_{P_{\mathcal{K}_m}}\left[f(X) \right] + \mathbb{E}_{P_*}\left[g(X) \mathbb{I}\{X \notin \mathcal{K}_m\} \right] - \mathbb{E}_{P_n}\left[f^{c_a}(X) \right] \\
\geq\ & \mathbb{E}_{P_{\mathcal{K}_m}}\left[f(X) \right] - \mathbb{E}_{P_n}\left[f^{c_a}(X) \right] - \epsilon\,(a+1) - \epsilon K^2 (a+2) \\
\geq\ & \mathcal{D}_m - \epsilon\left(1 + a + 1 + K^2(a+2)\right).
\end{aligned}$$

By the arbitrariness of ϵ, we complete the proof for the bounded cost function.

Appendix A.4.3 Unbounded Cost Function

The following lemma is useful for finishing the last part of the proof.

Lemma A4. *Let $c_a(\cdot) = \min\{a, c(\cdot)\}$. For any ϵ, let $P^{\epsilon}_{(a)}$ be an ϵ-optimizer for the problem*

$$\inf_{P\in\mathcal{P}(\mathbb{R}^d)} \left\{\mathcal{D}_{c_a}(P_n, P) : \mathbb{E}_{P_*}[f(X)] = \mathbb{E}_P[f(X)] \text{ for all } f \in \mathcal{B}(\mathbb{R}^d)\right\}.$$

Then, the set $\left\{P^{\epsilon}_{(a)}\right\}_{a=1}^{\infty}$ is relatively compact in the space $\mathcal{P}(\mathbb{R}^d)$ equipped with the topology of weak convergence.

Proof of Lemma A4. First, we have

$$\mathcal{D}_{c_a}(P_n, P^{\epsilon}_{(a)}) \le \mathcal{D}_{c_a}(P_n, P_*) + \epsilon \le \mathcal{D}_c(P_n, P_*) + \epsilon < \infty.$$

If the set $\left\{P^{\epsilon}_{(a)}\right\}_{a=1}^{\infty}$ is not relatively compact, then by Prokhorov's Theorem, there exists $\epsilon' > 0$ such that, for any compact set $\mathcal{K}$ and any $a_0 > 0$, we can find an $a > a_0$ with $P^{\epsilon}_{(a)}(\mathcal{K}) > \epsilon'$.

We pick $a_0 = \lceil (\mathcal{D}_c(P_n, P_*) + \epsilon)/\epsilon' \rceil$ and a sufficient large $\mathcal{K}$ such that

$$\inf_{x\in\mathcal{K}^c, y\in\{X_i\}_{i=1}^n} c(x, y) > a_0.$$

Then, for any $a > a_0$, we have

$$\mathcal{D}_{c_a}(P_n, P^{\epsilon}_{(a)}) > a_0\epsilon' \ge \mathcal{D}_c(P_n, P_*) + \epsilon,$$

which leads to a contradiction. □

Next, we define the space $\Pi\left(P_n, P_*, \mathcal{B}(\mathbb{R}^d)\right)$ as

$$\Pi\left(P_n, P_*, \mathcal{B}(\mathbb{R}^d)\right) := \left\{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : \pi(A\times\Omega) = P_n(A), \pi(\Omega\times A) = P(A) \right.$$
$$\left. \text{for every measurable set } A \subset \mathbb{R}^d, \text{ and } \mathbb{E}_P[f(X)] = \mathbb{E}_{P_*}[f(X)] \text{ for all } f \in \mathcal{B}(\mathbb{R}^d)\right\}.$$

We then have

$$R_n = \inf_{\pi\in\Pi(P_n, P_*, \mathcal{B}(\mathbb{R}^d))} \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x, y)\pi(\mathrm{d}x, \mathrm{d}y).$$

Now let I, I_a be respectively defined on $\Pi\left(P_n, P_*, \mathcal{B}(\mathbb{R}^d)\right)$ by

$$I_a[\pi] = \int_{\mathbb{R}^d\times\mathbb{R}^d} c_a(x, y)\pi(\mathrm{d}x, \mathrm{d}y) \qquad \text{and} \qquad I[\pi] = \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x, y)\pi(\mathrm{d}x, \mathrm{d}y).$$

By Appendix A.4.2, we obtain

$$\inf_{\pi\in\Pi(P_n, P_*, \mathcal{B}(\mathbb{R}^d))} I_a[\pi] = \sup_{f\in\mathcal{LB}(\mathbb{R}^d)} \left\{\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_{P_n}[f^{c_a}(X)]\right\}.$$

We conclude the argument by showing that

$$\inf_{\pi\in\Pi(P_n, P_*, \mathcal{B}(\mathbb{R}^d))} I[\pi] = \sup_a \inf_{\pi\in\Pi(P_n, P_*, \mathcal{B}(\mathbb{R}^d))} I_a[\pi] \tag{A.6}$$

and that, for each a,

$$\sup_{f\in\mathcal{LB}(\mathbb{R}^d)} \left\{\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_{P_n}[f^{c_a}(X)]\right\} \le \sup_{f\in\mathcal{LB}(\mathbb{R}^d)} \left\{\mathbb{E}_{P_*}[f(X)] - \mathbb{E}_{P_n}[f^{c}(X)]\right\}. \tag{A.7}$$

Then, by the combination of (A.6), (A.7) and the weak duality, we will have the desired results.

Since $\inf I_a$ is a nondecreasing sequence, bounded above by $\inf I$, we only need to prove that

$$\lim_{a\to\infty} \inf_{\pi\in\Pi(P_n, P_*, \mathcal{B}(\mathbb{R}^d))} I_a(\pi) \ge \inf_{\pi\in\Pi(P_n, P_*, \mathcal{B}(\mathbb{R}^d))} I(\pi).$$

Let π^{ϵ}_a be an optimal coupling between P_n and $P^{\epsilon}_{(a)}$ defined in Lemma A4. By the tightness of $\left\{P^{\epsilon}_{(a)}\right\}_{a=1}^{\infty}$, we have that the sequence $\{\pi^{\epsilon}_a\}_{a=1}^{\infty}$ is also tight. Therefore, by Prokhorov's Theorem,

we are able to extract a subsequence $\left\{\pi^{\epsilon}_{a_k}\right\}_{k=1}^{\infty}$, where $\pi^{\epsilon}_{a_k}$ converges weakly to a probability measure $\pi^{\epsilon}_{*} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ as $k \to \infty$, in the sense that for any bounded continuous function θ on $\mathbb{R}^d \times \mathbb{R}^d$

$$\int \theta(x,y) d\pi^{\epsilon}_{a_k}(\mathrm{d}x, \mathrm{d}y) \to \int \theta(x,y) d\pi^{\epsilon}_{*}(\mathrm{d}x, \mathrm{d}y).$$

From this we observe that $\pi^{\epsilon}_{*} \in \Pi\left(P_n, P_*, \mathcal{B}(\mathbb{R}^d)\right)$. Then, whenever $a \geq b$, one has

$$I_a\left[\pi^{\epsilon}_{a}\right] \geq I_b\left[\pi^{\epsilon}_{a}\right].$$

By the boundedness of $c_b(\cdot,\cdot)$, we obtain

$$\limsup_{a\to\infty} I_a\left[\pi^{\epsilon}_{a}\right] \geq \limsup_{a\to\infty} I_b\left[\pi^{\epsilon}_{a}\right] \geq I_b\left[\pi^{\epsilon}_{*}\right].$$

By monotone convergence, $I_b\left[\pi^{\epsilon}_{*}\right] \to I\left[\pi^{\epsilon}_{*}\right]$ as $b \to \infty$, and thus

$$\lim_{a\to\infty} \inf_{\pi\in\Pi(P_n,P_*,\mathcal{B}(\mathbb{R}^d))} I_a(\pi) \geq \limsup_{a\to\infty} I_a\left[\pi^{\epsilon}_{a}\right] - \epsilon \geq I\left[\pi^{\epsilon}_{*}\right] - \epsilon \geq \inf_{\pi\in\Pi(P_n,P_*,\mathcal{B}(\mathbb{R}^d))} I(\pi) - \epsilon.$$

Then, by the arbitrariness of ϵ, we have the desired results and conclude the proof.

Appendix A.5 Proof of Theorem 4

Define $\mathcal{D}_A(P,Q) = \mathcal{D}_c(P,Q)$ with cost function $c(x,y) = (x-y)^\top A(x-y)$ for any positive definite matrix A. Then, we have

$$R_n \leq \inf_{P\in\mathcal{P}(\mathbb{R}^d)} \left\{\mathcal{D}_I(P,P_n) : \mathbb{E}_P\left[f(\theta_i^\top X)\right] = \mathbb{E}_{P_*}\left[f(\theta_i^\top X)\right], \forall f \in \mathcal{C}_b(\mathbb{R}), \theta_1,\ldots,\theta_K \in \mathbb{R}^d\right\},$$

where $K \leq d$ and $\theta_1,\ldots,\theta_K$ are linearly independent. We first find orthonormal vectors $\theta_{K+1},\theta_{K+2},\ldots,\theta_d$ such that $\theta_1,\theta_2,\ldots,\theta_d$ are linearly independent and thus are a basis of $\mathbb{R}^d$. Let $Y_i = \theta_i^\top X$ for $i = 1,2,\ldots,d$, let P_*^Y denote the distribution of Y, and let P_n^Y denote the corresponding empirical distribution. Further let $C = [\theta_1,\theta_2,\ldots,\theta_d]^\top$, and then $Y = CX$. Therefore, we obtain

$$\begin{aligned} R_n &\leq \inf_{P\in\mathcal{P}(\mathbb{R}^d)} \left\{\mathcal{D}_I(P,P_n) : \mathbb{E}_P\left[f(\theta_i^\top X)\right] = \mathbb{E}_{P_*}\left[f(\theta_i^\top X)\right] \ \forall f \in \mathcal{C}_b(\mathbb{R}), i = 1,2,\ldots,K\right\} \\ &= \inf_{P\in\mathcal{P}(\mathbb{R}^d)} \left\{\mathcal{D}_A\left(P^Y,P_n^Y\right) : \mathbb{E}_{P^Y}[f(Y_i)] = \mathbb{E}_{P_*^Y}[f(Y_i)] \ \forall f \in \mathcal{C}_b(\mathbb{R}), i = 1,2,\ldots,K\right\} \end{aligned}$$

where $A = \left(CC^\top\right)^{-1}$. Then, notice that

$$\begin{aligned} \mathcal{D}_A\left(P^Y,P_n^Y\right) &= \inf_{\pi\in\mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d)} \left\{\left(\int (y-v)^\top A(y-v)\pi(\mathrm{d}y,\mathrm{d}v)\right)\right. \\ &\qquad \left. : \int_{v\in\mathbb{R}^d} \pi(\mathrm{d}y,\mathrm{d}v) = P^Y(\mathrm{d}y), \int_{y\in\mathbb{R}^d} \pi(\mathrm{d}y,\mathrm{d}v) = P_n^Y(\mathrm{d}v)\}\right\}. \end{aligned}$$

Let $\rho(A)$ denote the spectral radius of matrix A. We then have $\mathcal{D}_A\left(P^Y,P_n^Y\right) \leq \rho(A)\mathcal{D}_I\left(P^Y,P_n^Y\right)$ and

$$\begin{aligned} R_n &\leq \rho(A) \inf_{P\in\mathcal{P}(\mathbb{R}^d)} \left\{\mathcal{D}_I\left(P^Y,P_n^Y\right) : \mathbb{E}_{P^Y}[f(Y_i)] = \mathbb{E}_{P_*^Y}[f(X)] \ \forall f \in \mathcal{C}_b(\mathbb{R}), i = 1,\ldots,K\right\} \\ &= \rho(A)\sum_{i=1}^{K} \mathcal{W}_2^2\left(P_*^{Y_i},P_n^{Y_i}\right), \end{aligned} \tag{A.8}$$

where $P_*^{Y_i}$ and $P_n^{Y_i}$ are the push-forward measures of P_*^Y and P_n^Y from $\mathcal{P}(\mathbb{R}^d)$ to $\mathcal{P}(\mathbb{R})$ such that for any Borel set A in $\mathbb{R}$

$$\begin{aligned} P_*^{Y_i}(A) &= P_*^Y\left(\left\{x\in\mathbb{R}^d : x_i \in A\right\}\right), \\ P_n^{Y_i}(A) &= P_n^Y\left(\left\{x\in\mathbb{R}^d : x_i \in A\right\}\right). \end{aligned}$$

Notice that $\rho(A) = \rho\left(\left(C_K C_K^\top\right)^{-1}\right)$ for $C_K = [\theta_1,\theta_2,\ldots,\theta_K]^\top$. Therefore, $\rho(A)$ does not depend on our choices of $\theta_{K+1},\theta_{K+2},\ldots,\theta_d$. Finally, by Theorem 1 in [8], we obtain for the Wasserstein distance in the one-dimension case:

$$\mathbb{E}\left[\mathcal{W}_2^2\left(P_*^{Y_i},P_n^{Y_i}\right)\right] \leq C\left(\mathbb{E}\left[|Y_i|^{4+\epsilon}\right]\right)^{2/(4+\epsilon)}/\sqrt{n} \leq CM(P_*)^2 n^{-1/2}. \tag{A.9}$$

By substituting (A.9) into (A.8), we have the desired results.

Appendix B Appendix: Experiments

Appendix B.1 Marr Wavelet Basis for the f_j

We call $\phi(t)$ as the "mother" function for the wavelet basis. A function $f(t)$ is said to be written using a continuous wavelet basis as

$$f(t) = \int_s \int_u \frac{w_{s,u}}{\sqrt{s}} \phi\left(\frac{t-u}{s}\right) \mathrm{d}u\mathrm{d}s \quad \text{with } w_{u,s} = \int_t f(t) \frac{1}{\sqrt{s}} \phi\left(\frac{t-u}{s}\right) \mathrm{d}t,$$

where the index u is called the translation, index s the scaling, and $w_{u,s}$ the weights. Let $b_{u,s}(t) = \phi((t-u)/s)/\sqrt{s}$. Note that if $f(t)$ is a density then $w(u,s) = \mathbb{E}[b_{u,s}(t)]$. We will use a discrete version of this as our truncated basis, with a discrete set of $\{l = (u,s)\}_{l=1}^{L}$. Each f_j is thus represented as

$$f_j(v) = \sum_{l=1}^{L} w_{jl} b_l(v),$$

where the truncated sequence of the J terms gets us a finite dimensional variable $w = (w_{jl},\ j = 1, \ldots, K,\ l = 1, \ldots, L)$.

The discrete basis set is determined by user input on a desired domain $[-M, M]$ for the function approximation and a *granularity* value G. If $[-m_0, m_0]$ is the domain of the mother function ϕ, then the discrete pairs are determined as:

$$(u,s) = (km_0 2^{-g+1}, 2^g), \quad \forall\ k \in \{-K(g), \ldots, K(g)\},\ g = 0, \ldots, G, \tag{B.10}$$

where $K(g) = \left\lceil \frac{M}{m_0 2^{-g}} \right\rceil + 1$. These finite set of pairs of (u,s) over all g, k are used to constitute the index $l \in \{1, \ldots, L\}$.

We need a basis that yields relatively smooth values for the derivative:

$$\frac{\mathrm{d}f_j}{\mathrm{d}v}(v) = \sum_{l=1}^{L} w_{jl} \frac{\mathrm{d}b_l}{\mathrm{d}v}(v), \tag{B.11}$$

and thus we do not use the popular Haar basis, which yields $db_j/dv = 0$ a.e. We experiment with the Marr basis, also termed the (inverted) Mexican hat basis:

$$\phi(t) = \frac{2}{\sqrt{3}\pi^{1/4}}(1-t^2)e^{-t^2/2}, \quad \text{with} \quad \frac{\mathrm{d}\phi(t)}{\mathrm{d}t} = \frac{2}{\sqrt{3}\pi^{1/4}}(t^3 - 3t)e^{-t^2/2}. \tag{B.12}$$

Appendix B.2 Re-formulation of $f^c(x)$

For each x, we have that

$$f^c(x) = \sup_{\Delta} \left(\sum_{j=i}^{K} f_j(\theta_j^\top (x+\Delta)) - \|\Delta\|_2^2 \right).$$

By Substituting $z = C_K \Delta$, for C_K defined in Appendix A.5, we obtain

$$f^c(X_0) = \sup_{\Delta \in \mathbb{R}^d, v \in \mathbb{R}^K} \left(\sum_{j=i}^{K} f_j(\theta_j^\top x + v^{(j)})) - \|\Delta\|_2^2,\ s.t.\ v = C_K \Delta \right).$$

Keeping v fixed and maximizing only over Δ, we can see that the $\Delta^*(z)$ is the projection of the origin onto the linear subspace $C_K \Delta = z$. We can formally establish this by considering the Lagrangian formulation:

$$L(v, \Delta, \lambda) = \sum_{j=1}^{K} f_j(\theta_j^\top x + z^{(j)}) - \|\Delta\|_2^2 + \lambda^\top (C_K \Delta - z).$$

Setting up the first order optimality conditions, we have:

$$\begin{aligned}\nabla_z L &= \nabla_z f_j(C_K x + z) - \lambda &&= 0,\\ \nabla_\Delta L &= -2\Delta + C_K^\top \lambda &&= 0,\\ \nabla_\lambda L &= C_K \Delta - z &&= 0,\end{aligned}$$

where $\nabla_z f(C_K x + z) = [f_1'(\theta_1^\top x + z^{(1)}) \ , \ \ldots \ , \ f_K'(\theta_K^\top x + z^{(K)})]^\top$. Taking the last two equations, and recalling that $\Gamma_K = C_K C_K^\top$, we obtain that

$$\begin{aligned}\Delta &= \frac{1}{2} C_K^\top \lambda, & C_K \Delta &= \frac{1}{2}\Gamma_K \lambda = z,\\ \lambda &= 2(\Gamma_K)^{-1} z, & \Delta &= C_K^\top (\Gamma_K)^{-1} z,\\ \|\Delta\|_2^2 &= z^\top (\Gamma_K)^{-1} z.\end{aligned}$$

Substituting in the first equation renders the first order equation to be satisfied as follows:

$$\nabla_z f(C_K x + z) = 2(\Gamma_K)^{-1} z, \tag{B.13}$$

which is also the first order condition for maximizing

$$\sup_z \sum_{j=1}^K f_j(\theta_j^\top x + z^{(j)}) - z^\top (\Gamma_K)^{-1} z. \tag{B.14}$$

This is an optimization problem in z, a K-dimensional variable. Hence, the true complexity of the inner supremum is a K-dimensional problem. The rows of C_K are linearly independent by selection, so the inverse $(\Gamma_K)^{-1}$ exists.

Appendix B.3 Implementation of $\hat{R}_n$ Computation

The $\hat{R}_n$ problem has a suggestive sup-inf form: the w_{jl} variables attempt to emphasize the mass under P_* to maximize the expectation $\mathbb{E}_{P_*}$, while the z_n variables provide a mechanism for the P_n samples to attain the same high values by also moving themselves, thus negating the value of the first term but at a quadratic penalty cost. This suggests why $R_n \to 0$ as n grows and P_n is sampled from P_*: the variables z_n are able to attain the same expectations under P_n with low cost. On the other hand, if P_n comes from a different distribution than P^*, the weights w_{jl} have more leeway to emphasize the non-overlapping parts of P_*, thus driving the supremum higher.

The inner supremum over z is solved, as mentioned in the main body of the paper, using Newton iterations. Following (B.13), the Hessian can similarly be obtained as

$$\nabla_z^2 f(C_K x + z) - 2(\Gamma_K)^{-1}.$$

The problem (B.14) is a general non-linear optimization problem, and thus the Newton iterations return only locally optimal solutions. In order to obtain globally optimal solutions, this algorithm is restarted multiple times with randomized starting points. A particularly good set of initial starting points corresponds to samples from the P_* (in $\mathbb{R}^d$ space) since the purpose of the v (and equivalently Δ) is to successfuly improve its optimal function value by moving close to the support of P_*.

The (stochastic estimate of the) gradient with respect to w_{jl} for the stochastic approximation (SA) iterations of the outer optimization problem is obtained as:

$$G(w) = \frac{1}{M}\sum_{m=1}^M \left(\sum_{j=1}^K \sum_{l=1}^L b_l(\theta_j^\top X_m)\right) - \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^K \sum_{l=1}^L b_l(\theta_j^\top X_i + z_i^*)\right). \tag{B.15}$$

Appendix B.4 Algorithm

Here is the complete algorithm to solve the optimization problem defining $\hat{R}_n$ once for a given P_n.

- Given: set $\{X_i, \ i = 1, \ldots, n\}$ that form P_n, and a sampler for P_*.
- Given: gain sequence γ_r.

- Initialize wavelet weights $w^{(0)} = (w_{jl}^{(0)})$ uniformly from $[-1, 1]$.
- For $r = 1, \ldots$ (SA method for outer maximization of weights w_{jl})
  1. Sample $\{X_m,\ m = 1, \ldots, M\}$ from P_* to estimate expectations under P_*.
  2. Assemble m-th summand of the first term in (B.15) as:
$$\left(\nabla_w \left(\sum_{j=1}^{K} f_j(\theta_j^\top X_m)\right)\right)_{jl} = \sum_{j=1}^{K} b_l(\theta_j^\top X_m)$$
  3. For each $i = 1, \ldots, n$:
     (a) Estimate optimal Δ_i^* using deterministic Newton iterations and forming gradients from (B.11) and (B.12).
     (b) Return the i-th summand in the second term of (B.15) as
$$\left(\nabla_w \left(\sum_{j=1}^{K} f_j(\theta_j^\top (X_i + \Delta_i^*))\right)\right)_{jl} = \sum_{j=1}^{K} b_l\left(\theta_j^\top X_i + z_i^{(j)}\right),$$
     for $l = 1, \ldots, L$, and $j = 1, \ldots, K$.
  4. Assemble gradient for outer SA as given in (B.15) from the components in Steps (2) and (3b) above
  5. Set $w^{(r)} = w^{(r-1)} - \gamma_r G_r(w^{(r-1)})$

Appendix B.5 Experimental Setup Details

The P_* target distribution is set to be an equal weight mixture of four $d = 20$ dimensional unit-covariance Gaussians. The centers of the four Gaussians are $[-1, \ldots, -1]$, $[-1, 1, -1, \ldots, 1]$, $[1, -1, 1, \ldots, -1]$ and $[1, 1, 1, \ldots, 1]$. The P_*^{alt} is also an equal mixture of four Gaussians with their centers obtained by applying an arbitrarily sampled rotation matrix to the centers of P_*.

We use the Marr wavelet basis, setting $m_0 = 4.5$ and $G = 3$, which yields $L = 28$ from (B.10), and in turn 84 weight parameters w_{jl}.

The SA iterations for the outer optimization of w_{jl} were conducted with mini-batches of size 50. A gain sequence of $\gamma_r = 100/(100 + r)$ was used. For each empirical set P_n or P_n^{alt} (sampled from P_* or P_*^{alt}), we ran the SA algorithm 5 times to compute $\hat{R}_n$ or $\hat{R}_n^{\text{alt}}$, which are respectively defined by

$$\begin{aligned}
\hat{R}_n &:= \sup_{w_{jl}} \left[\mathbb{E}_{P_*}\left(\sum_{j=1}^{K}\sum_{l=1}^{L} w_{jl} b_l(\theta_j^\top X)\right)\right. \\
&\qquad \left. -\mathbb{E}_{P_n}\left(\sup_{z\in\mathbb{R}^K}\left(\sum_{j=i}^{K}\sum_{l=1}^{L} w_{jl} b_l(\theta_j^\top X + z^{(j)}) - z^\top (\Gamma_K)^{-1} z\right)\right)\right], \\
\hat{R}_n^{\text{alt}} &:= \sup_{w_{jl}} \left[\mathbb{E}_{P_*}\left(\sum_{j=1}^{K}\sum_{l=1}^{L} w_{jl} b_l(\theta_j^\top X)\right)\right. \\
&\qquad \left. -\mathbb{E}_{P_n^{\text{alt}}}\left(\sup_{z\in\mathbb{R}^K}\left(\sum_{j=i}^{K}\sum_{l=1}^{L} w_{jl} b_l(\theta_j^\top X + z^{(j)}) - z^\top (\Gamma_K)^{-1} z\right)\right)\right]
\end{aligned}$$

and took their averages. This procedure seeks to average out the noise experienced in the SA method due to the fixed batch size of 50 in estimating expectations under P_*.