THE UNIVERSITY OF CHICAGO

INFINITE-DIMENSIONAL INFERENCE AND LEARNING VIA OPTIMAL TRANSPORT

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
YOONHAENG HUR

CHICAGO, ILLINOIS
JUNE 2025

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisors, Tengyuan Liang and Yuehaw Khoo, for their guidance and support throughout my PhD journey. I am extremely fortunate to have been guided by their insights and expertise. I would like to thank Rina Foygel Barber, for not only being my committee member but also for providing wisdom whenever I faced challenges.

I would like to thank my collaborators, Wenxuan Guo and Chris Ryan, who I had the pleasure of working with on projects that are part of this thesis.

Thank you to my dear friends and colleagues—Omar Al-Ghattas, Subhodh Kotekal, and Kulunu Dharmakeerthi—for inspiring me with their talented insights and walking together through the ups and downs of the PhD journey.

My heartfelt gratitude goes to my parents and my sister, for their unwavering love and support. Their presence itself is a source of strength and motivation for me. For that, I will always be grateful.

ABSTRACT

This thesis studies new frameworks and methods for infinite-dimensional statistical inference and learning problems through the lens of optimal transport. Optimal transport is a branch of mathematics concerning how to transport one probability distribution to another while minimizing certain costs, which, as this thesis shows, can provide principled guidance for sophisticated modern data science problems. There are two overarching themes in this thesis: (1) the dynamic viewpoint of optimal transport that concerns the optimal path of transporting one probability distribution to another, and (2) quadratic extensions of the classical linear optimal transport problem. This thesis effectively utilizes these themes to delve into central problems in modern data science, including hypothesis testing, online learning, simulation-based inference, and missing value imputation. The results of this thesis contribute to the growing body of literature on optimal transport and its applications in statistics and machine learning, providing new insights and tools for researchers and practitioners in these fields.

CHAPTER 1

INTRODUCTION

Technological advances are continuously bringing new challenges to many scientific fields, calling for original ideas to understand and utilize data effectively. To develop intuitive and efficient algorithms to tackle such challenges, it is imperative to have a solid mathematical foundation that can help us understand complex systems through data-driven insights.

Optimal transport is a branch of mathematics concerning how to transport one probability distribution to another while minimizing certain costs, which dates back to Monge [1781]. At the intersection of multiple areas of mathematics, such as probability, geometry, and optimization, optimal transport can provide principled guidance for sophisticated modern data science problems.

This thesis presents new frameworks and methods for infinite-dimensional statistical inference and learning problems built on insights and principles from optimal transport. We first overview the mathematical foundations of optimal transport in Section 1.1. Then, Section 1.2 outlines the remaining chapters of the thesis and their contributions.

Notation We first introduce the notation that we use throughout the thesis. For a metric space $\mathcal{X}$, we denote its metric as $d_{\mathcal{X}}$ and write $\mathcal{P}(\mathcal{X})$ to denote the collection of all Borel probability measures on $\mathcal{X}$. We call a pair $(\mathcal{X}, \mu)$ a Polish probability space if $\mathcal{X}$ is a metric space that is complete and separable and $\mu \in \mathcal{P}(\mathcal{X})$. Given two Polish probability spaces $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$, we call a measurable map $T\colon \mathcal{X} \to \mathcal{Y}$ a transport map from μ to ν if $T_{\#}\mu = \nu$, where $T_{\#}\mu$ is the pushforward measure of μ under a measurable map $T\colon \mathcal{X} \to \mathcal{Y}$ defined by $T_{\#}\mu(B) = \mu\{x \in \mathcal{X} : T(x) \in B\}$ for any Borel set $B \subset \mathcal{Y}$, and the collection of all transport maps from μ to ν is denoted as $\mathcal{T}(\mu, \nu)$; we call $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ a coupling between μ and ν if $\gamma(A \times \mathcal{Y}) = \mu(A)$ and $\gamma(\mathcal{X} \times B) = \nu(B)$ for all Borel subsets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$, and we denote the collection of all such couplings as $\Pi(\mu, \nu)$.

1.1 A Brief Overview of Optimal Transport Theory

A major goal of optimal transport is minimizing the cost associated with the transport map between two Polish probability spaces, say $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$. Consider a measurable function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$; we view $c(x, y)$ as the cost associated with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. For each transport map $T \in \mathcal{T}(\mu, \nu)$, we interpret $c(x, T(x))$ as a unit cost incurred by mapping each $x \in \mathcal{X}$ to $T(x) \in \mathcal{Y}$. We define the average cost incurred by the transport map T as the integration of all the unit costs with respect to μ, that is, $\int_{\mathcal{X}} c(x, T(x)) \, \mathrm{d}\mu\,(x)$. Minimizing the cost over $\mathcal{T}(\mu, \nu)$ is referred to as the Monge problem [Monge, 1781], named after Gaspard Monge. We call $T^\star$ an optimal transport map if $T^\star$ is minimizer, that is,

$$T^\star \in \operatorname*{arg\,min}_{T \in \mathcal{T}(\mu,\nu)} \int_{\mathcal{X}} c(x, T(x)) \, \mathrm{d}\mu\,(x) \ ,$$

Another important optimal transport problem is minimizing the cost given by couplings. We define the average cost incurred by a coupling $\gamma \in \Pi(\mu, \nu)$ as the integration of the cost c with respect to γ, namely, $\int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\gamma\,(x, y)$. Minimizing this cost over $\Pi(\mu, \nu)$ is called the Kantorovich problem, credited to Leonid Kantorovich. We call $\gamma^\star$ an optimal coupling if

$$\gamma^\star \in \operatorname*{arg\,min}_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\gamma\,(x, y) \ ,$$

The two optimal transport problems are closely related: the Kantorovich problem is a relaxation of the Monge problem. To see this, for each $T \in \mathcal{T}(\mu, \nu)$, define a map $(\mathrm{Id}, T)\colon \mathcal{X} \to \mathcal{X} \times \mathcal{Y}$ by $(\mathrm{Id}, T)(x) = (x, T(x))$. One can verify $(\mathrm{Id}, T)_{\#}\mu \in \Pi(\mu, \nu)$. Therefore, if we define $\Pi_{\mathcal{T}} := \{(\mathrm{Id}, T)_{\#}\mu : T \in \mathcal{T}(\mu, \nu)\}$, then $\Pi_{\mathcal{T}} \subset \Pi(\mu, \nu)$ and thus

$$\inf_{T \in \mathcal{T}(\mu,\nu)} \int_{\mathcal{X}} c(x, T(x)) \, \mathrm{d}\mu\,(x) = \inf_{\gamma \in \Pi_{\mathcal{T}}} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\gamma\,(x, y) \geq \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\gamma\,(x, y) \ ,$$

where the first equality follows from change-of-variables. In other words, two optimal trans-

port problems share the same objective function as a function of couplings; however, the Kantorovich problem has a larger constraint set.

Unlike the Monge problem, the Kantorovich problem has favorable properties. First, the objective function is linear in γ. Moreover, $\Pi(\mu, \nu)$ is compact in the weak topology of Borel probability measures defined on $\mathcal{X} \times \mathcal{Y}$. This suggests that we can view the Kantorovich problem as an infinite-dimensional linear program.

Besides seeking optimal transport maps or couplings, another interesting aspect of optimal transport problems is that the least cost obtained from the Kantorovich problem can endow a metric structure among Polish probability spaces. For example, if $\mathcal{X} = \mathcal{Y}$ and $c = d_{\mathcal{X}}^2$, the square root of the solution of the Kantorovich problem defines a distance between μ and ν, known as the Wasserstein distance.

Definition 1.1.1. Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space that is complete and separable. For $p \geq 1$, we call

$$W_p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}^p(x, y) \, \mathrm{d}\gamma \, (x, y) \right)^{1/p}$$

the Wasserstein-p distance between $\mu, \nu \in \mathcal{P}(\mathcal{X})$.

1.2 Outline and Contributions

The rest of the thesis is organized into four chapters, each of which presents a new framework and method for infinite-dimensional statistical inference or learning problems using ideas and principles from optimal transport. There are two overarching themes: (1) the dynamic viewpoint of optimal transport that concerns the optimal path of transporting one probability distribution to another, and (2) quadratic extensions of the classical linear optimal transport problem, where the objective function is quadratic in the transport map or coupling.

Chapter 2 and Chapter 3 utilize the dynamic viewpoints of optimal transport. In Chapter 2, the notion of the optimal path is used to model weak distribution shifts and tackle a

relevant hypothesis testing problem, while in Chapter 3, this notion serves as a direction to update a decision variable in a new online learning problem.

Chapter 4 and Chapter 5 concerns different types of transport problems, where the objective function is quadratic in the transport map or coupling. Chapter 4 is built on the Gromov-Wasserstein distance, which is a non-convex quadratic type of transport problem between incomparable spaces. Chapter 5 studies a convex quadratic type of transport problem for missing value imputation and individualized inference.

The following overview summarizes the remaining chapters of the thesis and their contributions.

- Chapter 2 is based on Hur and Liang [2025]. Given two probability distributions P (null) and Q (alternative), we study the problem of detecting weak distribution shift deviating from the null P toward the alternative Q, where the level of deviation vanishes as a function of n, the sample size. We propose a model for weak distribution shifts via displacement interpolation between P and Q, drawing from the optimal transport theory. We study a hypothesis testing procedure based on the Wasserstein distance, derive sharp conditions under which detection is possible, and provide the exact characterization of the asymptotic Type I and Type II errors at the detection boundary using empirical processes. We demonstrate how the proposed testing procedure works in modeling and detecting weak distribution shifts in real data sets using two empirical examples: distribution shifts in consumer spending after COVID-19, and heterogeneity in the published p-values of statistical tests in journals across different disciplines.

- Chapter 3 is based on Guo et al. [2022], where we draw connections between online learning, optimal transport, and partial differential equations through an insight called the minimal selection principle, originally studied in the Wasserstein gradient flow setting by Ambrosio et al. [2005]. This allows us to extend the standard online learning framework to the infinite-dimensional setting seamlessly. Based on our framework, we

derive a novel method called the minimal selection or exploration (MSoE) algorithm to solve OLT problems using mean-field approximation and discretization techniques. In the displacement convex setting, the main theoretical message underpinning our approach is that minimizing transport cost over time (via the minimal selection principle) ensures optimal cumulative regret upper bounds. On the algorithmic side, our MSoE algorithm applies beyond the displacement convex setting, making the mathematical theory of optimal transport practically relevant to non-convex settings common in dynamic resource allocation.

- Chapter 4 is based on Hur et al. [2024], which introduces a new simulation-based inference procedure to model and sample from multi-dimensional probability distributions given access to i.i.d. samples, circumventing the usual approaches of explicitly modeling the density function or designing Markov chain Monte Carlo. Motivated by the seminal work on distance and isomorphism between metric measure spaces, we develop a new transform sampler to perform simulation-based inference, which estimates a notion of optimal alignments between two heterogeneous metric measure spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ from empirical data sets, with estimated maps that approximately push forward one measure μ to the other ν, and vice versa. We introduce a new notion called the Reversible Gromov-Monge (RGM) distance, providing mathematical formalism behind the new sampler. We study the statistical rate of convergence of the new transform sampler, along with several analytic properties of the RGM distance and operator viewpoints of transform sampling. Synthetic and real-world examples showcasing the effectiveness of the new sampler are also demonstrated.

- Chapter 5 is based on Hur and Liang [2024], where we introduce a new convexified matching method for missing value imputation and individualized inference inspired by computational optimal transport. Our method integrates features from mainstream imputation approaches: optimal matching, regression imputation, and synthetic con-

trol. We impute counterfactual outcomes based on convex combinations of observed outcomes, defined based on an optimal coupling between the treated and control data sets. The optimal coupling problem is considered a convex relaxation to the combinatorial optimal matching problem. We estimate granular-level individual treatment effects while maintaining a desirable aggregate-level summary by properly constraining the coupling. We construct individual confidence intervals for the estimated counterfactual outcomes. We devise fast iterative entropic-regularized algorithms to solve the optimal coupling problem that scales favorably when the number of units to match is large. Entropic regularization plays a crucial role in both inference and computation; it helps control the width of the individual confidence intervals and design fast optimization algorithms.

Additional Notation We use the following notation throughout the rest of the thesis. Let $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the Euclidean norm and inner product, respectively. For a square matrix A, let $\mathrm{tr}(A)$ denote its trace. For matrices A, B of the same dimension, $\langle A, B\rangle := \mathrm{tr}(A^\top B)$ is the Frobenius inner product; $\|A\| := \sqrt{\langle A, A\rangle}$ denotes the Frobenius norm of a matrix A. For any integer $n \in \mathbb{N}$, we define $[n] = \{1, \ldots, n\}$ and let $1_n = (1, \ldots, 1) \in \mathbb{R}^n$ denote the vector whose entries are all 1. Let $\Delta_n := \{a \in \mathbb{R}^n_+ : \sum_{i=1}^n a_i = 1\}$ denote the simplex of probability vectors and let $\Delta_n^+ := \{a \in \Delta_n : a_i > 0 \ \ \forall i\}$ denote the interior of Δ_n. For $a, b \in \mathbb{R}$, denote $a \wedge b = \min\{a, b\}$.

CHAPTER 2

DETECTING WEAK DISTRIBUTION SHIFTS VIA DISPLACEMENT INTERPOLATION

2.1 Introduction

Classic detection problem aims to distinguish a shifted distribution Q from the null distribution P based on data, formulated as a nonparametric goodness-of-fit test:

$$H_0 : X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P \quad \text{vs.} \quad H_1 : X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Q. \tag{2.1.1}$$

It is understood that the power of certain test statistics, such as Kolmogorov-Smirnov [Kolmogorov, 1933, Smirnov, 1948] or Anderson-Darling [Anderson and Darling, 1952], is asymptotically 1, implying that detection is possible if the sample size is large.

In many applications, one confronts situations where the signal in the alternative distribution is weaker. One natural formulation is to replace Q in H_1 (2.1.1) by a suitable interpolation scheme between P and Q. A common choice is the linear interpolation $(1-\varepsilon)P + \varepsilon Q$, or the so-called Huber's ε-contamination model [Huber, 1964], where one parametrizes $\varepsilon = \varepsilon_n \to 0$ as $n \to \infty$ to represent weak signals. This model represents that only a small ε-fraction of the data deviates from P, often seen in applications such as large-scale inference with high-throughput measurements [Efron, 2010]. For instance, microarray data—widely used in genetics and genomics—measure the expression level of thousands of genes, where only some portions are relevant to detecting certain diseases. See Donoho and Jin [2015] for a comprehensive overview of such applications. In those applications, weak signals are modeled as small perturbations in the frequencies of the histogram of the data, that is, deviations along the y-axis.

Detecting the presence of distribution shifts is also essential in social and economic appli-

cations. There, the main interest is often to quantify individual, heterogeneous responses to policies or other incentives. For instance, consider the economic impacts of the coronavirus pandemic (COVID-19) and government policies [Chetty et al., 2024]; when analyzing how the economy—measured by weekly statistics such as consumer spending—recovers from the shock caused by the pandemic, it is natural to consider individual, heterogeneous shifts to reflect that each individual adjusts the spending differently according to the income level and other characteristics. In this context, the individual responses can be modeled as perturbations along the x-axis of the data histogram, whereas the aforementioned perturbation along the y-axis (frequencies) rooted in engineering applications is arguably less informative. Another example is the study of the effect of financial or non-financial incentives given to students or parents to improve educational performance, measured by test scores [Fryer et al., 2015, Levitt et al., 2016]. There, the main interest is measuring the shifts in test score distributions in response to the incentives. It is worth noting that the shifts are arguably heterogeneous, depending on gender, race/ethnicity, and other demographic information [Levitt et al., 2016].

In the above examples, perturbations along the y-axis—the fraction of individuals deemed responsive to a given policy or incentive—may be of interest but are insufficient to model individual, heterogeneous responses. Unlike the earlier genetics example, where the identification of a handful of non-null genes showing distinct expression levels provides crucial scientific evidence of certain diseases, the goal of socioeconomic research is beyond simply finding a proportion of shifted observations; the fundamental interest is to quantify individual, heterogeneous responses to policy, which is more relevant to the shift along the x-axis of histograms.

This chapter proposes a different interpolation scheme for detecting weak distribution shifts motivated by the above. We study a natural test statistic motivated by optimal transport and conduct an exact study of the asymptotic power of the test. We represent

the signal strength as the level of deviation, rather than the proportion of deviated units. In a nutshell, we are interested in how much the data histogram shifts along the x-axis instead of the y-axis. To this end, we model weak signals using displacement interpolation [McCann, 1997] with a viewpoint from optimal transport theory [Villani, 2003], where the interpolation parameter ε represents a certain transport distance of data from the null P along the geodesics, thus characterizing the level of distribution shifts from P.

2.1.1 Displacement Interpolation and Optimal Transport

Let P and Q be Borel probability measures on $\mathbb{R}$ whose cumulative distribution functions are F and G, respectively; also, let F^{-1} and G^{-1} be their quantile functions, respectively. Displacement interpolation between P and Q is defined as

$$P_\varepsilon := ((1-\varepsilon)\mathrm{Id} + \varepsilon T)_{\#}P \quad \forall \varepsilon \in [0,1],$$

where $\mathrm{Id}\colon \mathbb{R} \to \mathbb{R}$ is the identity map and $T = G^{-1} \circ F$, namely, the composition of G^{-1} and F.[1] Recall that $S_{\#}P$ denotes the pushforward measure of P by a map $S\colon \mathbb{R} \to \mathbb{R}$ defined by $S_{\#}P(B) = P\{x \in \mathbb{R} : S(x) \in B\}$ for any Borel set $B \subset \mathbb{R}$. If P is absolutely continuous with respect to the Lebesgue measure, it is known that $T = G^{-1} \circ F$ is the unique monotone map such that $T_{\#}P = Q$. More importantly, it is also the unique solution to the following optimal transport problem [Brenier, 1991, Villani, 2003] provided P and Q have finite second moments, namely,

$$T = \operatorname*{arg\,min}_{\substack{S\colon \mathbb{R}\to\mathbb{R} \\ S_{\#}P = Q}} \int_{\mathbb{R}} |x - S(x)|^2 \, \mathrm{d}P(x)\,. \tag{2.1.2}$$

Intuitively, the constraint $S_{\#}P = Q$ means that a map S transports mass from the source distribution P to the target distribution Q by moving the infinitesimal mass at x in the support of P to $S(x)$ such that $\mathrm{d}P(x) = \mathrm{d}Q(S(x))$. Accordingly, (2.1.2) tells that the

1. It is possible to generalize displacement interpolation to $\mathbb{R}^d$ with $d > 1$; see Definition 3.2.3.

unique monotone map $T = G^{-1} \circ F$ provides the optimal way of transporting mass from P to Q, minimizing the squared transport distance $|x - T(x)|^2$ on average.

Now, for each x, let us view the segment $\{(1-\varepsilon)x + \varepsilon T(x) : \varepsilon \in [0,1]\}$ as the transport path from x to the destination $T(x)$ along its displacement $T(x) - x$. It turns out that transporting mass from the source distribution P along such a path gives rise to a geodesic connecting P and Q in the Wasserstein space [Ambrosio et al., 2005], namely, $\varepsilon = \frac{W_p(P,P_\varepsilon)}{W_p(P,Q)}$ for all $\varepsilon \in [0,1]$, where we note that the Wasserstein-p distance between probability measures on $\mathbb{R}$ can be expressed in terms of the quantile functions: for any probability measures μ, ν on $\mathbb{R}$ whose quantile functions are F_μ^{-1}, F_ν^{-1}, respectively, we have

$$W_p(\mu,\nu) := \left(\int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^p \, \mathrm{d}u \right)^{1/p}.$$

Observe that the quantile function of P_ε is $(1-\varepsilon)F^{-1} + \varepsilon G^{-1}$ as $P = (F^{-1})_{\#}U$ implies $P_\varepsilon = ((1-\varepsilon)\mathrm{Id} + \varepsilon G^{-1} \circ F)_{\#}((F^{-1})_{\#}U) = ((1-\varepsilon)F^{-1} + \varepsilon G^{-1})_{\#}U$, where U is the Lebesgue measure on $[0,1]$; see Figure 2.1 for details. Therefore, displacement interpolation represents the optimal path—shortest path under the distance W_p—for transporting mass from P to Q, where the parameter ε naturally characterizes the deviation from P via the relative distance $\varepsilon = \frac{W_p(P,P_\varepsilon)}{W_p(P,Q)}$.

We finish this section with a concrete example to contrast two interpolation schemes. Consider $P = N(0,1)$ and $Q = N(\tau, \sigma^2)$ for some $\tau, \sigma > 0$. Linear interpolation $(1-\varepsilon)P + \varepsilon Q$ is the mixture of two Gaussian distributions P, Q, where the parameter ε is essentially the frequency of signals from Q. Displacement interpolation is the optimal transport path from P to Q, where $T(x) = \sigma x + \tau$ for all $x \in \mathbb{R}$ and $P_\varepsilon := ((1-\varepsilon)\mathrm{Id} + \varepsilon T)_{\#}P = N(\varepsilon\tau, (1-\varepsilon+\varepsilon\sigma)^2)$; in words, the optimal path is simply to move each x to $\sigma x + \tau$. The resulting interpolation $N(\varepsilon\tau, (1-\varepsilon+\varepsilon\sigma)^2)$ suggests that the parameter ε provides an intuitive characterization of the signal strength which controls the level of distribution shift. Lastly, it is worth noting that displacement interpolation preserves the unimodality as the interpolation is always Gaussian,

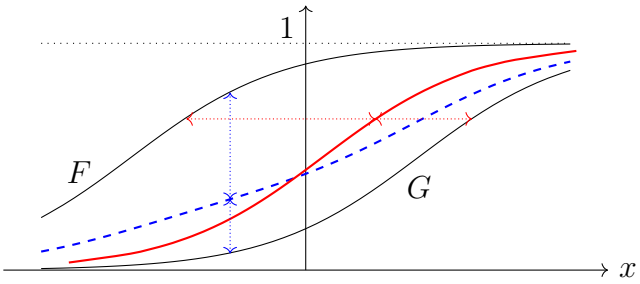


Figure 2.1: Illustration of two interpolation schemes. Linear interpolation $(1-\varepsilon)P+\varepsilon Q$ vertically combines the cumulative distribution functions F and G; namely, its cumulative distribution function (blue, dashed) is $(1-\varepsilon)F+\varepsilon G$. Meanwhile, displacement interpolation (red, solid) horizontally combines F and G, or equivalently, vertically combines the quantile functions F^{-1} and G^{-1}. In other words, the quantile function of $((1-\varepsilon)\mathrm{Id}+\varepsilon G^{-1}\circ F)_{\#}P$ is $(1-\varepsilon)F^{-1}+\varepsilon G^{-1}$.

whereas linear interpolation may result in two modes.

2.1.2 Distribution Shift as Displacement Interpolation

As discussed in the previous section, displacement interpolation constructs an optimal path transporting mass from P to Q, providing a natural notion of distribution shift from P to Q. Such a viewpoint has recently found several applications in machine learning and computer vision, where displacement interpolation serves as a method to optimally synthesize two objects, such as textures [Rabin et al., 2012], colors [Kolouri et al., 2017], and styles [Mroueh, 2020].

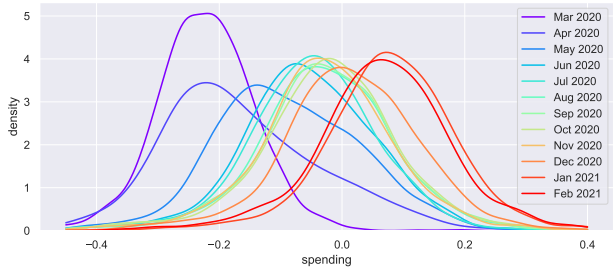
Before diving into theory and methodology, in this section, we adopt this viewpoint to model distribution shifts using real-world data; we see that displacement interpolation is better suited to model such distribution shifts than linear interpolation. Later, after laying out the theory, we revisit this empirical example to study the power of our testing procedure. To this end, we utilize the data from [Chetty et al., 2024], which studies the economic impacts of the coronavirus pandemic (COVID-19) and policy responses in the United States using a wide range of statistics, such as consumer spending, business revenues, employment

rates, and so on.[2] Here, we focus on the consumer spending data—recorded as seasonally adjusted percent changes based on anonymized card transactions data—and analyze how consumer expenditures have recovered from the steep plunge caused by COVID-19. Figure 2.2(a) shows the average consumer spending between January 2021 and June 2022. More specifically, we look at the distribution of monthly average consumer spending over 1,655 counties, where the spending is disaggregated based on the ZIP code where the cardholder lives. The monthly average is calculated by averaging the daily expenditures from the 16th of each month to the 15th of the following month. Figure 2.2(b) shows the smoothed histograms of monthly average spending for the first 12 months since March 2020, which clearly shows how the distribution has shifted in the increasing spending direction, namely, the spending is recovering from the shock of the pandemic. Meanwhile, Figure 2.2(c) shows the next 12 months from March 2021, where we can no longer observe such an evident distribution shift, suggesting the consumer spending has stabilized after the recovery; see also the corresponding period in Figure 2.2(a).

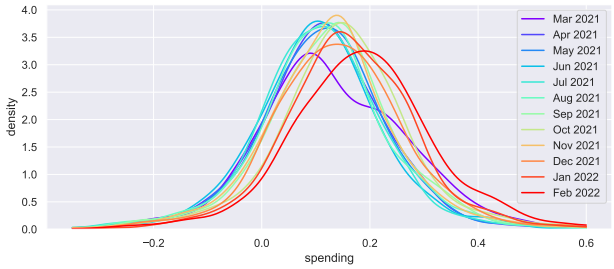
Empirically, we illustrate that displacement interpolation serves as a reasonable model for the distribution shift during the recovery period shown in Figure 2.2(b). To this end, we generate two interpolation paths, displacement interpolation and linear interpolation, and contrast them with the real data. First, let $\{P_{i/11}\}_{i=0}^{11}$ be the distributions of monthly spending by county during that period, namely, they correspond to the 12 histograms of Figure 2.2(b); here, P_0 and P_1 are the start and end of that period (from March 16, 2020 to April 15, 2020 and from February 16, 2021 to March 15, 2021, respectively). Then, we compute the relative Wasserstein-2 distances $\varepsilon_t = \frac{W_2(P_0,P_t)}{W_2(P_0,P_1)}$ and the relative Total Variation (TV) distances $\gamma_t = \frac{TV(P_0,P_t)}{TV(P_0,P_1)}$ for $t \in \{0, 1/11, \ldots, 10/11, 1\}$, as visualized in Figure 2.3(a). From these, we generate displacement interpolation $Q_t^{\text{dis}} = ((1-\varepsilon_t)\text{Id} + \varepsilon_t T)_{\#} P_0$, where T is the composition of the quantile function of P_1 and the cumulative distribution function of P_0;

2. Data are publicly available at `https://tracktherecovery.org/`.

(a) Daily average consumer spending



(b) Monthly spending: March 2020 + 1 year

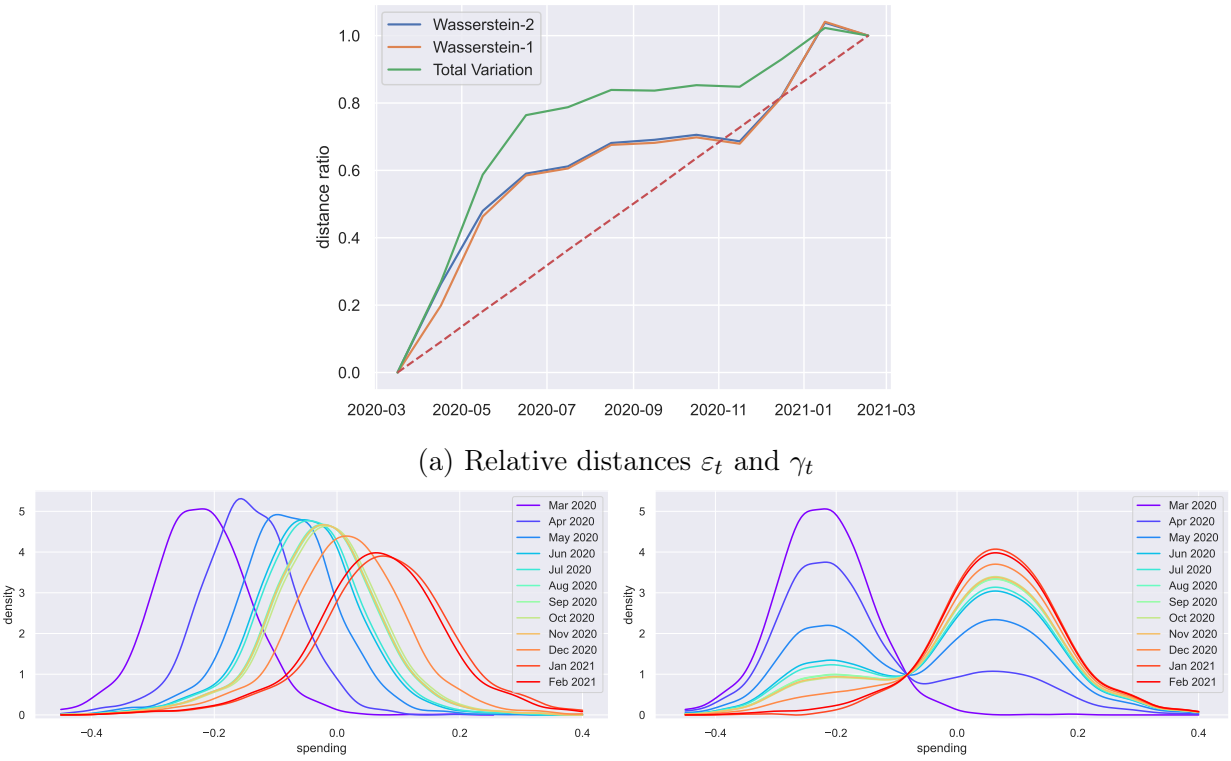


(c) Monthly spending: March 2021 + 1 year

Figure 2.2: (a) plots a time series of consumer spending from January 13, 2020 to June 5, 2022, where some noticeable events are marked: March 13, 2020 (national emergency declared) and April 15, 2020, January 4, 2021, and March 17, 2021 (first, second, and third stimulus payments start, respectively). (b) and (c) show the smoothed histograms of monthly average consumer spending by county from March 16, 2020 to March 15, 2021 and from March 16, 2021 to March 15, 2022, respectively.

similarly, we generate linear interpolation $Q_t^{\text{lin}} = (1-\gamma_t)P_0+\gamma_t P_1$. Essentially, Q_t^{dis} amounts to displacement interpolation between P_0 and P_1 that shifts from P_0 at the same rate as P_t under the Wasserstein-2 distance, namely, $W_2(P_0, Q_t^{\text{dis}}) = W_2(P_0, P_t)$; analogously, Q_t^{lin} corresponds to linear interpolation such that $TV(P_0, Q_t^{\text{lin}}) = TV(P_0, P_t)$. Figure 2.3(b) and Figure 2.3(c) show Q_t^{dis} and Q_t^{lin}, respectively. Comparing Figure 2.3(b) and Figure 2.3(c) with the real distribution shift in Figure 2.2(b), we can see that displacement interpolation provides a better approximation. Particularly, displacement interpolation preserves the unimodality of distributions as in Figure 2.2(b), while linear interpolation creates two modes, in reminiscence of the Gaussian example in the previous section.

Together with theoretical foundations in Section 2.1.1, the above example provides the empirical underpinnings of the displacement interpolation model for distribution shifts. Returning to the detection problem (2.1.1), in what follows, we will consider a detection problem where Q in H_1 is replaced by displacement interpolation $((1-\varepsilon)\mathrm{Id} + \varepsilon T)_{\#}P$ and propose a testing procedure based on the Wasserstein-2 distance. Later, we will apply the proposed testing procedure to revisit the above empirical example and analyze the power under the strong distribution shift during the recovery period.



(a) Relative distances ε_t and γ_t

(b) Displacement interpolation Q_t^{dis}

(c) Linear interpolation Q_t^{lin}

Figure 2.3: (a) plots the relative Wasserstein-2 distances $\varepsilon_t = \frac{W_2(P_0,P_t)}{W_2(P_0,P_1)}$ and the relative TV distances $\gamma_t = \frac{TV(P_0,P_t)}{TV(P_0,P_1)}$ for $t \in \{0, 1/11, \ldots, 1\}$, with the 45 degree line shown as a dashed line; relative Wasserstein-1 distances $\frac{W_1(P_0,P_t)}{W_1(P_0,P_1)}$ are plotted for reference as well, which almost coincide with ε_t. (b) and (c) show displacement interpolation $Q_t^{\mathrm{dis}} = ((1-\varepsilon_t)\mathrm{Id} + \varepsilon_t T)_{\#}P_0$ and linear interpolation $Q_t^{\mathrm{lin}} = (1-\gamma_t)P_0 + \gamma_t P_1$, respectively.

2.1.3 Problem Description

Motivated by the discussion above, we study the problem of detecting weak distribution shifts represented as displacement interpolation: given two distributions P and Q on $\mathbb{R}$ whose cumulative distribution functions are F and G, respectively,

$$H_0 : X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P \quad \text{vs.} \quad H_1 : X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} ((1-\varepsilon)\mathrm{Id} + \varepsilon G^{-1} \circ F)_{\#} P. \tag{2.1.3}$$

Here, F^{-1} and G^{-1} are the quantile functions of P and Q, respectively.

We propose a testing procedure based on the weighted Wasserstein distance between the empirical measure P_n—constructed by the observations $X_1, \dots, X_n$—and the null distribution P. Assuming a weak signal in a sense $\varepsilon = \varepsilon_n \to 0$ as $n \to \infty$, we derive sharp conditions under which detection is possible: (1) when $n^{1/2}\varepsilon_n \to 0$, detection is impossible; (2) when $n^{1/2}\varepsilon_n \to \infty$, the testing procedure has asymptotic power 1; (3) at the detection boundary $n^{1/2}\varepsilon_n \to$ constant $\in (0, \infty)$, sharp asymptotic Type I and Type II errors are analyzed using Gaussian processes.

2.1.4 Related Literature

Sparse Mixture Detection A popular approach to formulating the weak signal detection problem is to replace the alternative hypothesis of (2.1.1) with Huber's ε-contamination model $(1-\varepsilon)P + \varepsilon Q$, also known as sparse mixtures, where $\varepsilon = \varepsilon_n \to 0$ as $n \to \infty$. Donoho and Jin [2004] proposes Tukey's higher criticism as a test statistic and analyzes the asymptotic phase transition depending on the rate $\varepsilon = \varepsilon_n \to 0$ and the signal strength. A recent endeavor extending the higher criticism test to compare two large frequency tables is given in Donoho and Kipnis [2022]. See also Cai et al. [2011], Cai and Wu [2014] on optimal detection of general sparse mixtures.

Wasserstein Distances for Testing Our testing procedure is based on the Wasserstein distance, which has been studied extensively in the testing literature. For example, Munk and Czado [1998] and Del Barrio et al. [2005] study Goodness-of-Fit (GoF) testing using the Wasserstein distance between the null hypothesis and the empirical measure based on the observations; see Hallin et al. [2021] for an extension to the multivariate case. As mentioned earlier, the standard GoF testing is related to the classic detection problem (2.1.1), where the alternative distribution is fixed to some signal Q or a collection of distributions from a specific parametric family [de Wet, 2002, Csörgő, 2003].

Notation We use the following notation throughout the chapter. For positive sequences $\{a_n\}$ and $\{b_n\}$, denote $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. We write $Y_n \rightsquigarrow Y$ if a sequence $\{Y_n\}$ of random variables converges weakly to some random variable Y.

2.2 Main Results

Testing Procedure We propose a distance-based test statistic for the testing problem (2.1.3), motivated by optimal transport and displacement interpolation. More specifically, we compare P_n—the empirical measure based on the observations $X_1, \ldots, X_n$—with P under the following distance, called the weighted Wasserstein distance, and reject H_0 if it is larger than a specific critical value.

Definition 2.2.1. Let ω be a finite Borel measure on $(0, 1)$. For two distributions μ and ν on $\mathbb{R}$, we define the ω-weighted Wasserstein distance by

$$W_{2,\omega}(\mu, \nu) := \left(\int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^2 \, \mathrm{d}\omega(u) \right)^{1/2},$$

where F_μ^{-1}, F_ν^{-1} denote the quantile functions of μ, ν, respectively.

Remark 2.2.2. Note in Definition 2.2.1 that $W_{2,\omega} = W_2$ if ω is the Lebesgue measure.

Weighted versions of the Wasserstein distance are introduced in [de Wet, 2002, Csörgő, 2003]. Here, we introduce the weighted base measure as in the Anderson-Darling test because, for certain detection problems, the signal may hide unevenly among quantiles. For example, the signal is contained in the extreme quantiles in higher criticism [Donoho and Jin, 2004].

Asymptotic Phase Transition We analyze the testing error in the asymptotic regime where $\varepsilon = \varepsilon_n$ vanishes as $n \to \infty$. More precisely, we rewrite (2.1.3) as follows specifying dependency on the sample size n:

$$\begin{aligned} H_0^{(n)} &: X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P, \\ H_1^{(n)} &: X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} ((1-\varepsilon_n)\mathrm{Id} + \varepsilon_n G^{-1} \circ F)_{\#} P, \end{aligned}$$

where $\lim_{n\to\infty} \varepsilon_n = 0$, namely, $\varepsilon_n = o(1)$. Our main result shows that the asymptotic testing error behaves differently depending on the vanishing rate of ε_n, highlighting a phase transition phenomenon. In particular, we study a testing procedure such that the asymptotic Type I error is a given level $\alpha \in (0,1)$ by deriving the limit of the test statistic $W_{2,\omega}(P_n, P)$ under the null hypothesis, based on the fundamental limit theorem of the empirical quantile process: assuming F admits a positive density f,

$$\sqrt{n}(F_n^{-1}(u) - F^{-1}(u)) \rightsquigarrow \frac{\mathbf{B}_u}{f(F^{-1}(u))}, \tag{2.2.1}$$

where F_n^{-1} is the empirical quantile function based on $X_1, \dots, X_n$ from $H_0^{(n)}$ and $(\mathbf{B}_u)_{u\in[0,1]}$ is the standard Brownian bridge, namely, a mean-zero Gaussian process whose covariance satisfies $\mathbb{E}[\mathbf{B}_{u_1}\mathbf{B}_{u_2}] = u_1 \wedge u_2 - u_1 u_2$; rigorous asymptotic analysis is provided in the subsequent section. For such a testing procedure, we can characterize the asymptotic Type II

error based on the three phases of ε_n determined by

$$\lim_{n\to\infty} n^{1/2}\varepsilon_n = \begin{cases} 0, \\ \infty, \\ \gamma \in (0,\infty). \end{cases}$$

We formally state the main result as follows.

Theorem 2.2.3. *Suppose F has a density f that is continuous and bounded away from 0 on some compact interval I_F and is 0 on $\mathbb{R}\backslash I_F$, G^{-1} is bounded on $(0,1)$, and $G^{-1}\circ F$ is Lipschitz. Let ω be a finite Borel measure on $(0,1)$ that is absolutely continuous with respect to the Lebesgue measure such that $W_{2,\omega}(P,Q)\neq 0$. Also, let Ψ be the cumulative distribution function of*

$$\int_0^1 \left|\frac{\mathbf{B}_u}{f(F^{-1}(u))}\right|^2 \mathrm{d}\omega(u)\,, \tag{2.2.2}$$

where $(\mathbf{B}_u)_{u\in[0,1]}$ is the standard Brownian bridge. Fix $\alpha\in(0,1)$ and let C_α be the $(1-\alpha)$-th quantile of Ψ, that is, $\Psi(C_\alpha)=1-\alpha$. Consider the following testing procedure:

$$\textit{reject } H_0^{(n)} \textit{ if and only if } \; nW_{2,\omega}^2(P_n,P) > C_\alpha.$$

Then, the asymptotic Type I error is α, and the asymptotic Type II error is as follows.

(i) If $n^{1/2}\varepsilon_n\to 0$, the asymptotic Type II error is $1-\alpha$.

(ii) If $n^{1/2}\varepsilon_n\to\infty$, the asymptotic Type II error is 0.

(iii) If $n^{1/2}\varepsilon_n$ is a constant, say $n^{1/2}\varepsilon_n=\gamma>0$, the asymptotic Type II error is

$$\Psi_\gamma(C_\alpha-\gamma^2 W_{2,\omega}^2(P,Q)),$$

where Ψ_γ is the cumulative distribution function of

$$\int_0^1 \left| \frac{\mathbf{B}_u}{f(F^{-1}(u))} \right|^2 \mathrm{d}\omega(u) + 2\gamma \int_0^1 \frac{\mathbf{B}_u}{f(F^{-1}(u))} \cdot \Big(G^{-1}(u) - F^{-1}(u) \Big) \, \mathrm{d}\omega(u) \,. \quad (2.2.3)$$

Remark 2.2.4. Theorem 2.2.3 implies that the asymptotic testing error, namely, the sum of the asymptotic Type I and II errors, is 1 (undetectable) if $n^{1/2}\varepsilon_n \to 0$ and α (detectable) if $n^{1/2}\varepsilon_n \to \infty$. The phase transition occurs at the boundary if $n^{1/2}\varepsilon_n$ is a constant, where the testing error is determined by the constant $\gamma := n^{1/2}\varepsilon_n$ and $\Delta := W_{2,\omega}(P, Q)$, which denotes the signal strength. The detection boundary (iii) still holds if we replace $n^{1/2}\varepsilon_n = \gamma$ with $\lim_{n\to\infty} n^{1/2}\varepsilon_n = \gamma > 0$; see the formal proof of Theorem 2.2.3 which is deferred to Section 2.6. The main ideas of the analysis will be presented in the next section.

2.3 Asymptotic Analysis

In this section, we rigorously derive the asymptotic limit of the test statistic $W_{2,\omega}(P_n, P)$, under the null $H_0^{(n)}$ and the alternative $H_1^{(n)}$, respectively.

2.3.1 Preliminaries

Let $\ell^\infty(0,1)$ denote the set of all bounded functions defined on $(0,1)$, which is a Banach space under the uniform norm given by

$$\|h\|_\infty = \sup_{u\in(0,1)} |h(u)| \quad \forall h \in \ell^\infty(0,1).$$

We first provide the exact statement of (2.2.1) based on weak convergence in $\ell^\infty(0,1)$, see van der Vaart and Wellner [1996]; in what follows, for any random element Z in $\ell^\infty(0,1)$, let $Z(u)$ denote the random variable at coordinate $u \in (0,1)$.

Assumption 2.3.1. F has a density f that is continuous and bounded away from 0 on some compact interval I_F and is 0 on $\mathbb{R}\backslash I_F$.

Lemma 2.3.2. *Let F_n^{-1} be the empirical quantile function based on $X_1, \ldots, X_n$ that are i.i.d. from a cumulative distribution function F. Under Assumption 2.3.1, view $(\sqrt{n}(F_n^{-1} - F^{-1}))_{n\in\mathbb{N}}$ as a sequence of random elements in $\ell^\infty(0,1)$, then it converges weakly to some tight measurable random element $\mathbf{H}$ in $\ell^\infty(0,1)$, which we denote as*

$$\sqrt{n}(F_n^{-1} - F^{-1}) \rightsquigarrow \mathbf{H} \quad in \quad \ell^\infty(0,1). \tag{2.3.1}$$

The limit $\mathbf{H}$ satisfies the following.

(i) $\{\mathbf{H}(u) : u \in (0,1)\}$ is a mean-zero Gaussian process with covariance function

$$\mathbb{E}[\mathbf{H}(u_1)\mathbf{H}(u_2)] = \frac{u_1 \wedge u_2 - u_1 u_2}{f(F^{-1}(u_1))f(F^{-1}(u_2))} \quad \forall u_1, u_2 \in (0,1).$$

(ii) The sample path $u \mapsto \mathbf{H}(u)$ is continuous.

Remark 2.3.3. Lemma 1 is from Lemma 3.9.23 of [van der Vaart and Wellner, 1996]. Assumption 2.3.1 ensures that F^{-1} is bounded on $(0,1)$, thereby viewing $\sqrt{n}(F_n^{-1} - F^{-1})$ as a random element in $\ell^\infty(0,1)$. Though this assumption rules out distributions supported on the whole real line, such as normal distributions, we can still apply Assumption 2.3.1 to such distributions by restricting their support to a sufficiently large yet bounded interval. Alternatively, one may modify Lemma 2.3.2 by considering weak convergence in $\ell^\infty[\delta, 1-\delta]$ with a suitable $\delta > 0$. Such a modification provides limit theorems of the integration of $|F_n^{-1} - F^{-1}|^2$ on $[\delta, 1-\delta]$, often called the trimmed Wasserstein distance [Munk and Czado, 1998]. It is also possible to modify Lemma 2.3.2 by considering weak convergence in $L^2(0,1)$ under suitable assumptions on the behavior of F^{-1} near the endpoints 0 and 1 [Del Barrio et al., 2005].

Remark 2.3.4. In Lemma 2.3.2, the limit $\mathbf{H}$ is tight, meaning that for any $\varepsilon > 0$, we can find a compact subset K of $\ell^\infty(0,1)$ such that $\mathbb{P}(\mathbf{H} \in K) \geq 1 - \varepsilon$. Though this technicality is not used explicitly in the main analysis, it is required to apply the extended continuous mapping theorem, which we adapt from Theorem 1.11.1 of van der Vaart and Wellner [1996] and rewrite as Theorem 2.6.1.

Next, we consider the following integrated processes: letting $\mathbf{H}_n := \sqrt{n}(F_n^{-1} - F^{-1})$, define

$$A_n := \int_0^1 |\mathbf{H}_n(u)|^2 \, \mathrm{d}\omega(u) \,, \tag{2.3.2}$$

$$B_n := \int_0^1 |(G^{-1} \circ F)(F^{-1}(u) + n^{-1/2}\mathbf{H}_n(u)) - F^{-1}(u)|^2 \, \mathrm{d}\omega(u) \,, \tag{2.3.3}$$

$$C_n := \int_0^1 \mathbf{H}_n(u) \cdot \Big((G^{-1} \circ F)(F^{-1}(u) + n^{-1/2}\mathbf{H}_n(u)) - F^{-1}(u)\Big) \, \mathrm{d}\omega(u) \,. \tag{2.3.4}$$

Under the assumptions of Lemma 2.3.2, the limit of A_n is essentially the limit of the test statistic $nW_{2,\omega}^2(P_n, P)$ under the null. Later, we will use B_n and C_n to derive the limit of the test statistic under the alternatives. The following lemma derives the limits of the above processes, which we prove in Section 2.6.

Lemma 2.3.5. *Let F_n^{-1} be the empirical quantile function based on $X_1, \ldots, X_n$ that are i.i.d. from a cumulative distribution function F. Let ω be a finite Borel measure on $(0,1)$ that is absolutely continuous with respect to the Lebesgue measure. Under Assumption 2.3.1 and assuming G^{-1} is bounded on $(0,1)$, let $\mathbf{H}_n = \sqrt{n}(F_n^{-1} - F^{-1})$ and $\mathbf{H}$ be the random*

element mentioned in Lemma 2.3.2, then

$$A_n \rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \, \mathrm{d}\omega(u) \,, \tag{2.3.5}$$

$$B_n \rightsquigarrow \int_0^1 |G^{-1}(u) - F^{-1}(u)|^2 \, \mathrm{d}\omega(u) \,, \tag{2.3.6}$$

$$C_n \rightsquigarrow \int_0^1 \mathbf{H}(u) \cdot (G^{-1}(u) - F^{-1}(u)) \, \mathrm{d}\omega(u) \,, \tag{2.3.7}$$

where A_n, B_n, C_n are as in (2.3.2)*,* (2.3.3)*,* (2.3.4)*, respectively.*

2.3.2 Asymptotic Distributions

Now, we analyze the limit of the test statistic $W_{2,\omega}(P_n, P)$ under $H_0^{(n)}$. By (2.3.5) of Lemma 2.3.5, the following holds.

Proposition 2.3.6 (Limit under the null)**.** *For each $n \in \mathbb{N}$, let P_n be the empirical measure based on $X_1, \ldots, X_n$ following $H_0^{(n)}$. Let ω be a finite Borel measure on $(0,1)$ that is absolutely continuous with respect to the Lebesgue measure. Under Assumption 2.3.1,*

$$nW_{2,\omega}^2(P_n, P) \rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \, \mathrm{d}\omega(u) \,,$$

where $\mathbf{H}$ is the random element mentioned in Lemma 2.3.2.

Next, we analyze the limit of the test statistic under $H_1^{(n)}$.

Theorem 2.3.7 (Limit under the alternatives)**.** *For each $n \in \mathbb{N}$, let P_n be the empirical measure based on $X_1, \ldots, X_n$ following $H_1^{(n)}$. Let ω be a finite Borel measure on $(0,1)$ that is absolutely continuous with respect to the Lebesgue measure. Under Assumption 2.3.1 and assuming G^{-1} is bounded on $(0,1)$ and $G^{-1} \circ F$ is Lipschitz, we can characterize the limit of $nW_{2,\omega}^2(P_n, P)$ as follows.*

(i) If $n^{1/2}\varepsilon_n \to 0$,

$$nW_{2,\omega}^2(P_n, P) \rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \,\mathrm{d}\omega(u)\,. \tag{2.3.8}$$

(ii) If $n^{1/2}\varepsilon_n \to \infty$,

$$n^{1/2}\varepsilon_n^{-1}\left(W_{2,\omega}^2(P_n, P) - \varepsilon_n^2 W_{2,\omega}^2(P, Q)\right) \rightsquigarrow 2\int_0^1 \mathbf{H}(u)\cdot\left(G^{-1}(u) - F^{-1}(u)\right)\mathrm{d}\omega(u)\,. \tag{2.3.9}$$

(iii) If $n^{1/2}\varepsilon_n$ *is a constant, say* $n^{1/2}\varepsilon_n = \gamma > 0$,

$$\begin{aligned} &nW_{2,\omega}^2(P_n, P) - \gamma^2 W_{2,\omega}^2(P, Q) \\ &\qquad \rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \,\mathrm{d}\omega(u) + 2\gamma\int_0^1 \mathbf{H}(u)\cdot\left(G^{-1}(u) - F^{-1}(u)\right)\mathrm{d}\omega(u)\,. \end{aligned} \tag{2.3.10}$$

Remark 2.3.8. One can also show that the limit distribution in (2.3.9) is a mean-zero Gaussian distribution with variance

$$\int_0^1\int_0^1 \frac{4(u_1 \wedge u_2 - u_1u_2)}{f(F^{-1}(u_1))f(F^{-1}(u_2))}\cdot\left(G^{-1}(u_1) - F^{-1}(u_1)\right)\cdot\left(G^{-1}(u_2) - F^{-1}(u_2)\right)\mathrm{d}\omega(u_1)\,\mathrm{d}\omega(u_2)\,.$$

Remark 2.3.9. The Lipschitzness assumption on $G^{-1}\circ F$ is used for (2.3.9), but not for (2.3.8) and (2.3.10). Such an assumption is well studied in the literature; see Appendix A of Bobkov and Ledoux [2019].

2.4 Simulations

2.4.1 Phase Transition and Power Analysis

This section delivers results using two experiments to numerically verify Theorem 2.2.3. For both experiments, we fix $P = \mathrm{Unif}[0, 1]$ and let ω be the Lebesgue measure so that Ψ is the

cumulative distribution function of

$$\int_0^1 |\mathbf{B}_u|^2 \, \mathrm{d}u \, .$$

Such a distribution is well studied, and the quantile of Ψ is available, see Table 1 of Anderson and Darling [1952]; we choose $\alpha = 0.05$, then $C_\alpha \approx 0.46136$.

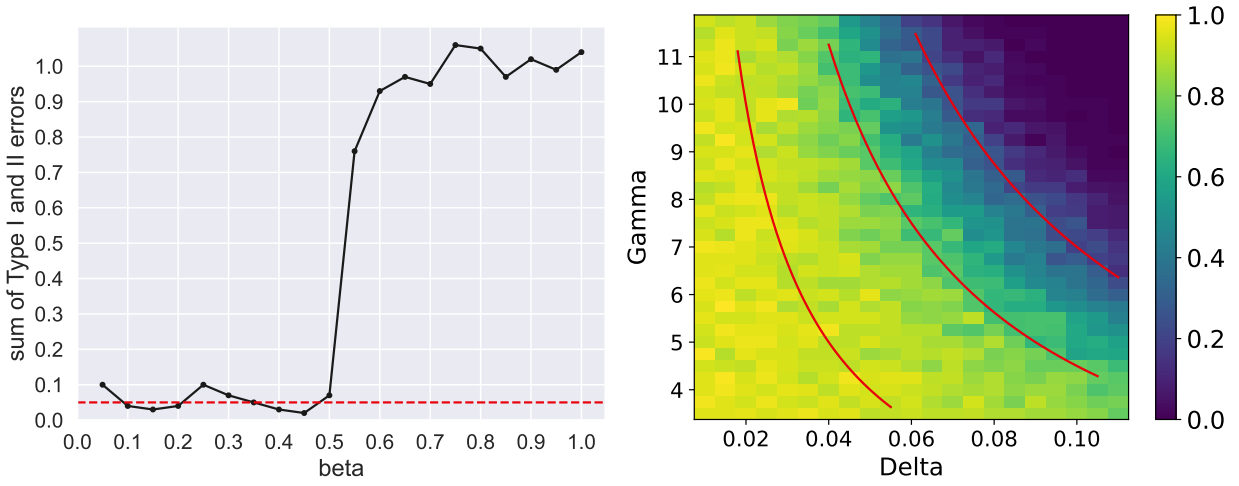
Experiment 1: Phase Transition In this experiment, we verify that the phase transition occurs at $\beta = 0.5$ for the scaling $\varepsilon_n = n^{-\beta}$. To this end, we fix $Q = N(0,1)$ and let $\varepsilon_n = n^{-\beta}$ with $n = 10^6$, and a range of $\beta \in \{0.05, 0.1, \ldots, 0.95, 1\}$. For each β, we compute $nW_2^2(P_n, P)$ using samples from $H_0^{(n)}$ and $H_1^{(n)}$, which we repeat 100 times, then compute the Type I and Type II errors using those 100 realizations, respectively. We plot the sum of Type I and Type II errors against β. Figure 2.4(a) confirms that the phase transition occurs at $\beta = 0.5$.

Experiment 2: Sharp Power Analysis We focus on the power behavior at the detection boundary, namely, $\varepsilon_n = \gamma n^{-0.5}$ with $n = 10^6$. Specifically, we analyze the Type II error based on two parameters representing the signal strength: γ and $\Delta := W_2(P, Q)$. By Theorem 2.2.3, the Type II error should be approximately $\Psi_\gamma(C_\alpha - \gamma^2\Delta^2)$. To vary Δ, we parametrized Q as follows: the quantile function of Q is $u \mapsto u + \frac{p}{2\pi}\sin(2\pi u)$, where we vary $p \in (0,1)$; note that this is a valid quantile function as it is monotonically increasing. Then,

$$W_2^2(P,Q) = \frac{p^2}{4\pi^2}\int_0^1 |\sin(2\pi u)|^2 \, \mathrm{d}u = \frac{p^2}{8\pi^2} = \Delta^2,$$

which results in $\Delta^2 \in (0, (8\pi^2)^{-1})$. We can easily generate Q's with desired signal strength Δ using this parametrization. For each pair (Δ, γ), where $\Delta \in \{0.01, 0.015, \ldots, 0.105, 0.11\}$ and $\gamma \in \{3.5, 3.75, \ldots, 11.5, 11.75\}$, we compute the Type II error and visualize it as a color map. Figure 2.4(b) visualizes Type II errors on the Δ-γ plane using colors; this essentially plots $\Psi_\gamma(C_\alpha - \gamma^2\Delta^2)$. Notice that the level set of Type II errors and the curve $\gamma \cdot \Delta = \text{constant}$

do not perfectly coincide because Type II errors can vary on the curve $\gamma \cdot \Delta = \text{constant}$, namely, $\Psi_\gamma(C_\alpha - \text{constant}^2)$ can change as γ varies.



(a) Sum of Type I and Type II errors

(b) Type II errors

Figure 2.4: (a) visualizes the sum of Type I and Type II errors of Experiment 1; the red dashed line represents the level $\alpha = 0.05$. (b) plots Type II errors of Experiment 2 as a color map, where the red solid curves represent $\gamma \cdot \Delta = \text{constant} \in \{0.2, 0.45, 0.7\}$

.

2.4.2 Comparison to Other Methods and Robustness Checks

This section first compares the power of the proposed procedure and the Kolmogorov-Smirnov (KS) test. Later, we carry out preliminary robustness checks on how the choice of the weight measure ω affects the power. Throughout this section, we again fix $P = \text{Unif}[0, 1]$ and $n = 10^6$.

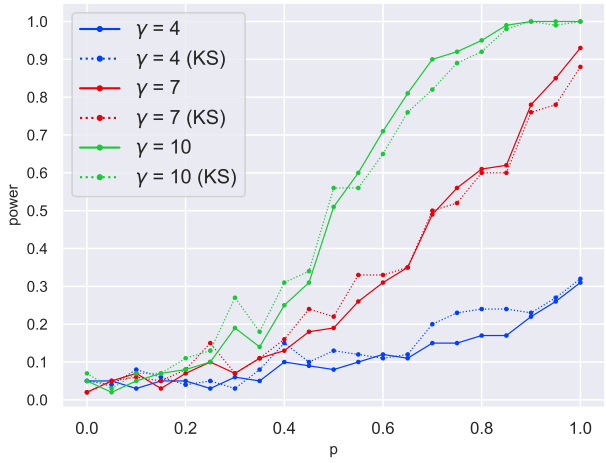
Example I First, we repeat the previous setting used for the boundary case: $\varepsilon_n = \gamma n^{-0.5}$ and the quantile function of Q is $u \mapsto u + \frac{p}{2\pi}\sin(2\pi u)$. This time, we parametrize $p \in \{0, 0.05, \ldots, 0.95, 1\}$ instead of $\Delta = W_2(P, Q)$ and restrict our interest to $\gamma \in \{4, 7, 10\}$. For each pair (p, γ), we compute the power of the proposed testing procedure—with ω being the Lebesgue measure—and the KS test; to ensure both are asymptotically level α tests with $\alpha =$

0.05, we use the critical value $C_\alpha \approx 0.46136$ for the proposed testing procedure as before and the critical value 1.36 for the KS test statistic $\sqrt{n} \cdot \mathrm{KS}(P_n, P) = \sqrt{n} \cdot \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$ based on Table 1 of [Shorack and Wellner, 2009]. Figure 2.5(a) shows the results. First, observe that for $\gamma \in \{7, 10\}$, the powers of both tests exceed 0.5 once the parameter p is above certain values, which means that both are better than a trivial test that randomly rejects the null with probability 0.5. In this case, the power of the proposed procedure is slightly larger than that of the KS test, as the solid lines are above the dotted lines. For $\gamma = 4$, the power of the KS test is mostly larger than that of the proposed procedure, as shown in the blue lines; in this case, however, the powers of both tests are far less than 0.5, implying that both tests practically failed. In other words, the signal strength $\gamma = 4$ is too weak for both tests to detect.

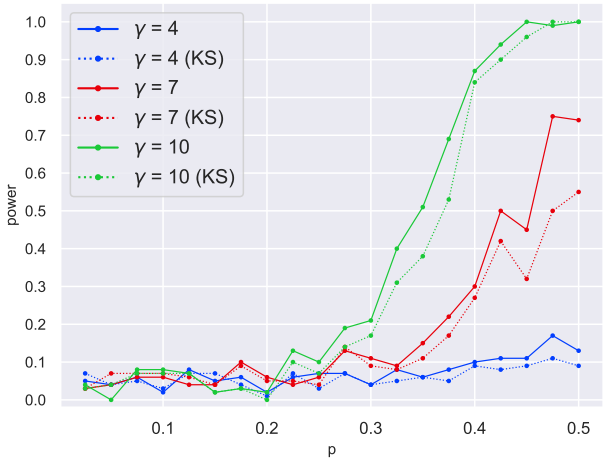
Example II Next, we consider a different setting where the quantile function of Q differs with that of $P = \mathrm{Unif}[0, 1]$ at tails. To this end, suppose the quantile function of Q is as follows:

$$\begin{cases} u + 0.45 \cdot \frac{2p}{\pi} \cos\left(\frac{\pi u}{2p}\right) & 0 \le u \le p, \\ u & p \le u \le 1 - p, \\ u - 0.45 \cdot \frac{2p}{\pi} \cos\left(\frac{\pi(1-u)}{2p}\right) & 1 - p \le u \le 1, \end{cases} \tag{2.4.1}$$

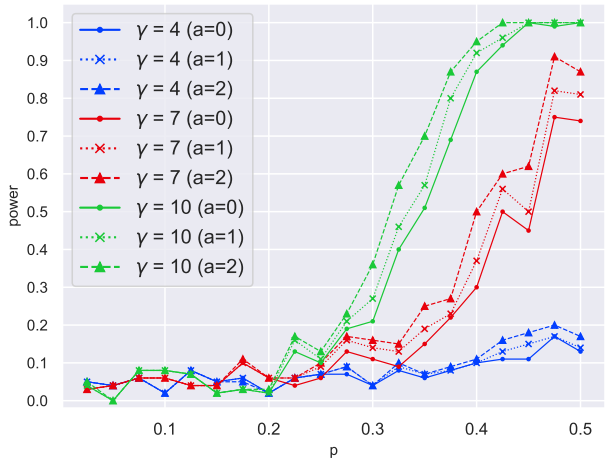
where we parametrize $p \in \{0.025, 0.05, \ldots, 0.5\}$. Note that the quantile function of Q deviates from the identity only at tails, namely, $[0, p] \cup [1 - p, 1]$. As in Example I, we keep using $\varepsilon_n = \gamma n^{-0.5}$ for $\gamma \in \{4, 7, 10\}$ and compute the powers for each pair (p, γ) which are shown in Figure 2.5(b). For $\gamma \in \{7, 10\}$, the results are similar to Example I; for sufficiently large p, both tests have enough powers. Now, for $\gamma = 4$, the power of the proposed procedure is slightly larger than that of the KS test; however, as in Example I, both tests are practically useless as their powers are too small; namely, $\gamma = 4$ is again too weak for them to detect.



(a) Powers (Example I)



(b) Powers (Example II)



(c) Powers (Example II) of different weight measures

Figure 2.5: (a) and (b) show the powers of the proposed testing procedure and the KS test in Example I and Example II, respectively; solid lines correspond to the proposed testing procedure, whereas dotted lines represent the KS test. (c) shows the powers of the proposed testing procedures in the setting of Example II, using the weight measure given by (2.4.2). By design, the solid lines of (b) and (c) coincide; both correspond to the Lebesgue measure, namely, $a = 0$ in (2.4.2).

The Role of the Weight Measure Lastly, we compare the powers by differing the weight measure ω. Let us keep considering the setting of Example II. As the quantile function Q mainly deviates from that of P at tails, it is reasonable to put more weights at tails for better performance when computing the test statistic. To verify this, consider a simple weight measure whose density with respect to the Lebesgue measure is a quadratic function:

for $a \geq 0$,

$$\frac{\mathrm{d}\omega(u)}{\mathrm{d}u} = a\left(u - \frac{1}{2}\right)^2 + 1 - \frac{a}{12}, \tag{2.4.2}$$

which satisfies $\int_0^1 \mathrm{d}\omega = 1$. Clearly, $a = 0$ corresponds to the Lebesgue measure; large a means more weights at both tails. We compute the powers of the proposed procedure for $a \in \{0, 1, 2\}$. Here, we again set critical values to ensure that all of them are asymptotically level $\alpha = 0.05$; for $a = 0$, we can reuse the previous critical value as before. For $a \in \{1, 2\}$, we estimate the quantile of the asymptotic distribution $\int_0^1 |\mathbf{B}_u|^2 \,\mathrm{d}\omega(u)$ by Monte Carlo simulations. Figure 2.5(c) shows the results. As expected, we obtain larger powers with the quadratic weight measure, namely, both $a = 1$ and $a = 2$ achieve better performance than the unweighted procedure $a = 0$; particularly, assigning more weights at tails ($a = 2$) yields the largest power. For $\gamma = 4$, though we have increased powers for the weighted procedures, they are still practically too weak to detect the signal, as discussed in the previous examples.

2.4.3 Application I: Distribution Shifts in Consumer Spending

We revisit the data example in Section 2.1.2 and apply the proposed testing procedure to study the power. Recall that $\{P_{i/11}\}_{i=0}^{11}$ are the distributions of monthly average spending by county during the recovery period between March 16, 2020 and March 15, 2021. For $t \in \{0, 1/11, \ldots, 10/11, 1\}$ and $n \in \{10, 50, 100, 500\}$, we construct the empirical measure P_t^n using n points that are i.i.d. from P_t and compute $W_2(P_0, P_t^n)$. Essentially, for each t and sample size n, we want to distinguish P_t from P_0 using a finite sample. To this end, we first estimate the quantile of $W_2(P_0, P_0^n)$ via Monte Carlo simulations to define a level $\alpha = 0.05$ test; this will give us a critical value, say, C_α^n. Then, for each $t > 0$, we compute $W_2(P_0, P_t^n)$ and reject the null—that is, data are from P_0—if it exceeds C_α^n; by repeating this for 100 times, we can evaluate the power of the test using simulations. The results are shown in the first 4 rows (Recovery Period) of Table 2.1. In this case, for $t \geq 3/11$, namely, after three months from March, 2020, we can detect the distribution shift from P_0

for any sample size n. In other words, we can tell there has been a significant recovery after three months; indeed, the shift is significant enough so we can tell the difference from P_0 by estimating P_t using only 10 randomly chosen counties instead of the total 1,655 counties. The first month after P_0, that is, for $t = 1/11$, the shift is relatively weaker compared to the subsequent months, so $n = 10$ yields power 0.56, which is not enough to detect the shift; in other words, for $t = 1/11$, we need at least $n = 50$ to distinguish it from P_0.

How about the power for the stable period? We repeat the above procedure by taking $\{P_{i/11}\}_{i=0}^{11}$ as the distributions during the stable period, namely, they amount to the histograms in Figure 2.2(c); again, P_0 and P_1 correspond to the start and end of this period (from March 16, 2021 to April 15, 2021 and from February 16, 2022 to March 15, 2022). Recall from Figure 2.2(c) that we no longer see a significant distribution shift as in the recovery period. The resulting powers are shown in the last 4 rows (Stable Period) of Table 2.1. Unlike the recovery period, we can observe the powers are much smaller for $n \leq 100$; particularly, most of the powers—except for a few periods with $n = 100$—are far below 0.5, meaning that we cannot detect a meaningful shift from P_0 using finite samples. For detection to be possible, we can see that we need a sufficiently large sample size $n = 500$.

In summary, the results in Table 2.1 essentially demonstrate that the proposed testing procedure can detect the distribution shifts during the recovery period as it is reasonably approximated by displacement interpolation as seen in Figure 2.2(b), while there are no such shifts to be captured by the proposed procedure during the stable period because the distributions are similar to each other as shown in Figure 2.2(c).

2.4.4 Application II: p-Value Heterogeneity Across Disciplines

We apply our testing procedure to the data set from Head et al. [2015], which collects p-values of statistical tests published in journals across different disciplines. In this example, we use our procedure to determine if the distribution of p-values differs depending on disciplines.

n	$11t$											
	1	2	3	4	5	6	7	8	9	10	11	
10	0.56	0.98	1	1	1	1	1	1	1	1	1	
50	0.98	1	1	1	1	1	1	1	1	1	1	**Recovery**
100	1	1	1	1	1	1	1	1	1	1	1	**Period**
500	1	1	1	1	1	1	1	1	1	1	1	
10	0.04	0.05	0.09	0.1	0.05	0.07	0.1	0.05	0.07	0.11	0.23	
50	0.2	0.23	0.36	0.38	0.22	0.18	0.1	0.14	0.09	0.4	0.75	**Stable**
100	0.57	0.45	0.66	0.66	0.45	0.29	0.43	0.33	0.16	0.76	0.96	**Period**
500	1	0.99	1	1	1	0.94	1	0.99	0.88	1	1	

Table 2.1: Powers for the recovery and stable periods.

To this end, we first collect p-values across 9 disciplines subsampled from total 21 disciplines, then take 80% of them as the null distribution P_{Null}, which consists of 90,950 observations; this is shown as the solid blue curve in Figure 2.6. Then, from the remaining 20%, we sample two disciplines "medical and health sciences" (P_{MH}) and "multidisciplinary" (P_{Mu}), which consist of 12,928 and 5,731 observations, respectively.[3] Figure 2.6 shows the cumulative distribution functions of the null P_{Null}, P_{MH}, and P_{Mu}, which are close to each other. This suggests that both P_{MH} and P_{Mu} are weak signals that are hard to distinguish from P_{Null} visually. To tackle this problem, we apply our testing procedure to $H_0 : P_{\text{MH}} = P_{\text{Null}}$. First, we compute the test statistic $T_{\text{MH}} := nW_2^2(P_{\text{MH}}, P_{\text{Null}})$, where $n := |P_{\text{MH}}| = 12{,}928$ observations. Then, setting $\alpha = 0.05$, we estimate the $(1-\alpha)$-th quantile of $nW_2^2(P_n, P_{\text{Null}})$, where P_n is the empirical measure based on $X_1, \ldots, X_n$ that are i.i.d. from P_{Null}, which can be done by resampling P_n. We obtain $T_{\text{MH}} = 0.004821$ which is larger than the estimated quantile 0.002918; also, we can estimate the p-value of T_{MH}, which is $0.005 \ll \alpha$, suggesting that we reject $H_0 : P_{\text{MH}} = P_{\text{Null}}$. We apply the same procedure to P_{Mu} with $n := |P_{\text{Mu}}| = 5{,}731$ observations. We obtain $T_{\text{Mu}} = 0.002334$, which is slightly below the estimated quantile 0.002412, and the p-value of T_{Mu} is $0.066 > \alpha$, implying that we cannot $H_0 : P_{\text{Mu}} = P_{\text{Null}}$ at level α. Therefore, our testing procedure provides a rigorous framework to detect weak signals that are otherwise indistinguishable, as visualized in Figure 2.6.

3. They have the largest number of observations among all 9 disciplines.

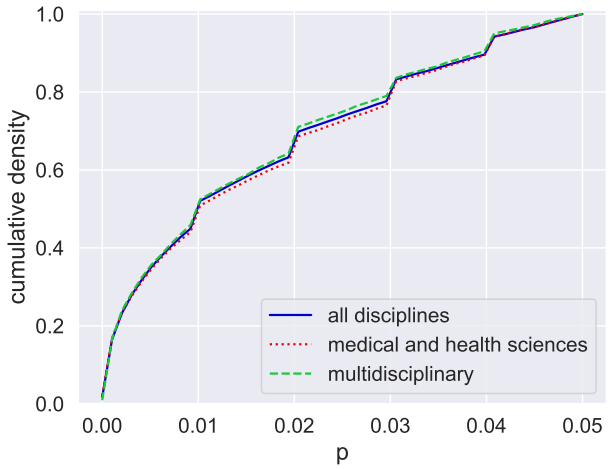


Figure 2.6: This figure plots the cumulative distribution functions of p-values (below 0.05) across nine disciplines subsampled from [Head et al., 2015], which is shown as the solid blue curve, along with two disciplines: "medical and health sciences" (red, dotted) and "multidisciplinary" (green, dashed).

2.5 Discussion

Comparison with Existing Methods Notice that the problem (2.1.3) can be viewed as an instance of general testing

$$H_0 : X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P \quad \text{vs.} \quad H_1 : X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Q_n,$$

where $Q_n \to P$ as $n \to \infty$; our problem is the case where Q_n is given as displacement interpolation between P and Q. Existing approaches to such a general problem characterize detection by utilizing the following (related) notions: likelihood ratio $\frac{\mathrm{d}Q_n}{\mathrm{d}P}$ (Section 13.10.1 of van der Vaart and Wellner [1996]) or Hellinger distance $H^2(P, Q_n)$ (Section 13.1 of Lehmann and Romano [2005]). The former approach—often referred to as Le Cam's third lemma—derives limit distributions based on asymptotic normality; the latter characterizes the detection boundary depending on the rate of $nH^2(P, Q_n)$. The first principle behind these methods is clear: quantify a distance/discrepancy between P, Q_n and characterize the limit. Our results also utilize such a first principle: we use weighted Wasserstein distances.

Though our method and the existing methods share a similar high-level idea, the existing methods are unsuitable for our problem for several reasons. First, though Le Cam's third lemma applies to any abstract setting under certain conditions, it does not lead to the exact characterization of testing errors unless there is a suitable parametric assumption. Second, the Hellinger distance is not preferred in analyzing the case where Q_n is given as displacement interpolation as its relationship with the interpolation parameter ε_n is not transparent. Moreover, the Hellinger distance-based characterization generally does not calculate testing errors at the detection boundary. On the other hand, our method motivated by weighted Wasserstein distances not only interplays well with displacement interpolation but also provides the exact characterization of testing errors, including the boundary case; moreover, unlike the likelihood ratio test, which requires information on both P and Q_n, our test is implementable as long as P is known.

Lifting Technical Assumptions As remarked in Remark 2.3.9, the Lipschitzness assumption on $G^{-1} \circ F$ is used in the detectable case ($n^{1/2}\varepsilon_n \to \infty$), but not in the other two cases. Removing the Lipschitzness assumption, there is no restriction on Q as long as its quantile function G^{-1} is bounded. Accordingly, our main results hold even when Q is discrete; in particular, there are cases where our method applies, but Le Cam's third lemma cannot because contiguity is not satisfied. As a concrete example, suppose $P = \mathrm{Unif}[0,1]$ and $Q = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. Then, $Q_n := ((1-\varepsilon_n)\mathrm{Id} + \varepsilon_n G^{-1} \circ F)_{\#}P$ is a uniform measure supported on the union of two disjoint intervals $U_n := [0,(1-\varepsilon_n)/2] \cup [(1+\varepsilon_n)/2, 1]$. One can verify that $Q_n^{\otimes n}$—the n tensor product of Q_n—is not contiguous with respect to $P^{\otimes n}$ because $Q_n^{\otimes n}(U_n^n) = 1$ while $P^{\otimes n}(U_n^n) = (1-\varepsilon_n)^n \to 0$ at the boundary $\varepsilon_n \asymp n^{-1/2}$. Hence, Le Cam's third lemma cannot be applied to derive the limit distribution for such a case, whereas our method can. Lastly, we briefly discuss Assumption 2.3.1. Though it is a reasonable assumption to impose from a practical viewpoint as discussed in Remark 2.3.3, it is natural to ask—from a technical viewpoint—whether such an assumption can be relaxed. To

circumvent Assumption 2.3.1, we can use an alternative pivotal test based on the weighted Wasserstein between Unif$[0,1]$ and the empirical measure based on $F(X_1), \ldots, F(X_n)$; then, we can show that similar theory holds.

Optimal Testing Procedure Displacement interpolation motivates the weighted Wasserstein distance as a natural test statistic for (2.1.3). While the main theory of this chapter focuses on the asymptotic power of the testing procedure based on the weighted Wasserstein distance, we believe there are several interesting theoretical questions to be addressed in future work. Particularly, one may ask whether the proposed procedure is optimal for testing (2.1.3). The simulation results in Section 2.4.2 suggest that optimality would depend on P and Q. Hence, it would be interesting to derive a minimax optimal procedure that minimizes the worst case testing error over some class of distributions P, Q, which we leave as important future work.

Extension to Two-Sample Testing The problem (2.1.3) itself and the proposed procedure essentially postulate that P is known. In the empirical applications presented in Section 2.4, we treated finite data points as the null distribution P and applied the proposed procedure. Though the purpose of such treatment was to demonstrate the applicability of the proposed procedure in simplified settings, the fact that the finite data points, say, $Y_1, \ldots, Y_m$, are usually noisy samples from some distribution P of interest is a crucial limitation in practice. To address this issue, one may consider a two-sample testing framework asking if $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are from the same distribution. We can still consider a local alternative given as displacement interpolation, say, $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} ((1-\varepsilon)\mathrm{Id} + \varepsilon G^{-1} \circ F)_{\#} P$ and $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} P$. The weighted Wasserstein distance between the empirical measures of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ is deemed a plausible test statistic under this framework. We leave the analysis of such a two-sample testing framework as future work.

2.6 Proofs

2.6.1 Proof of Lemma 2.3.5

Proof. Recall from Lemma 2.3.2 that $\mathbf{H}_n \rightsquigarrow \mathbf{H}$ in $\ell^\infty(0,1)$. Now, define

$$\ell_m^\infty(0,1) = \{h \in \ell^\infty(0,1) : h \text{ is measurable}\},$$

then one can verify that $\ell_m^\infty(0,1)$ is a closed subset of the Banach space $(\ell^\infty(0,1), \|\cdot\|_\infty)$. Also, as the sample path of $\mathbf{H}_n$ is always monotone and the sample path of $\mathbf{H}$ is always continuous, $\mathbf{H}_n$ and $\mathbf{H}$ take values in $\ell_m^\infty(0,1)$. Now, define a map $\mathcal{I}\colon \ell_m^\infty(0,1) \to \mathbb{R}$ by

$$\mathcal{I}(h) = \int_0^1 |h(u)|^2 \,\mathrm{d}\omega(u)\,.$$

We claim that $\mathcal{I}$ is continuous. To this end, consider a sequence $(h_n)_{n\in\mathbb{N}}$ in $\ell_m^\infty(0,1)$ and $h \in \ell_m^\infty(0,1)$ such that $\|h_n - h\|_\infty \to 0$. Then, $\sup_{n\in\mathbb{N}} \|h_n\|_\infty \le M$ for some $M > 0$, hence the dominated convergence theorem shows that $\mathcal{I}(h_n) \to \mathcal{I}(h)$. Now, applying Theorem 2.6.1 (stated below), we conclude that $\mathcal{I}(\mathbf{H}_n) \rightsquigarrow \mathcal{I}(\mathbf{H})$, which proves (2.3.5). Similarly, for each $n \in \mathbb{N}$, define a map $\mathcal{E}_n\colon \ell_m^\infty(0,1) \to \mathbb{R}$ by

$$\mathcal{E}_n(h) = \int_0^1 |(G^{-1}\circ F)(F^{-1}(u) + n^{-1/2}h(u)) - F^{-1}(u)|^2 \,\mathrm{d}\omega(u)\,.$$

For any sequence $(h_n)_{n\in\mathbb{N}}$ in $\ell_m^\infty(0,1)$ and $h \in \ell_m^\infty(0,1)$ such that $\|h_n - h\|_\infty \to 0$, we claim

$$\mathcal{E}_n(h_n) \to \int_0^1 |G^{-1}(u) - F^{-1}(u)|^2 \,\mathrm{d}\omega(u)\,. \tag{2.6.1}$$

As G^{-1} is continuous almost everywhere on $(0,1)$, we can see that $G^{-1}\circ F\circ(F^{-1}+n^{-1/2}h_n)$ converges to G^{-1} almost everywhere; as ω is absolutely continuous with respect to the Lebesgue measure, this convergence holds ω-almost everywhere as well. As G^{-1} and F^{-1}

are bounded on $(0,1)$, the dominated convergence theorem shows (2.6.1). By Theorem 2.6.1,

$$\mathcal{E}_n(\mathbf{H}_n) \rightsquigarrow \int_0^1 |G^{-1}(u) - F^{-1}(u)|^2 \,\mathrm{d}\omega(u)\,,$$

showing (2.3.6). Lastly, to prove (2.3.7), for each $n \in \mathbb{N}$, define a map $\mathcal{J}_n \colon \ell_m^\infty(0,1) \to \mathbb{R}$ by

$$\mathcal{J}_n(h) = \int_0^1 h(u) \cdot \Big((G^{-1} \circ F)(F^{-1}(u) + n^{-1/2} h(u)) - F^{-1}(u) \Big) \,\mathrm{d}\omega(u)\,.$$

Also, define $\mathcal{J} \colon \ell_m^\infty(0,1) \to \mathbb{R}$ by

$$\mathcal{J}(h) = \int_0^1 h(u) \cdot \Big(G^{-1}(u) - F^{-1}(u) \Big) \,\mathrm{d}\omega(u)\,.$$

Similarly to $\mathcal{I}$, one can verify that $\mathcal{J}$ is continuous. Also, for any sequence $(h_n)_{n\in\mathbb{N}}$ in $\ell_m^\infty(0,1)$ and $h \in \ell_m^\infty(0,1)$ such that $\|h_n - h\|_\infty \to 0$, we can show that $\mathcal{J}_n(h_n) \to \mathcal{J}(h)$ by means of the dominated convergence theorem. Therefore, $\mathcal{J}_n(\mathbf{H}_n) \rightsquigarrow \mathcal{J}(\mathbf{H})$ holds by Theorem 2.6.1, hence (2.3.7) holds. □

Theorem 2.6.1 (Extended Continuous Mapping Theorem)**.** *Let $\mathcal{H}$ and $\mathcal{K}$ be metric spaces. Let $\mathcal{H}_0$ be a Borel subset of $\mathcal{H}$ and consider a measurable map $\mathcal{I} \colon \mathcal{H}_0 \to \mathcal{K}$. Suppose there is a sequence $(\mathcal{I}_n)_{n\in\mathbb{N}}$ of maps from $\mathcal{H}_0$ to $\mathcal{K}$ satisfying the following: for any sequence $(h_n)_{n\in\mathbb{N}}$ in $\mathcal{H}_0$ converging to some $h \in \mathcal{H}_0$, the sequence $(\mathcal{I}_n(h_n))_{n\in\mathbb{N}}$ converges to $\mathcal{I}(h)$ in $\mathcal{K}$. If a sequence $(H_n)_{n\in\mathbb{N}}$ of random elements in $\mathcal{H}$ converges weakly to a tight measurable random element H in $\mathcal{H}$, where both H_n and H take values in $\mathcal{H}_0$, the sequence $(\mathcal{I}_n(H_n))_{n\in\mathbb{N}}$ of random elements in $\mathcal{K}$ converges weakly to the random element $\mathcal{I}(H)$ in $\mathcal{K}$.*

Remark 2.6.2. Theorem 2.6.1 is adapted from Theorem 1.11.1 of van der Vaart and Wellner [1996], where the latter uses separability instead of tightness. As tightness implies separability, we have stated Theorem 2.6.1 with tightness, which is sufficient in our analysis.

2.6.2 Proof of Theorem 2.3.7

Proof. For each $n \in \mathbb{N}$, let $\phi_n = (1-\varepsilon_n)\mathrm{Id} + \varepsilon_n G^{-1} \circ F$. As we are concerned with the limit distribution of $nW_{2,\omega}^2(P_n, P)$, we may assume $(X_n)_{n\in\mathbb{N}}$ is i.i.d. from P and let P_n be the empirical measure based on $\phi_n(X_1), \ldots, \phi_n(X_n)$ as $\phi_n(X_1), \ldots, \phi_n(X_n)$ also follow $H_1^{(n)}$. Now, let F_n^{-1} be the empirical quantile function based on $X_1, \ldots, X_n$, then

$$\begin{aligned}\phi_n \circ F_n^{-1} - F^{-1} &= (1-\varepsilon_n)F_n^{-1} + \varepsilon_n G^{-1} \circ F \circ F_n^{-1} - F^{-1} \\ &= (1-\varepsilon_n)n^{-1/2}\mathbf{H}_n + \varepsilon_n(G^{-1} \circ F \circ (F^{-1} + n^{-1/2}\mathbf{H}_n) - F^{-1}),\end{aligned}$$

where $\mathbf{H}_n := \sqrt{n}(F_n^{-1} - F^{-1})$. Hence, using A_n, B_n, C_n defined in (2.3.2), (2.3.3), (2.3.4), respectively, we have

$$W_{2,\omega}^2(P_n, P) = (1-\varepsilon_n)^2 n^{-1} A_n + \varepsilon_n^2 B_n + 2(1-\varepsilon_n)n^{-1/2}\varepsilon_n C_n.$$

Case I: Suppose $n^{1/2}\varepsilon_n \to 0$. Recall that we have shown in Lemma 2.3.5 that A_n, B_n, and C_n are weakly convergent. Note that

$$nW_{2,\omega}^2(P_n, P) = (1-\varepsilon_n)^2 A_n + (n^{1/2}\varepsilon_n)^2 B_n + 2(1-\varepsilon_n)(n^{1/2}\varepsilon_n)C_n.$$

As $(n^{1/2}\varepsilon_n)^2 B_n, (n^{1/2}\varepsilon_n)C_n \rightsquigarrow 0$, by Slutsky's theorem,

$$nW_{2,\omega}^2(P_n, P) \rightsquigarrow \lim_{n\to\infty} A_n = \int_0^1 |\mathbf{H}(u)|^2 \, \mathrm{d}\omega(u)\,.$$

Case II: Suppose $n^{1/2}\varepsilon_n \to \infty$ and note that

$$\begin{aligned}n^{1/2}\varepsilon_n^{-1}\left(W_{2,\omega}^2(P_n, P) - \varepsilon_n^2 W_{2,\omega}^2(P, Q)\right) &= (1-\varepsilon_n)^2(n^{1/2}\varepsilon_n)^{-1}A_n \\ &\quad + n^{1/2}\varepsilon_n(B_n - W_{2,\omega}^2(P, Q)) + 2(1-\varepsilon_n)C_n.\end{aligned}$$

First, notice that $(n^{1/2}\varepsilon_n)^{-1}A_n \rightsquigarrow 0$. We claim $n^{1/2}\varepsilon_n(B_n - W_{2,\omega}^2(P,Q)) \rightsquigarrow 0$. To this end, observe that

$$
\begin{aligned}
&B_n - W_{2,\omega}^2(P,Q) \\
&= \int_0^1 |(G^{-1}\circ F)(F^{-1}(u) + n^{-1/2}\mathbf{H}_n(u)) - F^{-1}(u)|^2 \,\mathrm{d}\omega(u) \\
&\quad - \int_0^1 |G^{-1}(u) - F^{-1}(u)|^2 \,\mathrm{d}\omega(u) \\
&= \int_0^1 |(G^{-1}\circ F)(F^{-1}(u) + n^{-1/2}\mathbf{H}_n(u)) - G^{-1}(u)|^2 \,\mathrm{d}\omega(u) \\
&\quad + 2\int_0^1 \Big((G^{-1}\circ F)(F^{-1}(u) + n^{-1/2}\mathbf{H}_n(u)) - G^{-1}(u)\Big)\cdot\Big(G^{-1}(u) - F^{-1}(u)\Big) \,\mathrm{d}\omega(u) \\
&=: B_n^o + 2B_n^*.
\end{aligned}
$$

Let L be the Lipschitz constant of $G^{-1}\circ F$, then

$$
n^{1/2}\varepsilon_n|B_n^*| \le n^{1/2}\varepsilon_n \int_0^1 Ln^{-1/2}|\mathbf{H}_n(u)|\cdot|G^{-1}(u) - F^{-1}(u)| \,\mathrm{d}\omega(u) \le L\varepsilon_n\sqrt{A_n}W_{2,\omega}^2(P,Q),
$$

where $\varepsilon_n\sqrt{A_n} \rightsquigarrow 0$ by Lemma 2.3.5, hence $n^{1/2}\varepsilon_n B_n^* \rightsquigarrow 0$. Similarly,

$$
n^{1/2}\varepsilon_n|B_n^0| \le n^{1/2}\varepsilon_n \int_0^1 L^2 n^{-1}|\mathbf{H}_n(u)|^2 \,\mathrm{d}\omega(u) = L^2 n^{-1/2}\varepsilon_n A_n \rightsquigarrow 0,
$$

hence $n^{1/2}\varepsilon_n B_n^o \rightsquigarrow 0$. Therefore, applying Slutsky's theorem, we have

$$
\begin{aligned}
n^{1/2}\varepsilon_n^{-1}\Big(W_{2,\omega}^2(P_n,P) - \varepsilon_n^2 W_{2,\omega}^2(P,Q)\Big) &\rightsquigarrow \lim_{n\to\infty} 2(1-\varepsilon_n)C_n \\
&= 2\int_0^1 \mathbf{H}(u)\cdot\Big(G^{-1}(u) - F^{-1}(u)\Big) \,\mathrm{d}\omega(u)\,.
\end{aligned}
$$

Case III: Suppose $n^{1/2}\varepsilon_n = \gamma$, then

$$
nW_{2,\omega}^2(P_n,P) = (1-\varepsilon_n)^2 A_n + \gamma^2 B_n + 2(1-\varepsilon_n)\gamma C_n = (1-\varepsilon_n)\cdot((1-\varepsilon_n)A_n + 2\gamma C_n) + \gamma^2 B_n.
$$

We claim that

$$(1-\varepsilon_n)A_n + 2\gamma C_n \rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \,\mathrm{d}\omega(u) + 2\gamma \int_0^1 \mathbf{H}(u) \cdot \Big(G^{-1}(u) - F^{-1}(u)\Big) \,\mathrm{d}\omega(u)\,.$$

We apply the same argument used in the proof of Lemma 2.3.5; for each $n \in \mathbb{N}$, define $\mathcal{K}_n \colon \ell_m^\infty(0,1) \to \mathbb{R}$ by

$$\begin{aligned}\mathcal{K}_n(h) = (1-\varepsilon_n) &\int_0^1 |h(u)|^2 \,\mathrm{d}\omega(u) \\ &+ 2\gamma \int_0^1 h(u) \cdot \Big((G^{-1} \circ F)(F^{-1}(u) + n^{-1/2}h(u)) - F^{-1}(u)\Big) \,\mathrm{d}\omega(u)\,.\end{aligned}$$

In the proof of Lemma 2.3.5, we have shown that

$$\mathcal{K}_n(h_n) \to \int_0^1 |h(u)|^2 \,\mathrm{d}\omega(u) + 2\gamma \int_0^1 h(u) \cdot \Big(G^{-1}(u) - F^{-1}(u)\Big) \,\mathrm{d}\omega(u)$$

for any sequence $(h_n)_{n\in\mathbb{N}}$ in $\ell_m^\infty(0,1)$ and $h \in \ell_m^\infty(0,1)$ such that $\|h_n - h\|_\infty \to 0$. Therefore, the extended continuous mapping theorem shows that

$$(1-\varepsilon_n)A_n + 2\gamma C_n = \mathcal{K}_n(\mathbf{H}_n) \rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \,\mathrm{d}\omega(u) + 2\gamma \int_0^1 \mathbf{H}(u) \cdot \Big(G^{-1}(u) - F^{-1}(u)\Big) \,\mathrm{d}\omega(u)\,.$$

Hence, by Slutsky's theorem,

$$\begin{aligned}&nW_{2,\omega}^2(P_n, P) \\ &= (1-\varepsilon_n) \cdot ((1-\varepsilon_n)A_n + 2\gamma C_n) + \gamma^2 B_n \\ &\rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \,\mathrm{d}\omega(u) + 2\gamma \int_0^1 \mathbf{H}(u) \cdot \Big(G^{-1}(u) - F^{-1}(u)\Big) \,\mathrm{d}\omega(u) + \gamma^2 W_{2,\omega}^2(P, Q).\end{aligned}$$

□

2.6.3 Proof of Theorem 2.2.3

Proof. For the asymptotic Type I error, we invoke Proposition 2.3.6: letting P_n be the empirical measure based on $X_1, \ldots, X_n$ from $H_0^{(n)}$,

$$\text{the asymptotic Type I error} = \lim_{n\to\infty} \mathbb{P}(nW_{2,\omega}^2(P_n, P) > C_\alpha) = \mathbb{P}(A > C_\alpha) = 1 - \Psi(C_\alpha) = \alpha,$$

where A is the limit distribution of $nW_{2,\omega}^2(P_n, P)$ under $H_0^{(n)}$, namely, $A =$ (2.2.2). Here, weak convergence implies the above limit as the open set (C_α, ∞) is a continuity set of Ψ. Therefore, the asymptotic Type I error is α.

Next, we compute the asymptotic Type II error: assuming P_n be the empirical measure based on $X_1, \ldots, X_n$ from $H_1^{(n)}$,

$$\text{the asymptotic Type II error} = \lim_{n\to\infty} \mathbb{P}(nW_{2,\omega}^2(P_n, P) \le C_\alpha).$$

If $n^{1/2}\varepsilon_n \to 0$, we have shown in Theorem 2.3.7 that the limit distribution of $nW_{2,\omega}^2(P_n, P)$ is exactly A, hence

$$\lim_{n\to\infty} \mathbb{P}(nW_{2,\omega}^2(P_n, P) \le C_\alpha) = \mathbb{P}(A \le C_\alpha) = \Psi(C_\alpha) = 1 - \alpha.$$

If $n^{1/2}\varepsilon_n \to \infty$, we apply Slutsky's theorem by multiplying $n^{-1/2}\varepsilon_n^{-1}$ to both sides of (2.3.9), which yields

$$\varepsilon_n^{-2}\left(W_{2,\omega}^2(P_n, P) - \varepsilon_n^2 W_{2,\omega}^2(P, Q)\right) \rightsquigarrow 0.$$

In other words, $\varepsilon_n^{-2} W_{2,\omega}^2(P_n, P)$ converges to a constant $W_{2,\omega}^2(P, Q) > 0$ under $H_1^{(n)}$, hence

$$\lim_{n\to\infty} \mathbb{P}(nW_{2,\omega}^2(P_n, P) \le C_\alpha) = \lim_{n\to\infty} \mathbb{P}(\varepsilon_n^{-2} W_{2,\omega}^2(P_n, P) \le n^{-1}\varepsilon_n^{-2} C_\alpha) = 0.$$

If $n^{1/2}\varepsilon_n = \gamma > 0$, the limit distribution of $T_n := nW_{2,\omega}^2(P_n, P) - \gamma^2 W_{2,\omega}^2(P, Q)$ under $H_1^{(n)}$

is (2.2.3) by (2.3.10), hence

$$\lim_{n\to\infty} \mathbb{P}(nW_{2,\omega}^2(P_n, P) \le C_\alpha) = \lim_{n\to\infty} \mathbb{P}(T_n \le C_\alpha - \gamma^2 W_{2,\omega}^2(P, Q)) = \Psi_\gamma(C_\alpha - \gamma^2 W_{2,\omega}^2(P, Q)).$$

Here, weak convergence implies the above limit as the set $(-\infty, C_\alpha]$ is a continuity set of Ψ_γ. □

Remark 2.6.3. As noted in Remark 2.2.4, we may replace the detection boundary $n^{1/2}\varepsilon_n = \gamma > 0$ with $\lim_{n\to\infty} n^{1/2}\varepsilon_n = \gamma > 0$ by modifying the proofs of Theorem 2.3.7 and 2.2.3.

CHAPTER 3

ONLINE LEARNING TO TRANSPORT VIA THE MINIMAL SELECTION PRINCIPLE

3.1 Introduction

Online learning and online convex optimization offer an elegant framework for regret minimization under worst-case sequences [Gordon, 1999, Zinkevich, 2003, Shalev-Shwartz, 2012, Orabona, 2021]. Principles for these methods have been discovered independently in decision theory, game theory, learning theory, and convex optimization [Cesa-Bianchi and Lugosi, 2006]. For problems with a finite-dimensional decision variable, extensive studies have been conducted to understand online algorithms that achieve small regret, either in Euclidean or non-Euclidean settings. Such algorithms usually rely on some notion of (sub-)gradients to iteratively improve the finite-dimensional decision variable. In the simplest Euclidean setting, let $\ell_t \colon \mathbb{R}^d \to \mathbb{R}$ be a smooth convex function and $x_t \in \mathbb{R}^d$ be a finite-dimensional decision variable, both indexed by time $t \in \mathbb{N}$. Online gradient descent with step size η satisfies, for any $z \in \mathbb{R}^d$,

$$2\eta\big(\overbrace{\ell_t(x_t) - \ell_t(z)}^{\text{regret term}}\big) - \big(\overbrace{\|x_t - z\|^2 - \|x_{t+1} - z\|^2}^{\text{telescoping term}}\big) \leq \overbrace{\|x_t - x_{t+1}\|^2}^{\text{transport cost}} \leq \overbrace{\|\eta \nabla \ell_t(x_t)\|^2}^{\text{gradient in Euclidean norm}} . \tag{3.1.1}$$

In a simple non-Euclidean setting, let $p_t \in \Delta_d$ be a probability distribution on a d-discrete decision and $\ell_t \in \mathbb{R}^d$ be a sequence of cost vectors where each element $\ell_t[k]$ denotes the cost associated with the k-th action and $\ell_t(p_t) := \sum_{k=1}^d \ell_t[k] p_t[k]$ thus $\nabla \ell_t(\cdot) \equiv \ell_t$. Online

mirror descent (specifically, exponentiated gradient) satisfies for any $q \in \Delta_d$

$$\eta\big(\overbrace{\ell_t(p_t) - \ell_t(q)}^{\text{regret term}}\big) - \big(\overbrace{d_{\mathrm{KL}}(q||p_t) - d_{\mathrm{KL}}(q||p_{t+1})}^{\text{telescoping term}}\big) \le \overbrace{d_{\mathrm{KL}}(p_t||p_{t+1})}^{\text{transport cost}} \le \overbrace{\|\eta\nabla\ell_t\|_{p_t}^2}^{\text{gradient in local norm}}, \tag{3.1.2}$$

where the local norm is defined as $\|\ell_t\|_{p_t}^2 := \sum_{k=1}^d (\ell_t[k])^2 p_t[k]$. It is immediate to spot the resemblance between (3.1.1) and (3.1.2). The telescoping term marks the progress towards the competing decision (z or q), and the first right-hand side of the expression describes some notion of "transport cost" induced by different geometries.

It is natural to wonder how the theoretical and algorithmic principles behind these finite-dimensional cases extend to the infinite-dimensional case, specifically when the decision variable is a probability distribution over a generic metric space. Such a question is of practical importance. Several dynamic resource allocation problems in operations research involve these infinite-dimensional objects, e.g., routing a network of drones over airspace or allocating drivers across a city topology in ridesharing. Theoretically, the infinite-dimensional object can be viewed as a generalization of both (3.1.1) (x from finite to infinite-dimensional) and (3.1.2) (p from discrete to continuous measure). As hinted in the previous paragraph where small "transport cost" terms guarantee small regret, optimal transport theory [Villani, 2003, Ambrosio et al., 2005] plays a pivotal role in extending online learning to the infinite-dimensional case. This chapter draws connections between online learning, optimal transport, and partial differential equations (PDEs) using an insight called the minimal selection principle. Eventually, we derive new algorithms based on this insight.

This chapter investigates online learning using a toolbox set out by Brenier, who brought perspectives from PDEs, geometry, and functional analysis to the study of Optimal transport (OT) around 1987 [Brenier, 1987, 1989]. Let us first introduce our infinite-dimensional online optimization problem in the language of OT and then lay out the connection between OT and

online learning. Consider the following **Online Learning to Transport (OLT)** problem: let $\mathcal{P}(\mathbb{R}^d)$ be the space of probability measures on $\mathbb{R}^d$ where, in each round, $t = 1, \ldots, T$,

- an adversary chooses an energy functional $\mathcal{E}_t \colon \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ without revealing it;
- the player commits a probability measure $\mu_t \in \mathcal{P}(\mathbb{R}^d)$ as the decision variable;
- the adversary reveals the energy functional, and the player suffers the loss $\mathcal{E}_t(\mu_t)$.

The player needs to decide the next μ_{t+1} based on μ_t and all historical information and aims to minimize cumulative regret from facing the adversary. The main conceptual message we discover in OLT is that optimally transporting the probability measures $\mu_1 \to \mu_2 \to \ldots \to \mu_T$ with respect to an appropriate cost enables small cumulative regret.

To provide a glimpse of the angle that inspired us to connect online learning and OT, we follow a viewpoint taken in Benamou and Brenier [2000] and Otto [2001]. All concepts introduced in this paragraph have rigorous definitions in Section 3.2, so we proceed at a high level here. Mimicking the finite-dimensional case, OT theory provides a way to calculate a type of (sub-)gradient of $\mathcal{E}_t$ over $\mathcal{P}(\mathbb{R}^d)$ when equipped with a certain metric. This (sub-)gradient is used to update $\mu_t \to \mu_{t+1}$. To enrich this analogy, we consider a continuous/infinitesimal time analog. In the finite-dimensional cases (3.1.1) and (3.1.2), the first right-hand side in a continuous time analog would quantify movement according to the ODE $\frac{\mathrm{d}x_t}{\mathrm{d}t} = -\nabla \ell_t(x_t)$ where $\left\| \frac{\mathrm{d}x_t}{\mathrm{d}t} \right\|^2 = \|\nabla \ell_t(x_t)\|^2$. In the infinite-dimensional case, the density ρ_t (of μ_t w.r.t. the Lebesgue measure) evolves according to the PDE $\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \boldsymbol{\xi}_t)$ with a vector field $\boldsymbol{\xi}_t \colon \mathbb{R}^d \to \mathbb{R}^d$ from the "Fréchet subdifferential" $\partial \mathcal{E}_t(\mu_t)$. The infinitesimal transport cost is[1]

$$\left\| \frac{\partial \rho_t}{\partial t} \right\|_{\rho_t}^2 = \inf_{\boldsymbol{\xi} \in L^2(\mu_t; \mathbb{R}^d)} \left\{ \int \|\boldsymbol{\xi}(x)\|^2 \, \mathrm{d}\mu_t(x) \;\; : \;\; \boldsymbol{\xi} \in \partial \mathcal{E}_t(\mu_t) \right\} . \tag{3.1.3}$$

1. The curious reader may identify the above resemblance to the local norm in (3.1.2): here the local Riemannian geometry quantifies the transport cost.

This motivates the *minimal selection principle* of Ambrosio et al. [2005]: among all the possible vector fields that belong to the Fréchet subdifferential, we aim to select one whose kinetic energy (captured by the integral $\int \|\boldsymbol{\xi}(x)\|^2 \, \mathrm{d}\mu_t(x)$) is lowest. Assuming a notion of convexity of $\mathcal{E}_t$ along Riemannian geodesics, we formally show that minimizing the transport cost over time using the minimal selection principle yields the smallest cumulative regret upper bounds in the OLT problem.

Equipped with the minimal selection principle, we propose a novel algorithm in Section 3.3 that we call *minimal selection or exploration (MSoE)* to solve the variational optimization problem in (3.1.3) and, in turn, solve OLT using only zeroth-order partial feedback similar to the bandit setting. To make this algorithm numerically tractable, we use a discrete analog of the minimal selection principle using mean-field approximations. It is noteworthy that our MSoE algorithm works beyond the convex setting, which shows how elegant theory from OT is practically relevant to certain non-convex settings common in robust dynamic resource allocation. Simple toy examples demonstrating the algorithm's empirical performance are discussed in Section 3.7.

A key feature of this chapter is the natural simplicity of its arguments and algorithmic principles that become apparent once the framework connecting online learning and OT is established. Several directions for extending our framework are discussed at the end of the chapter.

Related Work This chapter is at the intersection of two active research fields: online learning and optimal transport. For the former, due to space limits, we cannot do fair justice to credit all contributions properly; see Orabona [2021] for a comprehensive survey. Here, we give a selective overview. Several improvements of online gradient descent (3.1.1) using an adaptive step size have been proposed [Streeter and McMahan, 2010, McMahan and Streeter, 2010] with a further modification via scale-freeness [Orabona and Pál, 2015, Orabona and Pál, 2018]. Expression (3.1.2) can be derived using different principles such

as exponentiated gradient [Kivinen and Warmuth, 1997], online mirror descent (OMD), follow-the-regularized-leader (FTRL) [Shalev-Shwartz and Singer, 2007, Abernethy et al., 2008]. Modifications of FTRL via adaptive regularization have been introduced [van Erven et al., 2011, De Rooij et al., 2014, Orabona and Pál, 2015]. General first-order Riemannian optimization methods have been investigated in Zhang and Sra [2016]. Using martingale tools, a theoretical framework for sequential prediction has been studied in Rakhlin and Sridharan [2014]. The typical regret bound for online learning with K-discrete actions scales as $\sqrt{T \log K}$ in the full information setting, and as $\sqrt{TK \log K}$ in the bandit/partial feedback setting. Therefore, naive generalizations of these bounds to the infinite-dimensional case ($K \to \infty$) result in diverging bounds. It is thus unclear whether $\sqrt{T}$-regret is achievable in the infinite-dimensional case. We employ tools from optimal transport to resolve this issue and obtain optimal bounds.

Beyond well-established analytic results of OT, computational [Cuturi, 2013, Genevay et al., 2016, Altschuler et al., 2017] and statistical [Weed and Bach, 2019, Liang, 2021, Weed and Berthet, 2019, Liang, 2019, Hütter and Rigollet, 2021] aspects have been emerging research areas at the intersection of OT, probability distribution estimation, and numerical sampling. At the same time, OT has been a versatile tool for various applied tasks such as image retrieval [Rubner et al., 2000], computational linguistics [Kusner et al., 2015], and domain adaptation [Courty et al., 2017]. One of the successful applications of OT is generative sampling, such as GANs [Goodfellow et al., 2014]. OT has improved generative sampling in several ways: by modifying GANs using OT-based probability metrics [Arjovsky et al., 2017, Genevay et al., 2018] or by utilizing dual formulations of OT [Seguy et al., 2018, Makkuva et al., 2020]. More recently, further improvements [Bunne et al., 2019, Hur et al., 2024] have been made by considering the Gromov-Wasserstein [Mémoli, 2011], a generalization of OT ideas for identifying isomorphism in metric measure spaces.

Notation We use the following notation throughout the chapter. Let $\mathcal{P}^r(\mathbb{R}^d)$ denote the subset of $\mathcal{P}(\mathbb{R}^d)$ of Borel probability measures that are absolutely continuous with respect to the Lebesgue measure $\mathcal{L}^d$ on $\mathbb{R}^d$. For $\mu \in \mathcal{P}^r(\mathbb{R}^d)$, we write $\mu = \rho \cdot \mathcal{L}^d$ to signify that ρ is a density function of μ with respect to $\mathcal{L}^d$. We let δ_x denote the Dirac measure for $x \in \mathbb{R}^d$ and $\mathbf{i}\colon \mathbb{R}^d \to \mathbb{R}^d$ the identity map $\mathbf{i}(x) = x$ for all $x \in \mathbb{R}^d$. For $x \in \mathbb{R}$, let $[x]_+ = \max(x, 0)$.

3.2 Optimal Transport, Minimal Selection Principle, and Regret Bound

In this section, we present our main theoretical results for OLT. As noted earlier, the key to our analysis is a connection with optimal transport theory. We start by introducing the notion of Wasserstein space to structure the decision space of OLT. Recall from (3.1.1) and (3.1.2) that a notion of difference (transport cost) between two consecutive decisions plays a role in deriving a regret bound. To apply this idea to OLT, we utilize the Wasserstein distance between probability measures (decisions in OLT). Next, we discuss differential calculus over Wasserstein space, which serves as a key building block for deriving a regret bound for OLT. In Wasserstein space, we first define a subdifferential and then select an element associated with the smallest size, measured by the local Riemannian geometry of the Wasserstein space. Such an element, which we call a minimal selection, is defined by a variational problem and serves as a functional gradient. Based on this, we make transparent a strategy to obtain $\sqrt{T}$-regret that generalizes seamlessly to the infinite-dimensional setting. Lastly, we mention a connection between OLT and Wasserstein gradient flow.

3.2.1 Wasserstein Space

One of the most important consequences of optimal transport theory is that we can define a distance between $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ by finding a coupling that gives the smallest transport cost.

Recall that the Wasserstein distance W_2 on $\mathcal{P}(\mathbb{R}^d)$ as $W_2^2(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\gamma(x,y)$. The Wasserstein distance W_2 is indeed a distance over $\mathcal{P}_2(\mathbb{R}^d)$, where $\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}\mu(x) < \infty\}$. From this, we obtain a metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ of probability measures called the Wasserstein space.

One important property of W_2 is that we can find a unique optimal coupling under certain regularity conditions. Here, an optimal coupling is any coupling $\gamma \in \Pi(\mu,\nu)$ associated with the smallest transport cost, that is, $W_2^2(\mu,\nu) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \,\mathrm{d}\gamma(x,y)$. Proposition 2.1 of Villani [2003] tells that $\Pi_o(\mu,\nu)$, the set of all optimal couplings, is always nonempty. The following result states that $\Pi_o(\mu,\nu)$ is a singleton if μ is absolutely continuous with respect to the Lebesgue measure; moreover, such an optimal coupling is concentrated on the graph of some unique map.

Lemma 3.2.1. *Let $\mathcal{P}_2^r(\mathbb{R}^d) = \mathcal{P}_2(\mathbb{R}^d) \cap \mathcal{P}^r(\mathbb{R}^d)$. Given $\mu \in \mathcal{P}_2^r(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists a unique optimal coupling $\gamma \in \Pi(\mu,\nu)$, namely $\Pi_o(\mu,\nu) = \{\gamma\}$. Also, there exists a unique map $\mathbf{t}_\mu^\nu \colon \mathbb{R}^d \to \mathbb{R}^d$ such that γ is concentrated on the graph of $\mathbf{t}_\mu^\nu$, that is,*

$$\gamma(\{(x,y) \in \mathbb{R}^d \times \mathbb{R}^d : y = \mathbf{t}_\mu^\nu(x)\}) = 1 \,.$$

This implies $(\mathbf{t}_\mu^\nu)_{\#}\mu = \nu$ and $W_2^2(\mu,\nu) = \int_{\mathbb{R}^d} \|x - \mathbf{t}_\mu^\nu(x)\|^2 \,\mathrm{d}\mu(x)$.

Remark 3.2.2. This is a part of Brenier's theorem [Brenier, 1991], which contains further properties of $\mathbf{t}_\mu^\nu$ such as monotonicity. See Theorem 2.12 of Villani [2003] for details.

3.2.2 Displacement Convexity and Subdifferential Calculus

Having defined the decision space for OLT, we explore how to minimize a loss function over this space. As noted earlier, the key is finding an object analogous to the gradient in the finite-dimensional setting. Ambrosio et al. [2005] achieves this by rigorously defining notions of convexity and differential calculus on Wasserstein space. We reproduce a few relevant

results from Ambrosio et al. [2005]. Throughout, we consider a functional $\mathcal{E}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$ with a nonempty domain $D(\mathcal{E}) := \{\mu \in \mathcal{P}_2^r(\mathbb{R}^d) : \mathcal{E}(\mu) < \infty\} \neq \emptyset$; for a concise summary, we restrict the domain to $\mathcal{P}_2^r(\mathbb{R}^d)$, see Section 10.1 of Ambrosio et al. [2005] for more details. First, we discuss convexity of $\mathcal{E}$. To import convexity into $\mathcal{P}_2^r(\mathbb{R}^d)$, we need a concept that replaces the concept of 'line segment' in a vector space. McCann [1997] introduces the displacement interpolation for connecting two elements of $\mathcal{P}_2(\mathbb{R}^d)$ and defines the convexity based on it.

Definition 3.2.3 (Displacement interpolation and Convexity)**.** We define the displacement interpolation between $\mu, \nu \in \mathcal{P}_2^r(\mathbb{R}^d)$ as a curve $(\pi_\eta^{\mu\to\nu})_{\eta\in[0,1]}$ in $\mathcal{P}_2^r(\mathbb{R}^d)$ such that

$$\pi_\eta^{\mu\to\nu} := ((1-\eta)\mathbf{i} + \eta \mathbf{t}_\mu^\nu)_{\#}\mu \ .$$

We say $\mathcal{E}$ is displacement convex if $\mathcal{E}(\pi_\eta^{\mu\to\nu}) \leq (1-\eta)\mathcal{E}(\mu) + \eta\mathcal{E}(\nu)$ for all $\mu, \nu \in D(\mathcal{E})$ and $\eta \in [0,1]$.

Remark 3.2.4. To see that the displacement interpolation $(\pi_\eta^{\mu\to\nu})_{\eta\in[0,1]}$ serves as a segment connecting μ and ν one can verify that $\pi_0^{\mu\to\nu} = \mu$, $\pi_1^{\mu\to\nu} = \nu$, and $\pi_\eta^{\mu\to\nu} \in \mathcal{P}_2^r(\mathbb{R}^d)$ for all $\eta \in (0,1)$. See Chapter 7 of Ambrosio et al. [2005] for details.

Next, we introduce the Fréchet subdifferential, which delineates the differentiable structure on the Wasserstein space. Recall that derivatives or subgradients of functionals defined on a Hilbert space are linear functionals over that space. To mimic their features, we define the subdifferential at $\mu \in \mathcal{P}_2^r(\mathbb{R}^d)$ by means of a Hilbert space $L^2(\mu; \mathbb{R}^d)$ as follows.

Definition 3.2.5 (Fréchet Subdifferential)**.** For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, let $L^2(\mu; \mathbb{R}^d)$ be the collection of vector fields $\boldsymbol{\xi}\colon \mathbb{R}^d \to \mathbb{R}^d$ satisfying $\|\boldsymbol{\xi}\|^2_{L^2(\mu;\mathbb{R}^d)} := \int_{\mathbb{R}^d} \|\boldsymbol{\xi}(x)\|^2 \,\mathrm{d}\mu(x) < \infty$. For a lower semi-continuous functional $\mathcal{E}$ and $\mu \in D(\mathcal{E})$, we say that $\boldsymbol{\xi} \in L^2(\mu; \mathbb{R}^d)$ belongs to the Fréchet

subdifferential $\partial\mathcal{E}(\mu)$ if

$$\liminf_{\nu\to\mu} \frac{\mathcal{E}(\nu) - \mathcal{E}(\mu) - \int_{\mathbb{R}^d}\langle \boldsymbol{\xi}(x), \mathbf{t}_\mu^\nu(x) - x\rangle \,\mathrm{d}\mu(x)}{W_2(\nu,\mu)} \geq 0 \,, \tag{3.2.1}$$

where $\liminf_{\nu\to\mu}$ is based on the convergence under W_2.

Definition 3.2.5 is a local definition as one can see from (3.2.1). For a displacement convex functional, however, Fréchet subdifferentials admit a global characterization. Moreover, as the next result shows, we can find a unique member of the Fréchet subdifferential with the smallest norm; see Section 10.1 of Ambrosio et al. [2005].

Lemma 3.2.6 (Minimal Selection Principle)**.** *Let $\mathcal{E}$ be a lower semi-continuous and displacement convex functional, then a vector field $\boldsymbol{\xi}$ belongs to the Fréchet subdifferential $\partial\mathcal{E}(\mu)$ if and only if*

$$\mathcal{E}(\nu) - \mathcal{E}(\mu) \geq \int_{\mathbb{R}^d} \langle \boldsymbol{\xi}(x), \mathbf{t}_\mu^\nu(x) - x\rangle \,\mathrm{d}\mu(x) \quad \forall \nu \in \mathcal{P}_2^r(\mathbb{R}^d) \,.$$

In this case, $\partial\mathcal{E}(\mu)$ has a unique element $\partial^o\mathcal{E}(\mu)$ called the minimal selection *with the smallest norm in the following sense:*

$$\partial^o\mathcal{E}(\mu) = \arg\min\{\|\boldsymbol{\xi}\|^2_{L^2(\mu;\mathbb{R}^d)} : \boldsymbol{\xi} \in \partial\mathcal{E}(\mu)\} \,.$$

As we shall see in Theorem 3.2.10, the minimal selection $\partial^o\mathcal{E}(\mu)$ plays a role analogous to a gradient in providing a regret bound for OLT comparable to (3.1.1) and (3.1.2). We conclude this subsection with two canonical examples of the minimal selection principle.

Example 3.2.7 (Potential Functional)**.** Given a lower semi-continuous function $V\colon \mathbb{R}^d \to (-\infty,\infty]$, we call $\mathcal{V}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty,\infty]$ a potential functional associated with V if

$$\mathcal{V}(\mu) := \int_{\mathbb{R}^d} V(x) \,\mathrm{d}\mu(x)$$

and $\partial^o \mathcal{V}(\mu) = \nabla V$ if V is convex and satisfies some regularity conditions; see Section 10.4 of Ambrosio et al. [2005].

Example 3.2.8 (Interaction Functional)**.** Given a lower semi-continuous function $W\colon \mathbb{R}^d \to [0,\infty)$, we call $\mathcal{W}\colon \mathcal{P}_2(\mathbb{R}^d) \to [0,\infty]$ an interaction functional associated with W if

$$\mathcal{W}(\mu) := \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} W(x-y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)$$

and $\partial^o \mathcal{W}(\mu) = \nabla W * \rho$, where $\mu = \rho \cdot \mathcal{L}^d$, if W is convex and satisfies some regularity conditions; see Section 10.4 of Ambrosio et al. [2005]. Here, $*$ denotes a convolution of vector field ∇W and density ρ.

3.2.3 Regret Bound for OLT

We are ready to derive a regret bound for the OLT problem. The following result, often referred to as Evolution Variational Inequality (EVI) in the PDE literature, clarifies why the minimal selection principle is key to obtaining a regret bound.

Lemma 3.2.9 (EVI)**.** *Let $\mathcal{E}$ be a lower semi-continuous and displacement convex functional. Let $\mu \in D(\mathcal{E})$ and $\boldsymbol{\xi} \in \partial\mathcal{E}(\mu)$. Fix $\eta > 0$ and define $\mu_\eta := (\mathbf{i} - \eta\boldsymbol{\xi})_{\#}\mu$. Then, for any $\nu \in \mathcal{P}_2^r(\mathbb{R}^d)$,*

$$\mathcal{E}(\mu) - \mathcal{E}(\nu) \le \frac{W_2^2(\mu,\nu) - W_2^2(\mu_\eta,\nu)}{2\eta} + \frac{\eta}{2}\int_{\mathbb{R}^d} \|\boldsymbol{\xi}(x)\|^2\,\mathrm{d}\mu(x)\ . \tag{3.2.2}$$

Proof. Recall from Lemma 3.2.1 that $W_2^2(\mu,\nu) = \int_{\mathbb{R}^d} \|x - \mathbf{t}_\mu^\nu(x)\|^2\,\mathrm{d}\mu(x)$. Let $(\mathbf{i} - \eta\boldsymbol{\xi}, \mathbf{t}_\mu^\nu)$ be the map from $\mathbb{R}^d$ to $\mathbb{R}^{2d}$ that maps $x \mapsto (x - \eta\boldsymbol{\xi}(x), \mathbf{t}_\mu^\nu(x))$, then $(\mathbf{i} - \eta\boldsymbol{\xi}, \mathbf{t}_\mu^\nu)_{\#}\mu \in \Pi(\mu_\eta,\nu)$ and thus

$$W_2^2(\mu_\eta,\nu) = \inf_{\gamma\in\Pi(\mu_\eta,\nu)} \int_{\mathbb{R}^d} \|x-y\|^2\,\mathrm{d}\gamma(x,y) \le \int_{\mathbb{R}^d} \|(x - \eta\boldsymbol{\xi}(x)) - \mathbf{t}_\mu^\nu(x)\|^2\,\mathrm{d}\mu(x)\ .$$

Hence,

$$\begin{aligned}\frac{W_2^2(\mu,\nu)-W_2^2(\mu_\eta,\nu)}{2\eta} &\geq \frac{1}{2\eta}\int_{\mathbb{R}^d}\left(\|x-\mathbf{t}_\mu^\nu(x)\|^2-\|x-\eta\boldsymbol{\xi}(x)-\mathbf{t}_\mu^\nu(x)\|^2\right)\mathrm{d}\mu(x)\\ &= \int_{\mathbb{R}^d}\langle\boldsymbol{\xi}(x),x-\mathbf{t}_\mu^\nu(x)\rangle\,\mathrm{d}\mu(x)-\frac{\eta}{2}\int_{\mathbb{R}^d}\|\boldsymbol{\xi}(x)\|^2\,\mathrm{d}\mu(x)\\ &\geq \mathcal{E}(\mu)-\mathcal{E}(\nu)-\frac{\eta}{2}\int_{\mathbb{R}^d}\|\boldsymbol{\xi}(x)\|^2\,\mathrm{d}\mu(x)\quad(\because\text{Lemma 3.2.6})\,.\end{aligned}$$

□

From (3.2.2), it is now obvious that the minimal selection $\boldsymbol{\xi}=\partial^o\mathcal{E}(\mu)$ gives the smallest upper bound. Also, notice the similarity of (3.1.1), (3.1.2), and (3.2.2); the last term on the right-hand side of (3.2.2) corresponds to the norm of a gradient in (3.1.1) and (3.1.2). The minimal selection enables us to import bounding techniques in (3.1.1) and (3.1.2) into the OLT problem. Based on this connection, we propose a strategy to tackle the OLT: for each time t, the player finds the minimal selection $\boldsymbol{\xi}_t=\partial^o\mathcal{E}_t(\mu_t)$ and updates $\mu_{t+1}=(\mathbf{i}-\eta\boldsymbol{\xi}_t)_\#\mu_t$. Using (3.2.2), we obtain the following regret bound.

Theorem 3.2.10 (Regret Bound: Discrete-Time)**.** *Assume $\mathcal{E}_t$ of the OLT is displacement convex for all $t\in[T]$ and $T\in\mathbb{N}$. The player selects $\boldsymbol{\xi}_t=\partial^o\mathcal{E}_t(\mu_t)$ and updates $\mu_{t+1}=(\mathbf{i}-\eta\boldsymbol{\xi}_t)_\#\mu_t$ for each round t, where $\eta>0$ is a fixed step size. Then, for any $\nu\in\mathcal{P}_2^r(\mathbb{R}^d)$,*

$$\sum_{t=1}^T\mathcal{E}_t(\mu_t)-\sum_{t=1}^T\mathcal{E}_t(\nu)\leq\frac{W_2^2(\mu_1,\nu)-W_2^2(\mu_{T+1},\nu)}{2\eta}+\frac{\eta}{2}\sum_{t=1}^T\int_{\mathbb{R}^d}\|\boldsymbol{\xi}_t(x)\|^2\,\mathrm{d}\mu_t(x)\,. \quad (3.2.3)$$

Remark 3.2.11. Under mild conditions such as μ_t and ν are supported on a bounded domain and $\boldsymbol{\xi}_t$ are uniformly bounded, we have $W_2(\mu_t,\nu)\leq D$ and $\|\boldsymbol{\xi}_t\|_{L^2(\mu_t;\mathbb{R}^d)}\leq L$ for all t for some constants $D,L>0$. Then, the right-hand side of (3.2.3) is upper bounded by $\frac{D^2}{2\eta}+\frac{\eta L^2T}{2}$, which becomes $DL\sqrt{T}$ for $\eta=\frac{D}{L\sqrt{T}}$ and yields $\sqrt{T}$-regret.

Remark 3.2.12. We clarify that the EVI is an existing technique in the PDE literature and has already been used for different purposes, for instance, see Dieuleveut et al. [2020] and

Salim et al. [2020]. We emphasize, however, that Theorem 3.2.10 utilizes such a technique for the first time in the context of the online learning given a sequence of adversarial functionals $(\mathcal{E}_t)_{t=1}^T$, and an arbitrary reference measure ν.

3.2.4 Connections to Wasserstein Gradient Flow

We conclude this section with a connection between online learning and Wasserstein gradient flow through a lens of continuous-time OLT, further highlighting why it is natural to employ insights from PDEs. In the infinitesimal limit as $\eta \to 0$, the algorithm solving OLT amounts to a Wasserstein gradient flow. To see this, consider the steepest descent on Wasserstein space according to the energy functional $\mathcal{E}_t$ at time t, $\mu_{t+\eta} := \arg\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{E}_t(\mu) + \frac{1}{2\eta} W_2^2(\mu, \mu_t)$. In the infinitesimal limit, the steepest descent defines continuous-time evolution of probability measures $(\mu_t)_{t\in[0,T]}$, naturally described by PDEs. The following continuity equation is the natural counterpart of discrete update of the steepest descent: letting $\mu_t = \rho_t \cdot \mathcal{L}^d$, solve

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \boldsymbol{\xi}_t) \quad \text{where } \boldsymbol{\xi}_t \in \partial \mathcal{E}_t(\mu_t) \,. \tag{3.2.4}$$

Therefore, it is clear that our online OLT is a discrete analogue of the online Wasserstein gradient flow, with a time-varying energy functional $\mathcal{E}_t$. A solution $(\mu_t)_{t\in[0,T]}$ enjoys the following regret bound; see Section 3.6.1 for the proof.

Theorem 3.2.13 (Regret Bound: Continuous-Time)**.** *Assume $\mathcal{E}_t$ is displacement convex for all $t \in [0,T]$. Under mild regularity assumptions, a solution $(\mu_t)_{t\in[0,T]}$ to* (3.2.4) *satisfies*

$$\int_0^T \mathcal{E}_t(\mu_t)\,\mathrm{d}t - \int_0^T \mathcal{E}_t(\nu)\,\mathrm{d}t \le \frac{W_2^2(\mu_0,\nu) - W_2^2(\mu_T,\nu)}{2} \quad \forall \nu \in \mathcal{P}_2^r(\mathbb{R}^d) \,. \tag{3.2.5}$$

Remark 3.2.14. If $\mathcal{E}_t$ is time-invariant, say $\mathcal{E}_t = \mathcal{E}$ for all $t \in [0,T]$, then (3.2.4) amounts to finding a solution to the standard Wasserstein gradient flow given $\mathcal{E}$ (Chapter 11 of Ambrosio et al. [2005]); minimizing the regret is exactly minimizing a functional $\mathcal{E}$ by solving

a Wasserstein gradient flow. Note that the RHS is time-independent, namely, we have bounded total regret and fast rate as $1/T$.

3.3 Minimal Selection Algorithm with Zeroth-Order Information

In this section, we propose a more practical framework for OLT and methods to solve it based on only zeroth-order, partial feedback. First, let us briefly explain our motivation for studying this practical framework. Recall that the strategy we studied in the previous section to solve OLT was to find the minimal selection $\boldsymbol{\xi}_t = \partial^o \mathcal{E}_t(\mu_t)$ and update μ_t to $(\mathbf{i} - \eta\boldsymbol{\xi}_t)_{\#}\mu_t$ for all t. The EVI gave us a regret bound as in Theorem 3.2.10. However, such a strategy is difficult to implement in practice. Finding $\partial^o \mathcal{E}_t(\mu_t)$ requires the player to access the Fréchet subdifferential $\partial \mathcal{E}_t(\mu_t)$ for each t. Such information is not directly available as nature typically only reveals the loss $\mathcal{E}_t(\mu_t)$ and not the entire collection of vector fields in $\partial^o \mathcal{E}_t(\mu_t)$. For instance, suppose $\mathcal{E}_t$ is a potential functional associated with $V_t \colon \mathbb{R}^d \to \mathbb{R}$, then $\nabla V_t = \partial^o \mathcal{E}_t(\mu_t)$ due to Example 3.2.7. It is not obvious how the player can determine the vector field ∇V_t based only on the zeroth-order information V_t.

As a result, we need a more reasonable setting to implement a strategy based on the minimal selection principle. We achieve this goal by querying only through zeroth-order information, drawing a parallel to the bandit/partial feedback setting with finite arms. To make our discussion concrete, we focus on the case where $\mathcal{E}_t = \mathcal{V}_t$ (potential functional) that occurs frequently in applications; the loss functional is a potential functional associated with some $V_t \colon \mathbb{R}^d \to \mathbb{R}$ for all $t \in \mathbb{N}$. An extension to the interaction functional, another common setting in applications, will be discussed later in this section. Under this framework, nature reveals $\mathcal{V}_t(\mu_t)$ by telling the values of V_t **only** on the support of μ_t and some fixed set of grid points. As we will see shortly, such a setting enables the player to execute a strategy based on the minimal selection principle, thereby obtaining a regret bound based on the EVI. In other words, instead of observing the gradient ∇V_t (minimal selection), the player

approximates it using only zeroth-order information. As such, our formulation bridges the gap between the sophisticated theory in the previous section and its practical realization.

3.3.1 OLT with Zeroth-Order Information and Minimal Selection Algorithm

We formally define our online learning problem. Fix a domain $\Omega \subset \mathbb{R}^d$ and a set $Z \subset \Omega$ of grid points or hubs, say $Z = \{z_1, \dots, z_n\}$, known to the player.[2] At round $t \in \mathbb{N}$,

- the player chooses a discrete measure $\mu_t := \frac{1}{m} \sum_{j=1}^m \delta_{x_j^t}$, where we call $x_1^t, \dots, x_m^t \in \Omega$ decision points,
- nature reveals a lower semi-continuous function $V_t \colon \mathbb{R}^d \to \mathbb{R}$ only on $\{x_j^t\}_{j \in [m]} \cup Z$, namely the zeroth-order information on the player's decision points and the fixed grid points, and the player suffers a loss given as the potential functional associated with V_t, that is, $\mathcal{V}_t(\mu_t) = \frac{1}{m} \sum_{j=1}^m V_t(x_j^t)$ (Example 3.2.7).

Here, we view the grid points as the canonical locations of the domain Ω. Meanwhile, decision points x_j^t are any elements of Ω, not necessarily the grid points. By using the zeroth-information $V_t(z_i)$, the player competes against the grid points, meaning that she aims to choose her decision points that would incur smaller losses than the grid points would do. Accordingly, we consider a regret $\sum_{t=1}^T \mathcal{V}_t(\mu_t) - \sum_{t=1}^T \mathcal{V}_t(\nu)$ for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ supported on Z.

Inspired by the minimal selection principle (Lemma 3.2.6), we propose an algorithm for this learning problem.

2. For instance, z_i's could be hub points (in ridesharing applications), be stochastically sampled based on some reference measure, or be grid points for discretization.

Definition 3.3.1 (Minimal Selection Algorithm)**.** After round $t \in \mathbb{N}$, the player solves

$$\min_{\xi_1,\dots,\xi_m \in \mathbb{R}^d} \quad \frac{1}{m}\sum_{j=1}^{m} \|\xi_j\|^2 \tag{3.3.1}$$

$$\text{subject to} \quad V_t(x_j^t) - V_t(z_i) \le \langle \xi_j, x_j^t - z_i \rangle \quad \forall (i,j) \in [n] \times [m] \,. \tag{3.3.2}$$

After obtaining a minimizer $(\xi_1^t, \dots, \xi_m^t)$, the player updates $x_j^{t+1} = x_j^t - \eta \xi_j^t$, where $\eta > 0$ is a suitable step size.

Remark 3.3.2. Note that this is a convex program. Also, there exists a unique minimizer $(\xi_1^t, \dots, \xi_m^t)$ provided the constraint (3.3.2) is feasible.

Not surprisingly, this algorithm amounts to the minimal section principle in a discrete setting. Recall that $\|\boldsymbol{\xi}\|^2_{L^2(\mu_t;\mathbb{R}^d)} = \frac{1}{m}\sum_{j=1}^m \|\boldsymbol{\xi}(x_j^t)\|^2$ for any $\boldsymbol{\xi} \in L^2(\mu_t;\mathbb{R}^d)$, hence the objective function (3.3.1) corresponds to the norm of an element of the Fréchet subdifferential $\partial \mathcal{V}_t(\mu_t)$. Meanwhile, constraint (3.3.2) amounts to finding a certified subgradient of V_t at x_j^t. Recall from Example 3.2.7 that $\partial^o \mathcal{V}_t(\mu_t) = \nabla V_t$. Essentially, this algorithm aims to find a minimizer $\xi_j^t = \nabla V_t(x_j^t)$. Also, the update rule $x_j^{t+1} = x_j^t - \eta \xi_j^t$ corresponds to $\mu_{t+1} = (\mathbf{i} - \eta \boldsymbol{\xi}_t)_{\#}\mu_t$ in Theorem 3.2.10, where $\boldsymbol{\xi}_t(x_j^t) = \xi_j^t$.

Now, we can utilize the EVI to obtain a regret bound. By adapting the proof of Lemma 3.2.9, we obtain the following result similar to Theorem 3.2.10; see Section 3.5.1 for the proof.

Theorem 3.3.3. *Assume $\Omega = \mathbb{R}^d$. Let μ_t and ξ_j^t be the output of the minimal selection algorithm. If V_t is convex for all $t \in \mathbb{N}$, then for any measure $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ supported on Z and any $\eta > 0$,*

$$\sum_{t=1}^{T} \mathcal{V}_t(\mu_t) - \sum_{t=1}^{T} \mathcal{V}_t(\nu) \le \frac{W_2^2(\mu_1,\nu) - W_2^2(\mu_{T+1},\nu)}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\left(\frac{1}{m}\sum_{j=1}^{m}\|\xi_j^t\|^2\right) \,. \tag{3.3.3}$$

Remark 3.3.4. Remark that the RHS of the above equation is precisely the minimal objective value in (3.3.1). Additionally, if $\Omega \subsetneq \mathbb{R}^d$, the update rule $x_j^{t+1} = x_j^t - \eta \xi_j^t$ may cause $x_j^{t+1} \notin \Omega$. We can prevent this via a projection step. See Section 3.3.3 for details.

Remark 3.3.5. We emphasize that the key of the aforementioned zeroth-order framework is that the decision points can take any values in Ω, allowing the support of μ_t to change continuously on Ω. On the contrary, if we had restricted the support of μ_t to be contained in Z, the decision points would have moved discontinuously over Z. In applications such as real-time dynamic resource allocation, such a discontinuous movement is impractical as one needs to allocate the decision points (resources such as drones or drivers) within a short period of time. In conclusion, our framework is more suitable in practice due to the continuously varying support of μ_t. At the same time, our framework preserves the infinite-dimensional nature.

3.3.2 Beyond Convex Case: Exploration and Regret Bound

If V_t is non-convex, constraint set (3.3.2) might be infeasible. We propose a simple modification of the algorithm using exploration and provide a regret bound that extends beyond the convex case.

Note that we may consider the previous learning problem at each decision point separately. The minimal selection algorithm amounts to solving, for each $j \in [m]$,

$$\min_{\xi_j \in \mathbb{R}^d} \quad \|\xi_j\|^2 \quad \text{subject to} \quad V_t(x_j^t) - V_t(z_i) \leq \langle \xi_j, x_j^t - z_i \rangle \quad \forall i \in [n] \; . \tag{3.3.4}$$

If the above problem is not feasible, we sample a stochastic, isotropic Gaussian vector with a suitable scaling and update x_j^t along this exploration direction.

Definition 3.3.6 (Minimal Selection or Exploration (MSoE) Algorithm)**.** After round $t \in \mathbb{N}$, let S^t be a subset of $[m]$ such that $j \in S^t$ if there is $\xi_j \in \mathbb{R}^d$ satisfying $V_t(x_j) - V_t(z_i) \leq$

$\langle \xi_j, x_j^t - z_i \rangle$ for all $i \in [n]$. For $j \in S^t$, the player solves (3.3.4) and updates $x_j^{t+1} = x_j^t - \eta \xi_j^t$ using the unique solution ξ_j^t to (3.3.4). For $j \notin S^t$, the player samples a Gaussian vector $g_j^t \sim N(0, I_d)$ and updates

$$x_j^{t+1} = x_j^t - \sqrt{\eta \frac{\max_{i \in [n]} [V_t(x_j^t) - V_t(z_i)]_+}{d}} g_j^t \, .$$

We assume all the Gaussian vectors are independent of each other.

In short, we modify the minimal selection algorithm to let the feasible decision points move along minimal selection directions as before, while infeasible decision points fully explore in a random direction. The fact that the discretization step size scales with $\sqrt{\eta}$ is motivated by Euler discretization of Langevin dynamics. The adaptive step size factor $\max_{i \in [n]} [V_t(x_j^t) - V_t(z_i)]_+$ follows the relative potential difference between infeasible decision points and grid points. For the MSoE algorithm, an expected regret bound can be derived that depends on the fraction of infeasible points; see Section 3.5.2 for the proof.

Theorem 3.3.7. *Assume $\Omega = \mathbb{R}^d$. Let μ_t, ξ_j^t, and g_j^t be the output of the MSoE algorithm in Definition 3.3.6. For any V_t and any measure $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ supported on Z,*

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^T \mathcal{V}_t(\mu_t) - \sum_{t=1}^T \mathcal{V}_t(\nu)\right] \leq \frac{W_2^2(\mu_1, \nu)}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathbb{E}\left[\frac{1}{m} \sum_{j \in S^t} \|\xi_j^t\|^2\right] \\ + \frac{3}{2} \sum_{t=1}^T \mathbb{E}\left[\frac{1}{m} \sum_{j \notin S^t} \max_{i \in [n]} [V_t(x_j^t) - V_t(z_i)]_+\right] \, . \end{aligned} \tag{3.3.5}$$

From (3.3.5), we can expect a moderate regret bound provided that the proportion $\frac{|[m] \setminus S^t|}{m}$ of infeasible decision points is small. We prove that this is indeed the case under some assumptions on the sequence of functions V_t. To see this, we first define a feasible region.

Definition 3.3.8. For $t \in \mathbb{N}$, let

$$R^t := \{x \in \Omega : \exists \xi \text{ such that } V_t(x) - V_t(z_i) \leq \langle \xi, x - z_i \rangle \ \ \forall i \in [n]\}$$

and for each $x \in R^t$ define

$$\xi_x = \arg\min_{\xi \in \mathbb{R}^d} \|\xi\|^2 \quad \text{subject to } \ V_t(x) - V_t(z_i) \leq \langle \xi, x - z_i \rangle \ \ \forall i \in [n] \ .$$

Lastly, for $\eta > 0$ let $R^t_\eta := \{x - \eta \xi_x : x \in R^t\}$.

The set R^t contains the points for which the minimal selection algorithm is feasible, hence $j \in S^t$ if and only if $x^t_j \in R^t$. Also, ξ_x is well-defined for $x \in R^t$ as discussed in Remark 3.3.2. Lastly, R^t_η denotes the region obtained by updating points in R^t according to the minimal selection algorithm. Now, we introduce an assumption to control the proportion of infeasible decision points.

Assumption 3.3.9 (Non-expansion condition)**.** There exists $\eta > 0$ so that $R^t_\eta \subseteq R^{t+1}$ for all $t \in \mathbb{N}$.

In other words, this assumption guarantees that any feasible point at some round is still feasible after the update; $j \in S^t$ implies $j \in S^{t+1}$. Hence, the proportion $\frac{|[m] \setminus S^t|}{m}$ of infeasible points does not grow. We impose a further assumption that guarantees this proportion decreases.

Assumption 3.3.10 (Shrinking condition)**.** There exist $\eta > 0$ and $\gamma \in (0, 1)$ such that for all $t \in \mathbb{N}$,

$$\inf_{x \in \Omega \setminus R^t} \Pr_{g \sim N(0, I_d)} \left(x - \sqrt{\eta \frac{\max_{i \in [n]} [V_t(x) - V_t(z_i)]_+}{d}} g \in R^{t+1} \right) \geq \gamma \ . \tag{3.3.6}$$

In short, each infeasible point can be updated to a feasible point by a random direction

with probability at least γ.[3] Combining this assumption with Assumption 3.3.9, we can show that the proportion $\frac{|[m]\backslash S^t|}{m}$ of infeasible decision points shrinks by a factor $1-\gamma$ in each iteration. This observation yields the following regret bound. For simplicity, we assume each V_t is uniformly bounded; see Section 3.5.3 for the proof.

Theorem 3.3.11 (Shrinking fraction of infeasible decision points)**.** *Assume $\Omega = \mathbb{R}^d$. Let μ_t, ξ_j^t, and g_j^t be the output of the MSoE algorithm. If Assumptions 3.3.9 and 3.3.10 are satisfied with parameters $\eta, \gamma > 0$, then for any measure $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ supported on Z,*

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathcal{V}_t(\mu_t) - \sum_{t=1}^{T}\mathcal{V}_t(\nu)\right] \le \frac{W_2^2(\mu_1,\nu)}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{m}\sum_{j\in S^t}\|\xi_j^t\|^2\right] + 3B\gamma^{-1}\frac{|[m]\backslash S^1|}{m}\,, \tag{3.3.7}$$

where $B = \max_{t\in[T]}\sup_{x\in\Omega}|V_t(x)|$.

3.3.3 Further Extensions

We briefly discuss a few extensions of our learning problem and the minimal selection algorithm.

Relaxed Minimal Selection Algorithm We introduce another modification of the minimal selection algorithm using slack variables. After round $t \in \mathbb{N}$, for each $j \in [m]$, the player solves

$$\min_{\xi_j\in\mathbb{R}^d, s_j\ge 0} \quad \|\xi_j\|^2 + \frac{2}{\eta}s_j \tag{3.3.8}$$

$$\text{subject to} \quad -s_j + V_t(x_j^t) - V_t(z_i) \le \langle \xi_j, x_j^t - z_i\rangle \;\; \forall i \in [n]\,. \tag{3.3.9}$$

3. Assumptions 3.3.9 and 3.3.10 are about relations between two adjacent functions V_t and V_{t+1}. We present a simple example of V_t satisfying these assumptions in Section 3.6.2.

After finding minimizers (ξ_j^t, s_j^t), the player updates $x_j^{t+1} = x_j^t - \eta\xi_j^t$, where $\eta > 0$ is a suitable step size. Now that variables are ξ_j and s_j, the constraint (3.3.9) is always feasible. The resulting optimization problem is still convex and admits a unique solution (ξ_j^t, s_j^t). Also, for $j \in S^t$, that is, for a decision point for which the minimal selection algorithm is feasible, one can easily verify that the relaxed minimal selection algorithm boils down to the minimal selection algorithm. Moreover, we can obtain regret bounds similar to (3.3.3) or (3.3.5); see (3.6.1) and (3.6.3) in Section 3.6.3.

Interaction Functional We consider an extension of our problem to a different class of loss functional. Suppose we replace the potential functional with the interaction functional. Nature reveals a lower semi-continuous function $W^t\colon \mathbb{R}^d \to [0,\infty)$ only on $\{x_j^t\}_{j\in[m]} \cup Z$, that is, the player knows $W^t(x_j^t - z_i)$ for all $(i,j) \in [n]\times[m]$. The player suffers a loss given as the interaction functional associated with W^t (Example 3.2.8):

$$\mathcal{W}_t(\mu_t) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} W^t(x-y)\,\mathrm{d}\mu_t(x)\,\mathrm{d}\mu_t(y) = \frac{1}{m^2}\sum_{j,k=1}^{m} W^t(x_j^t - x_k^t)\,. \tag{3.3.10}$$

In this case, we define the minimal selection algorithm as follows: after finishing round $t \in \mathbb{N}$, the player solves

$$\min_{\xi_1,\dots,\xi_m\in\mathbb{R}^d} \quad \frac{1}{m}\sum_{j=1}^{m}\|\xi_j\|^2 \tag{3.3.11}$$

$$\text{subject to} \quad \frac{1}{m}\sum_{k=1}^{m} W^t(x_j^t - x_k^t) - \min_{k\in[n]} W^t(z_i - z_k) \le \langle \xi_j, x_j^t - z_i\rangle \quad \forall (i,j)\in[n]\times[m]\,. \tag{3.3.12}$$

Again, we have a convex program that admits a unique solution $(\xi_1^t,\dots,\xi_m^t)$ provided the constraint (3.3.12) is feasible. Using the EVI, we can derive a regret bound similar to (3.3.3); see Theorem 3.6.3 in Section 3.6.3.

Projection We briefly mention a projection step to ensure $x_j^t \in \Omega$. The idea is to change the update rule from $x_j^{t+1} = x_j^t - \eta \xi_j^t$ to $x_j^{t+1} = P_\Omega(x_j^t - \eta \xi_j^t)$, where $P_\Omega \colon \mathbb{R}^d \to \Omega$ is defined as

$$P_\Omega(x) = \operatorname*{arg\,min}_{\omega \in \Omega} \|x - \omega\| .$$

Recall that the projection P_Ω is well-defined provided Ω is closed and convex. Using this projection step, we can always ensure decision points are contained in Ω. Moreover, this modification does not affect the regret bound (3.3.3); see Theorem 3.6.4 in Section 3.6.3.

3.4 Discussion

In this chapter, we have introduced and studied an infinite-dimensional online learning problem called the Online Learning to Transport (OLT) problem, where the decision variable is a probability measure on a fully nonparametric space, the Wasserstein space. Leveraging tools from optimal transport theory, we have established several theoretical results regarding the OLT problem; equipping the decision space with the Wasserstein distance, we have utilized the evolution variational inequality and the minimal selection principle to derive $\sqrt{T}$-regret bound for the OLT problem. Also, we have studied a practical framework for the OLT problem using zeroth-order information, where we develop a concrete algorithm called the Minimal Selection or Exploration (MSoE) that is numerically tractable and works beyond the convex setting.

Lastly, we mention a few directions for future research. First, theoretical results in Section 3.2.3 may be improved with a stronger notion of convexity. For instance, one might study if $\log(T)$-regret is achievable based on λ-convexity (Definition 9.1.1 of Ambrosio et al. [2005]) of $\mathcal{E}_t$. Another important direction is to modify the minimal selection strategy with adaptive step size η_t. This chapter shows that $\sqrt{T}$-regret is achievable for a fixed step size depending on some universal constants as mentioned Remark 3.2.11. Hence, to remedy this limitation,

we might need to design an adaptive step size η_t possibly in terms of $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}$ (or their norms). Meanwhile, on the practical side, it would be interesting to see how the zeroth-order information framework and the MSoE algorithm work in real-life examples; for instance, we may compare our method with existing methods for real driver allocation datasets, which will shed light on developing even more practical framework and algorithms.

3.5 Proofs of the Regret Bounds in Section 3.3

3.5.1 Proof of Theorem 3.3.3

By convexity of V_t, the constraint set (3.3.2) is always feasible, hence ξ_j^t is well-defined. Let $\nu = \sum_i a_i \delta_{z_i}$ for some $a_1, \ldots, a_n \geq 0$ such that $\sum_{i=1}^n a_i = 1$. We can find an optimal coupling between μ_t and ν, which takes the following form: $\sum_{i=1}^n \sum_{j=1}^m \pi_{ji} \delta_{(x_j^t, z_i)}$, where $\sum_{i=1}^n \pi_{ji} = \frac{1}{m}$ and $\sum_{j=1}^m \pi_{ji} = a_i$. Note that $\sum_{i=1}^n \sum_{j=1}^m \pi_{ji} \delta_{(x_j^{t+1}, z_i)}$ is a coupling between ν and μ_{t+1} as well. Hence,

$$\begin{aligned}
W_2^2(\mu_{t+1}, \nu) - W_2^2(\mu_t, \nu) &\leq \sum_{i=1}^n \sum_{j=1}^m \|x_j^{t+1} - z_i\|^2 \pi_{ji} - \sum_{i=1}^n \sum_{j=1}^m \|x_j^t - z_i\|^2 \pi_{ji} \\
&= \sum_{i=1}^n \sum_{j=1}^m \left(\eta^2 \|\xi_j^t\|^2 - 2\eta \langle \xi_j^t, x_j^t - z_i \rangle \right) \pi_{ji} \\
&\leq \sum_{i=1}^n \sum_{j=1}^m \left(\eta^2 \|\xi_j^t\|^2 - 2\eta (V_t(x_j^t) - V_t(z_i)) \right) \pi_{ji} \\
&= \eta^2 \left(\frac{1}{m} \sum_{j=1}^m \|\xi_j^t\|^2 \right) - 2\eta \left(\mathcal{V}_t(\mu_t) - \mathcal{V}_t(\nu) \right) .
\end{aligned}$$

Therefore, we have

$$\mathcal{V}_t(\mu_t) - \mathcal{V}_t(\nu) \leq \frac{W_2^2(\mu_t, \nu) - W_2^2(\mu_{t+1}, \nu)}{2\eta} + \frac{\eta}{2} \cdot \frac{1}{m} \sum_{j=1}^m \|\xi_j^t\|^2 , \tag{3.5.1}$$

and (3.3.3) follows by summing (3.5.1) iteratively.

3.5.2 Proof of Theorem 3.3.7

We first prove the following:

$$\begin{aligned}\mathbb{E}\left[\mathcal{V}_t(\mu_t) - \mathcal{V}_t(\nu) \mid \mathcal{F}_t\right] \leq \frac{\mathbb{E}\big[W_2^2(\mu_t,\nu) - W_2^2(\mu_{t+1},\nu) \mid \mathcal{F}_t\big]}{2\eta} \\ + \frac{\eta}{2}\cdot\frac{1}{m}\sum_{j\in S^t}\|\xi_j^t\|^2 + \frac{3}{2}\cdot\frac{1}{m}\sum_{j\notin S^t}\max_{i\in[n]}[V_t(x_j^t) - V_t(z_i)]_+ \,,\end{aligned} \tag{3.5.2}$$

where $\mathcal{F}_t$ denotes the σ-field generated by $\{g_j^s : \forall s < t \;\text{ and }\; \forall j \in S^s\}$ for $t > 1$ and $\mathcal{F}_1$ is the trivial σ-field. We write $x_j^{t+1} = x_j^t - v_j^t$, where

$$v_j^t = \sqrt{\eta\frac{\max_{i\in[n]}[V_t(x_j^t) - V_t(z_i)]_+}{d}}\, g_j^t \quad \text{if } \; j \notin S^t$$

and $v_j^t = \eta\xi_j^t$ otherwise. As in the proof of Theorem 3.3.3, let $\sum_{i=1}^n\sum_{j=1}^m \pi_{ji}\delta_{(x_j^t,z_i)}$ be an optimal coupling between μ_t and $\nu = \sum_i a_i\delta_{z_i}$. Again, it is a coupling between ν and μ_{t+1}, hence

$$\begin{aligned}W_2^2(\mu_{t+1},\nu) - W_2^2(\mu_t,\nu) &\leq \sum_{i=1}^n\sum_{j=1}^m \|x_j^{t+1} - z_i\|^2\pi_{ji} - \sum_{i=1}^n\sum_{j=1}^m \|x_j^t - z_i\|^2\pi_{ji} \\ &= \sum_{i=1}^n\sum_{j=1}^m \Big(\|v_j^t\|^2 - 2\langle v_j^t, x_j^t - z_i\rangle\Big)\pi_{ji} \\ &= \frac{1}{m}\sum_{j=1}^m \|v_j^t\|^2 - 2\sum_{i=1}^n\sum_{j=1}^m \langle v_j^t, x_j^t - z_i\rangle\pi_{ji}\,.\end{aligned}$$

Now, taking the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_t]$, we have

$$\mathbb{E}[W_2^2(\mu_{t+1},\nu) - W_2^2(\mu_t,\nu) \mid \mathcal{F}_t] \leq \frac{1}{m}\sum_{j=1}^m \mathbb{E}\Big[\|v_j^t\|^2 \mid \mathcal{F}_t\Big] - 2\sum_{i=1}^n\sum_{j=1}^m \mathbb{E}\Big[\langle v_j^t, x_j^t - z_i\rangle\pi_{ji} \mid \mathcal{F}_t\Big]\,.$$

Now, we upper bound the last two terms. For $j \in S^t$, we have $\mathbb{E}[\|v_j^t\|^2 \mid \mathcal{F}_t] = \eta^2\|\xi_j^t\|^2$ because $v_j^t = \xi_j^t$ is measurable with respect to $\mathcal{F}_t$. Meanwhile, for $j \notin S^t$, since $\mathbb{E}\|g_j^t\|^2 = d$,

$$\mathbb{E}\big[\|v_j^t\|^2 \mid \mathcal{F}_t\big] = \eta \max_{i\in[n]}[V_t(x_j^t) - V_t(z_i)]_+ \,.$$

Hence,

$$\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left[\|v_j^t\|^2 \mid \mathcal{F}_t\right] = \frac{\eta^2}{m}\sum_{j\in S^t}\|\xi_j^t\|^2 + \frac{\eta}{m}\sum_{j\notin S^t}\max_{i\in[n]}[V_t(x_j^t) - V_t(z_i)]_+ \,. \tag{3.5.3}$$

Next, for $j \in S^t$, recall that

$$\langle v_j^t, x_j^t - z_i\rangle \geq \eta\big(V_t(x_j^t) - V_t(z_i)\big) \,.$$

For $j \notin S^t$, since $\mathbb{E}g_j^t = 0$,

$$\mathbb{E}\left[\langle v_j^t, x_j^t - z_i\rangle \mid \mathcal{F}_t\right] = 0 \geq \eta\big(V_t(x_j^t) - V_t(z_i)\big) - \eta\max_{i\in[n]}[V_t(x_j^t) - V_t(z_i)]_+ \,.$$

Hence,

$$\begin{aligned}
&-2\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}\left[\langle v_j^t, x_j^t - z_i\rangle\pi_{ji} \mid \mathcal{F}_t\right] \\
&\leq -2\eta\sum_{i=1}^{n}\sum_{j\in S^t}\big(V_t(x_j^t) - V_t(z_i)\big)\pi_{ji} \\
&\quad - 2\eta\sum_{i=1}^{n}\sum_{j\notin S^t}\big(V_t(x_j^t) - V_t(z_i)\big)\pi_{ji} + 2\eta\sum_{i=1}^{n}\sum_{j\notin S^t}\max_{i\in[n]}[V_t(x_j^t) - V_t(z_i)]_+\pi_{ji} \\
&= -2\eta\sum_{i=1}^{n}\sum_{j=1}^{m}\big(V_t(x_j^t) - V_t(z_i)\big)\pi_{ji} + \frac{2\eta}{m}\sum_{j\notin S^t}\max_{i\in[n]}[V_t(x_j^t) - V_t(z_i)]_+ \\
&= -2\eta\left(\mathcal{V}_t(\mu_t) - \mathcal{V}_t(\nu)\right) + \frac{2\eta}{m}\sum_{j\notin S^t}\max_{i\in[n]}[V_t(x_j^t) - V_t(z_i)]_+ \,.
\end{aligned}$$

Combining this result with (3.5.3) and using $\mathbb{E}\left[\mathcal{V}_t(\mu_t) - \mathcal{V}_t(\nu) \mid \mathcal{F}_t\right] = \mathcal{V}_t(\mu_t) - \mathcal{V}_t(\nu)$, we obtain (3.5.2). Now, taking the expectation to the both sides of (3.5.2) and sum iteratively to obtain (3.3.5).

3.5.3 Proof of Theorem 3.3.11

First, using uniform boundedness of V_t, we have

$$\sum_{j \notin S^t} \mathbb{E}\left[\max_{i \in [n]} [V_t(x_j^t) - V_t(z_i)]_+\right] \leq 2B \sum_{j=1}^{m} 1_{j \notin S^t} \,.$$

Combining this with (3.5.2) and taking the expectation, we have

$$\mathbb{E}\left[\mathcal{V}_t(\mu_t) - \mathcal{V}_t(\nu)\right] \leq \frac{\mathbb{E}\big[W_2^2(\mu_t, \nu) - W_2^2(\mu_{t+1}, \nu)\big]}{2\eta} + \frac{\eta}{2}\mathbb{E}\left[\frac{1}{m}\sum_{j \in S^t} \|\xi_j^t\|^2\right] + \frac{3B}{m}\sum_{j=1}^{m} \mathbb{E}[1_{j \notin S^t}] \,.$$

Summing this over $t \in [T]$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{V}_t(\mu_t) - \sum_{t=1}^{T} \mathcal{V}_t(\nu)\right] \leq \frac{W_2^2(\mu_1, \nu)}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{m}\sum_{j \in S^t} \|\xi_j^t\|^2\right] + \frac{3B}{m}\sum_{t=1}^{T}\sum_{j=1}^{m} \mathbb{E}[1_{j \notin S^t}] \,. \tag{3.5.4}$$

Now, we upper bound the last term on the right-hand side. Since $j \notin S^t$ implies $j \notin S^{t-1}$, we have

$$\begin{aligned}
&\mathbb{E}\left[1_{j\notin S^t} \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[1_{j\notin S^{t-1} \text{ and } j\notin S^t} \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[1_{x_j^t \notin R^t} \mid \mathcal{F}_{t-1}\right] 1_{x_j^{t-1}\in\Omega\setminus R^{t-1}} \\
&= \Pr_{g\sim N(0,I)}\left(x_j^{t-1} - \sqrt{\eta\frac{\max_{i\in[n]}[V_{t-1}(x)-V_{t-1}(z_i)]_+}{d}}g \notin R^t\right) 1_{x_j^{t-1}\in\Omega\setminus R^{t-1}} \\
&\le (1-\gamma) 1_{x_j^{t-1}\in\Omega\setminus R^{t-1}} \\
&= (1-\gamma) 1_{j\notin S^{t-1}} \,,
\end{aligned}$$

where the inequality is due to Assumption 3.3.10. Therefore, taking the conditional expectation recursively according to the filtration, we have $\mathbb{E}[1_{j\notin S^t}] \le (1-\gamma)^{t-1} 1_{j\notin S^1}$. Hence,

$$\sum_{j=1}^m \sum_{t=1}^T \mathbb{E}[1_{j\notin S^t}] \le (1 + (1-\gamma) + \cdots + (1-\gamma)^{T-1}) \sum_{j=1}^m 1_{j\notin S^1} \le \gamma^{-1} |[m]\setminus S^1| \,.$$

Combining this with (3.5.4), we obtain (3.3.7).

3.6 Supplementary Explanations

3.6.1 Proof of Theorem 3.2.13

The proof uses differentiability of Wasserstein distance (Theorem 8.13 of Villani [2003]). If $\boldsymbol{\xi}_t(x)$ is a C^1 function of x and t that is globally bounded, for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, one can calculate that

$$\left.\frac{\mathrm{d}W_2^2(\mu,\mu_t)}{\mathrm{d}t}\right|_{t=s} = 2\int \left\langle x - \mathbf{t}_{\mu_s}^{\mu}(x), \boldsymbol{\xi}_s(x)\right\rangle \mathrm{d}\mu_s(x) \le 2[\mathcal{E}_s(\mu) - \mathcal{E}_s(\mu_s)] \,.$$

The inequality is due to the characterization of Fréchet subdifferential in Lemma 3.2.6. Integrating over $s \in [0, T]$, we obtain the regret bound.

3.6.2 Assumptions 3.3.9 and 3.3.10

First, we illustrate the notions of feasible and infeasible regions for some loss functions in one dimension. In Figure3.1, we plot a w-shape loss function V. For simplicity, we set two global minimizers z_1, z_2 as grid points, and x_1, x_2 as player's decision points. One can see that x_1 is feasible where the minimal selection subdifferential ξ_1 is the slope of the blue line l_1, passing through $(x_1, V(x_1))$, $(z_1, V(z_1))$. The decision point x_2, lying in the barrier between two global minimizers, is infeasible: for any line passing through $(x_2, V(x_2))$, it is impossible to have smaller values than the loss function at both grid points z_1, z_2. The red lines in the figure are two attempts. With this intuition, one can verify that the feasible region is $(-\infty, z_1] \cup [z_2, +\infty)$, whereas the infeasible region is (z_1, z_2).

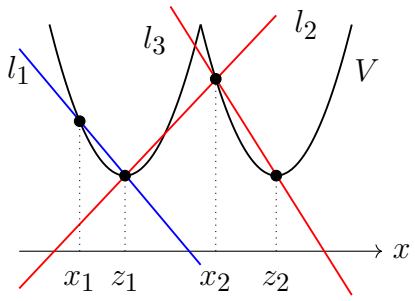


Figure 3.1: An illustration of a w-shape non-convex loss $V : \mathbb{R} \to \mathbb{R}$. Two global minimizers z_1, z_2 serve as grid points. x_1, x_2 are player's decision points.

Next, we showcase a series of w-shape non-convex functions that satisfy Assumptions 3.3.9, 3.3.10.

Lemma 3.6.1. *Consider loss functions*

$$V_t(x) = \begin{cases} a^t(x+1)^2, & \text{if } x < 0, \\ a^t(x-1)^2, & \text{if } x \geq 0, \end{cases}$$

and grid points $-1 = z_1 < z_2 < \cdots < z_n = 1$ *for some arbitrary* n*. If there exists* $\varepsilon > 0$ *such that* $a^t \geq \varepsilon$*,* $a^t\eta < \frac{1}{2}$ *for all* t*, then Assumptions 3.3.9, 3.3.10 hold.*

Proof. The feasible region is $R^t = (-\infty, -1] \cup [1, \infty)$ for all t. We first verify Assumption 3.3.9. Without loss of generality, we consider positive feasible points $x^t = 1 + \delta^t \in R^t$ for some $\delta^t \geq 0$. The minimal selection of subdifferential returns $\xi_x = 2a^t\delta^t$. Therefore $x^{t+1} = x^t - 2\eta a^t\delta^t = 1 + (1 - 2a^t\eta)\delta$. Under the condition $a^t\eta < \frac{1}{2}$, we have $x^{t+1} \geq 1$, i.e., $R^t_\eta \subset R^{t+1}$.

Now, we verify Assumption 3.3.10. For any infeasible $x^t \in [0, 1)$, update by random exploration

$$x^{t+1} = x_t - \sqrt{\eta V_t(x^t)}g \ ,$$

where $g \sim \mathcal{N}(0, 1)$. Then,

$$\begin{aligned} \mathbb{P}(x^{t+1} \in R^{t+1}) &\geq \mathbb{P}(x^t - \sqrt{\eta V_t(x^t)}g \geq 1) \\ &= \mathbb{P}(-\sqrt{\eta a^t}(1 - x^t)g \geq 1 - x^t) \\ &= \mathbb{P}\left(g \leq -\frac{1}{\sqrt{a^t\eta}}\right) \\ &\geq \mathbb{P}\left(g \leq -\frac{1}{\sqrt{\varepsilon\eta}}\right) \ , \end{aligned}$$

where the last inequality is due to the condition $a^t \geq \varepsilon$. Note that the lower bound above holds for any $x \in [0, 1)$. By a symmetric argument, we conclude that Assumption 3.3.10 is satisfied with $\gamma = \mathbb{P}(g \leq -1/\sqrt{\varepsilon\eta})$. □

3.6.3 Results in Section 3.3.3

Here, we provide detailed explanations on the results presented in Section 3.3.3. First, using a similar EVI technique, we derive the following regret bound for the relaxed minimal selection algorithm.

Theorem 3.6.2. *Assume $\Omega = \mathbb{R}^d$. Let μ_t, ξ_j^t, and s_j^t be the output of the relaxed minimal selection algorithm. For any V_t and any measure $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ supported on Z,*

$$\sum_{t=1}^{T} \mathcal{V}_t(\mu_t) - \sum_{t=1}^{T} \mathcal{V}_t(\nu) \leq \frac{W_2^2(\mu_1, \nu) - W_2^2(\mu_{T+1}, \nu)}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \frac{1}{m} \sum_{j=1}^{m} \left(\|\xi_j^t\|^2 + \frac{2s_j^t}{\eta} \right) . \quad (3.6.1)$$

Proof. As in the proof of Theorem 3.3.3, let $\sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ji} \delta_{(x_j^t, z_i)}$ be an optimal coupling between μ_t and $\nu = \sum_i a_i \delta_{z_i}$. Again, it is a coupling between ν and μ_{t+1}, hence

$$\begin{aligned} W_2^2(\mu_{t+1}, \nu) - W_2^2(\mu_t, \nu) &\leq \sum_{i=1}^{n} \sum_{j=1}^{m} \|x_j^{t+1} - z_i\|^2 \pi_{ji} - \sum_{i=1}^{n} \sum_{j=1}^{m} \|x_j^t - z_i\|^2 \pi_{ji} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\eta^2 \|\xi_j^t\|^2 - 2\eta \langle \xi_j^t, x_j^t - z_i \rangle \right) \pi_{ji} \\ &\leq \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\eta^2 \|\xi_j^t\|^2 + 2\eta s_j^t - 2\eta (V_t(x_j^t) - V_t(z_i)) \right) \pi_{ji} \\ &= \frac{\eta^2}{m} \sum_{j=1}^{m} \left(\|\xi_j^t\|^2 + \frac{2s_j^t}{\eta} \right) - 2\eta \left(\mathcal{V}_t(\mu_t) - \mathcal{V}_t(\nu) \right) . \end{aligned}$$

Hence,

$$\mathcal{V}_t(\mu_t) - \mathcal{V}_t(\nu) \leq \frac{W_2^2(\mu_t, \nu) - W_2^2(\mu_{t+1}, \nu)}{2\eta} + \frac{\eta}{2} \cdot \frac{1}{m} \sum_{j=1}^{m} \left(\|\xi_j^t\|^2 + \frac{2s_j^t}{\eta} \right) \quad (3.6.2)$$

(3.6.1) follows by summing (3.6.2) iteratively. □

In fact, we can further upper bound (3.6.2). Note that $\xi_j = 0$ and $s_j = \max_{i \in [n]} [V_t(x_j^t) -$

$V_t(z_i)]_+$ satisfy the constraint (3.3.9), hence the minimizer (ξ_j^t, s_j^t) satisfies

$$\|\xi_j^t\|^2 + \frac{2s_j^t}{\eta} \le \frac{2\max_{i\in[n]}[V_t(x_j^t) - V_t(z_i)]_+}{\eta} .$$

Using this upper bound for $j \notin S^t$, we have

$$\begin{aligned}\mathcal{V}_t(\mu_t) - \mathcal{V}_t(\nu) \le \frac{W_2^2(\mu_t,\nu) - W_2^2(\mu_{t+1},\nu)}{2\eta} \\ + \frac{\eta}{2}\cdot\frac{1}{m}\sum_{j\in S^t}\|\xi_j^t\|^2 + \frac{1}{m}\sum_{j\notin S^t}\max_{i\in[n]}[V_t(x_j^t) - V_t(z_i)]_+ .\end{aligned} \tag{3.6.3}$$

Notice that this upper bound is comparable to (3.5.2). Therefore, the relaxed minimal selection algorithm and the minimal selection or exploration algorithm admit essentially the same upper bound.

Next, we present a regret bound for the case where the loss is given according to the interaction functional.

Theorem 3.6.3. *Assume $\Omega = \mathbb{R}^d$ and consider the game with the loss* (3.3.10) *discussed in Section 3.3.3. Let μ_t and ξ_j^t be the output of the minimal selection algorithm with the constraint* (3.3.12)*. If W^t is convex, for any measure $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ supported on Z,*

$$\sum_{t=1}^T \mathcal{W}_t(\mu_t) - \sum_{t=1}^T \mathcal{W}_t(\nu) \le \frac{W_2^2(\mu_1,\nu) - W_2^2(\mu_{T+1},\nu)}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T\left(\frac{1}{m}\sum_{j=1}^m\|\xi_j^t\|^2\right) . \tag{3.6.4}$$

Proof. As in the proof of Theorem 3.3.3, let $\sum_{i=1}^n\sum_{j=1}^m \pi_{ji}\delta_{(x_j^t, z_i)}$ be an optimal coupling between μ_t and $\nu = \sum_i a_i\delta_{z_i}$. Again, it is a coupling between ν and μ_{t+1}, hence

$$\begin{aligned}W_2^2(\mu_{t+1},\nu) - W_2^2(\mu_t,\nu) &\le \sum_{i=1}^n\sum_{j=1}^m \|x_j^{t+1} - z_i\|^2\pi_{ji} - \sum_{i=1}^n\sum_{j=1}^m\|x_j^t - z_i\|^2\pi_{ji} \\ &= \sum_{i=1}^n\sum_{j=1}^m\left(\eta^2\|\xi_j^t\|^2 - 2\eta\langle\xi_j^t, x_j^t - z_i\rangle\right)\pi_{ji} .\end{aligned}$$

Using (3.3.12),

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m} -2\eta\langle \xi_j^t, x_j^t - z_i\rangle \pi_{ji} &\le \sum_{i=1}^{n}\sum_{j=1}^{m} -2\eta \left(\frac{1}{m}\sum_{k=1}^{m} W^t(x_j^t - x_k^t) - \min_{k\in[n]} W^t(z_i - z_k)\right)\pi_{ji} \\ &= -2\eta\left(\frac{1}{m^2}\sum_{j,k=1}^{m} W^t(x_j^t - x_k^t) - \sum_{i=1}^{n} a_i \cdot \min_{k\in[n]} W^t(z_i - z_k)\right) \\ &\le -2\eta\left(\frac{1}{m^2}\sum_{j,k=1}^{m} W^t(x_j^t - x_k^t) - \sum_{i=1}^{n} a_i \sum_{k=1}^{n} a_k W^t(z_i - z_k)\right) \\ &= -2\eta\left(\mathcal{W}_t(\mu_t) - \mathcal{W}_t(\nu)\right) \ .\end{aligned}$$

Therefore, we have

$$\mathcal{W}_t(\mu_t) - \mathcal{W}_t(\nu) \le \frac{W_2^2(\mu_t,\nu) - W_2^2(\mu_{t+1},\nu)}{2\eta} + \frac{\eta}{2}\cdot\frac{1}{m}\sum_{j=1}^{m}\|\xi_j^t\|^2 \ , \tag{3.6.5}$$

and (3.6.4) follows by summing (3.6.5) iteratively. □

Lastly, we prove that the projection step does not affect the regret bound.

Theorem 3.6.4. *Assume Ω is closed and convex. Let μ_t and ξ_j^t be the output of the minimal selection algorithm with a modified update rule $x_j^{t+1} = P_\Omega(x_j^t - \eta\xi_j^t)$. If V_t is convex, for any measure $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ supported on Z,*

$$\sum_{t=1}^{T}\mathcal{V}_t(\mu_t) - \sum_{t=1}^{T}\mathcal{V}_t(\nu) \le \frac{W_2^2(\mu_1,\nu) - W_2^2(\mu_{T+1},\nu)}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\left(\frac{1}{m}\sum_{j=1}^{m}\|\xi_j^t\|^2\right) \ .$$

Proof. As in the proof of Theorem 3.3.3, let $\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ji}\delta_{(x_j^t, z_i)}$ be an optimal coupling between μ_t and $\nu = \sum_i a_i\delta_{z_i}$. Again, it is a coupling between ν and μ_{t+1}, hence

$$W_2^2(\mu_{t+1},\nu) - W_2^2(\mu_t,\nu) \le \sum_{i=1}^{n}\sum_{j=1}^{m}\|x_j^{t+1} - z_i\|^2\pi_{ji} - \sum_{i=1}^{n}\sum_{j=1}^{m}\|x_j^t - z_i\|^2\pi_{ji} \ .$$

One property of the projection P_Ω is that $\|x_j^{t+1} - \omega\| \leq \|x_j^t - \eta\xi_j^t - \omega\|$ for all $\omega \in \Omega$. Thus,

$$\begin{aligned}
W_2^2(\mu_{t+1}, \nu) - W_2^2(\mu_t, \nu) &\leq \sum_{i=1}^n \sum_{j=1}^m \|x_j^{t+1} - z_i\|^2 \pi_{ji} - \sum_{i=1}^n \sum_{j=1}^m \|x_j^t - z_i\|^2 \pi_{ji} \\
&\leq \sum_{i=1}^n \sum_{j=1}^m \|x_j^t - \eta\xi_j^t - z_i\|^2 \pi_{ji} - \sum_{i=1}^n \sum_{j=1}^m \|x_j^t - z_i\|^2 \pi_{ji} \\
&= \sum_{i=1}^n \sum_{j=1}^m \left(\eta^2 \|\xi_j^t\|^2 - 2\eta \langle \xi_j^t, x_j^t - z_i \rangle\right) \pi_{ji} \\
&\leq \eta^2 \left(\frac{1}{m} \sum_{j=1}^m \|\xi_j^t\|^2\right) - 2\eta \left(\mathcal{V}_t(\mu_t) - \mathcal{V}_t(\nu)\right) \ ,
\end{aligned}$$

where the last inequality directly follows as in the proof of Theorem 3.3.3. Therefore, the regret bounds in Theorem 3.3.3 still hold. □

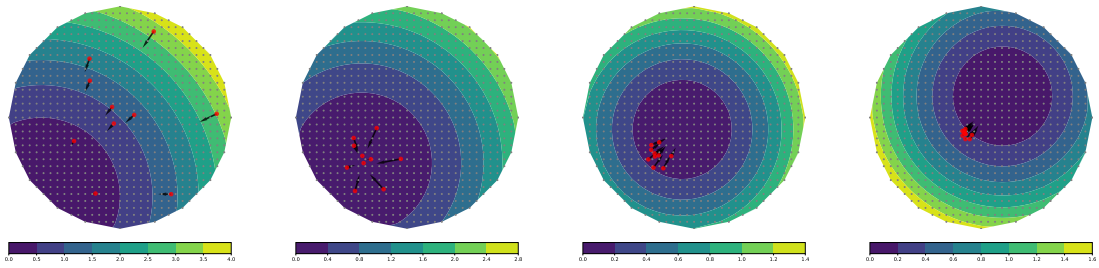
Remark 3.6.5. We can apply this projection step to the MSoE algorithm and the relaxed minimal selection algorithm as well; we modify their update rules by combining them with P_Ω. Then, the regret bounds in Theorems 3.3.7 and 3.6.2 still hold.

3.7 Simulations

In this section, we examine the empirical performance of the minimal selection and the MSoE algorithms. Here, we consider a toy example where $\Omega = \{x \in \mathbb{R}^2 : \|x\| \leq 1\}$, $m = 10$, and Z consists of uniform grid points of Ω with $n = 797$, obtained by forming uniform grid points of $[-1, 1]^2$ and choose a subset of included in Ω. Code for reproducing all the results below is provided in the supplementary material.

Convex Case First, we consider a simple quadratic function $V_t(x) = \|x - u_t\|^2$, where $u_1 = (-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ and $u_t = u_1 + (0.15t, 0.15t)$. For better understanding, we may imagine a situation, where we deploy m drones to track a moving target u_t using the signals V_t (distances to the target) captured at Z (fixed stations with sensors) and $\{x_j^t\}_{j \in [m]}$ (drones

with mobile sensors). Figure 3.2 shows how the minimal selection algorithm moves the decision points. At $t = 1$ (Figure 3.2(a)), the decision points, which are randomly initialized, start moving towards the darker region based on ξ_j^t's from the minimal selection algorithm. In this simple toy example, we can see that the decision points quickly gather around the minimum of V_t (the target) and follow it.

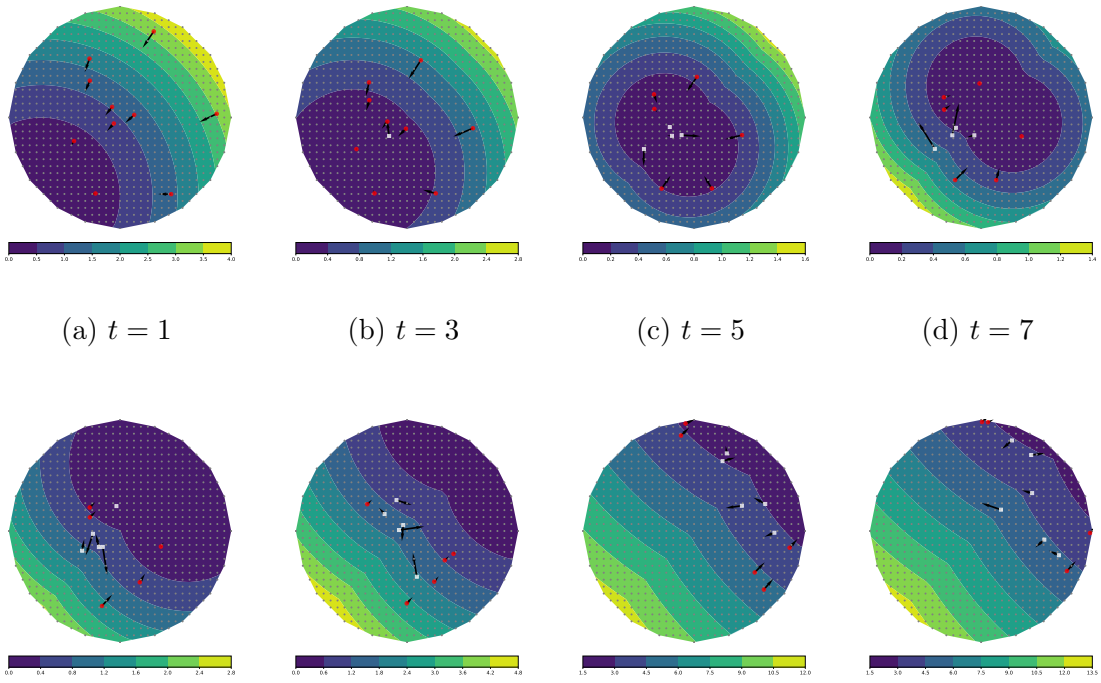


(a) $t = 1$ (b) $t = 3$ (c) $t = 5$ (d) $t = 7$

Figure 3.2: Minimal selection algorithm for convex V_t with step size $\eta = 0.2$. The gray dots and the red circles denote Z and $\{x_j^t\}_{j\in[m]}$, respectively. The black solid arrows show ξ_j^t's. The contour regions represent the level of V_t (darker = smaller as shown in the horizontal colorbars).

Non-Convex Case As a simple non-convex example, consider $V_t(x) = \min\{\|x-u_t\|^2, \|x-v_t\|^2\}$, where $u_1 = v_1 = (-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$, $u_t = u_1 + (0.165t, 0.11t)$, and $v_t = v_1 + (0.11t, 0.165t)$. As in the convex case, we may interpret u_t and v_t as moving targets, while the signal is the distance to a closer target. Figure 3.3 shows how the MSoE algorithm works. As opposed to the convex case, we can check that there are infeasible decision points ($j \notin S^t$ as in Definition 3.3.6) after $t = 3$. The MSoE algorithm let such points move along random directions, thereby continuing the tracking situation; the plain minimal selection would have ended up stopping all the decision points, losing the targets. Although infeasible points might get far away from the targets as in Figure 3.3(e) and Figure 3.3(f), once they get into a feasible region, they start moving again towards the darker area based on ξ_j^t's from the

minimal selection. Figure 3.3(g) and Figure 3.3(h) show that all the decision points somehow get closer to the darker area after repeating the minimal selection and exploration.



(a) $t = 1$ (b) $t = 3$ (c) $t = 5$ (d) $t = 7$

(e) $t = 9$ (f) $t = 12$ (g) $t = 18$ (h) $t = 19$

Figure 3.3: MSoE algorithm for non-convex case with step size $\eta = 0.05$. The red circles denote the feasible decision points ($j \in S^t$) and the white squares denote the infeasible decision points ($j \notin S^t$).

CHAPTER 4

REVERSIBLE GROMOV-MONGE SAMPLER FOR SIMULATION-BASED INFERENCE

4.1 Introduction

One of the central tasks in statistics is to model and sample from a multi-dimensional probability distribution. Classic statistics approaches this problem by fitting a model to the target distribution and then sampling from a fitted model via Markov Chain Monte Carlo (MCMC) techniques. Although such model-based methods are widely used, MCMC sampling often entails several technicalities. Beyond diagnosing whether the chain mixes, obtaining i.i.d. samples from MCMC methods is complex as one has to control correlations among successive samples or run parallel chains.

An alternative approach available in statistics, reserved for the one-dimensional case, is usually referred to as the (inverse) *transform sampling*. Such an approach circumvents the calling for a parametric or nonparametric density and directly designs a sampler by transforming a simple uniform distribution. The idea is simple: one can transform a uniform measure $\mu = \mathrm{Unif}([0,1])$ to any one-dimensional target probability measure ν leveraging the following monotonic transformation $T\colon [0,1] \to \mathbb{R}$ called the inverse Cumulative Distribution Function (CDF),

$$T(x) = \inf\{y \in \mathbb{R} \ : \ \nu((-\infty, y]) \geq x\} \, . \tag{4.1.1}$$

Recall that the pushforward measure $T_{\#}\mu$ is defined by $T_{\#}\mu(S) = \mu(\{x : T(x) \in S\})$ for any Borel set $S \subseteq \mathbb{R}$. Then, one can easily check that $T_{\#}\mu = \nu$; namely, with a draw from the one-dimensional uniform distribution $x \sim \mu$, the transformed sample $T(x)$ has the target probability distribution ν.

The transform sampling idea can be extended to the multi-dimensional setting: given a target probability measure ν supported on $\mathcal{Y}$, one can specify a probability measure μ on $\mathcal{X}$, which is easy to sample from such as a multivariate Gaussian, and then find a measurable map $T\colon \mathcal{X} \to \mathcal{Y}$ such that $T_{\#}\mu = \nu$, where the pushforward measure $T_{\#}\mu$ is defined analogously to the one-dimensional case above. Such a map T—called a *transport map* from μ to ν—transforms i.i.d. samples from μ into i.i.d. samples from ν. Over the past few years, the generative modeling literature in machine learning has been actively employing such transform sampling ideas by identifying $T_{\#}\mu = \nu$ through the following minimization:

$$\min_{T \in \mathcal{F}} \mathcal{L}(T_{\#}\mu, \nu) \ , \tag{4.1.2}$$

where $\mathcal{F}$ is a class of maps from $\mathcal{X}$ to $\mathcal{Y}$ parametrized by neural networks and $\mathcal{L}$ measures certain discrepancies between two distributions. Different choices of $\mathcal{L}$ have led to various models such as the Jensen-Shannon divergence for Generative Adversarial Networks (GANs) [Goodfellow et al., 2014], the Wasserstein-1 distance for Wasserstein-GAN [Arjovsky et al., 2017], and the Maximum Mean Discrepancy (MMD) for MMD-GAN [Dziugaite et al., 2015, Li et al., 2015]. One caveat is that there can be infinitely many transport maps from μ to ν; for instance, when $\mu = \nu = \mathrm{Unif}([0,1])$, define $T\colon [0,1] \to [0,1]$ by $T(x) = |2x - 1|$, then the n-fold compositions of T are valid transport maps for all $n \in \mathbb{N}$. In other words, finding a map T satisfying $T_{\#}\mu = \nu$ is an over-identified problem, where (4.1.2) may have infinitely many minimizers. Though all minimizers are equivalent in terms of transform sampling, not all are equally preferred in light of Occam's razor principle: one wishes to select simple, desirable transport maps among the over-identified set $\{T : T_{\#}\mu = \nu\}$.

Inductive biases tackle the aforementioned over-identified problem by restricting the search to transport maps with desirable properties. In this context, meaningful progress has been made based on Optimal Transport (OT) theory [Taghvaei and Jalali, 2019, Makkuva et al., 2020]. The OT theory aims to identify an optimal transformation T, quantified by

the transportation cost of moving mass from μ to ν; for instance, when μ and ν lie in the same space $\mathbb{R}^d$, each transport map T is associated with the transport cost $C(T) := \int_{\mathbb{R}^d} \|x - T(x)\|^2 \, \mathrm{d}\mu(x)$. Brenier [1991] proved that, under mild regularity conditions, there exists a unique minimizer $T^\star$ of C among all transport maps, namely,

$$T^\star = \underset{T_\# \mu = \nu}{\arg\min} \, C(T) \; . \tag{4.1.3}$$

More importantly, $T^\star$ is the gradient of some convex function. On the one hand, Brenier's result extends the one-dimensional (inverse) transform sampling to the multi-dimensional case. When $d = 1$ and $\mu = \mathrm{Unif}([0, 1])$, the inverse CDF map in (4.1.1) turns out to be exactly $T^\star$; for $d > 1$, the multi-dimensional map $T^\star \colon \mathbb{R}^d \to \mathbb{R}^d$ is the gradient of a convex function, generalizing monotonic functions on the real line to multi-dimensions. On the other hand, Brenier's result naturally initiates an inductive bias in transform sampling: instead of searching any transport map, one may find $T^\star$, the optimal one with the smallest cost. To contrast this with the plain transform sampling (4.1.2), let us rewrite (4.1.3) using a suitable Lagrangian multiplier $\lambda > 0$ to enforce the equality constraint $T_\# \mu = \nu$:

$$\min_{T \in \mathcal{F}} C(T) + \lambda \cdot \mathcal{L}(T_\# \mu, \nu) \; . \tag{4.1.4}$$

Now, we can see that (4.1.4) incorporates an additional objective function of T—the transport cost—in (4.1.2), thereby introducing an inductive bias towards the optimal transport map that achieves the minimum of C. Moreover, the Lagrangian provides an implementable formulation in practice to leverage OT ideas in generative modeling.

Such an OT-based approach, however, can be cumbersome in practice if the target ν is a high-dimensional embedding of some low-dimensional distribution. For instance, let ν be the distribution of handwritten digit images from the MNIST data set on $\mathbb{R}^{784}$.[1] To use the

1. Images are normalized and fit into a 28×28 pixel bounding box ($\mathbb{R}^{28 \times 28} \equiv \mathbb{R}^{784}$) [LeCun et al., 1998].

above OT-based approach, one must choose μ on $\mathbb{R}^{784}$ and find a map $T\colon \mathbb{R}^{784} \to \mathbb{R}^{784}$. However, the support of ν is intrinsically low-dimensional (roughly $\mathbb{R}^{15}$ as in Facco et al. [2017]); hence, other transform samplers with $\mathcal{X} = \mathbb{R}^{15}$ yielding $T\colon \mathbb{R}^{15} \to \mathbb{R}^{784}$ are more efficient than the OT-based method in terms of estimating T and computing $T(X)$ for $X \sim \mu$.

We propose and study a new transform sampler combining the best of both worlds: it introduces beneficial inductive biases like the OT approach, while operating when $\mathcal{X}$ and $\mathcal{Y}$ are heterogeneous spaces. The key to our approach is to utilize a notion of isomorphism and the Gromov-Wasserstein (GW) distance between μ and ν. Given two cost functions $c_{\mathcal{X}}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $c_{\mathcal{Y}}\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the GW distance [Mémoli, 2011, Chowdhury and Mémoli, 2019] is

$$\mathrm{GW}(\mu, \nu) := \inf_{\gamma \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} \left(c_{\mathcal{X}}(x, x') - c_{\mathcal{Y}}(y, y')\right)^2 \mathrm{d}\gamma\,(x, y)\,\mathrm{d}\gamma\,(x', y') \right)^{1/2}, \tag{4.1.5}$$

where $\Pi(\mu, \nu)$ is the set of couplings between μ and ν. GW aims to match the cost functions defined on two heterogeneous spaces, intending to identify an isomorphism, namely, a transport map T such that $c_{\mathcal{X}}(x, x') = c_{\mathcal{Y}}(T(x), T(x'))$ for all $x, x' \in \mathcal{X}$. Inspired by these, one can define the following objective function of T to replace the transport cost C:

$$Q(T) := \int_{\mathcal{X}} \int_{\mathcal{X}} (c_{\mathcal{X}}(x, x') - c_{\mathcal{Y}}(T(x), T(x')))^2 \,\mathrm{d}\mu\,(x)\,\mathrm{d}\mu\,(x')\,. \tag{4.1.6}$$

One way to design a transform sampler with inductive biases toward isomorphisms (the target when $\min_T Q(T) = 0$) is to utilize the Lagrangian form (4.1.4), but with Q instead of C. It turns out, however, that the objective function Q—which is quadratic in μ—results in several modeling subtleties when considering its plug-in estimation based on finite i.i.d. samples from μ and ν, which might not be favorable in practice (see Remark 4.3.3). To circumvent such issues, we develop a transform sampler with a different objective function,

which we will detail in Section 4.3.

Organization The rest of the chapter is organized as follows. First, we briefly review other related studies omitted in the discussion above; Section 4.2 briefly outlines some preliminary background on the Gromov-Wasserstein distance. Then, Section 4.3 introduces the primary methodology of this chapter, where we develop a new transform sampler based on the new notion called the Reversible Gromov-Monge (RGM) distance. Section 4.4 delineates the main theoretical result, providing the statistical rate of convergence for the plug-in estimation. Section 4.5 discusses further theoretical perspectives on the new sampler. Synthetic and real-world examples showcasing the effectiveness of the new sampler are demonstrated in Section 4.6 as a proof of concept. The supplementary material collects details of the results in Sections 4.4, 4.5, and 4.6 along with relevant discussions.

Notation We use the following notation throughout the chapter. Given a set $\mathcal{X}$ and a function $f\colon \mathcal{X} \to \mathbb{R}$, let $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$ denote the sup norm. For a sequence of numbers $a(n), b(n) \in \mathbb{R}$, we use $a(n) \precsim b(n)$ to denote the relationship that $a(n)/b(n) \leq C$ for all $n \in \mathbb{N}$ with some universal constant $C > 0$.

4.1.1 Related Literature

Inferring the underlying probability distributions from data has been a central problem in statistics and unsupervised machine learning since the invention of histograms by Pearson a century ago. Classic mathematical statistics explicitly models the density function in a parametric or a nonparametric way [Silverman, 1986], and studies the minimax optimality of directly estimating such density functions [Stone, 1982]. It is also unclear how to proceed to sample from a possibly improper[2] density estimator, even with an optimal estimator at hand. One may employ Markov Chain Monte Carlo (MCMC) techniques for sampling from

2. Here we mean that the estimated density is not always non-negative and integrates to one.

specific models. However, on the computational front, it is highly non-trivial how to ensure the mixing properties of MCMC for a designed sampler [Robert and Casella, 1999].

A recent trend in unsupervised machine learning is to learn complex, high-dimensional distributions via (deep) generative models, either explicitly by parametrizing the sufficient statistics of the exponential families [Doersch, 2016, Kingma and Welling, 2013], or implicitly by parametrizing the pushforward map transporting distributions [Dziugaite et al., 2015, Goodfellow et al., 2014], with a focus on tractability in computation. Surprisingly, though lacking theoretical underpinning and optimality, generative models perform well empirically in large-scale applications where classical statistical procedures are destined to fail. There has been a growing literature on understanding distribution estimation with the implicit framework, with more general metrics and target distribution classes, to name a few, Muandet et al. [2017], Li et al. [2015], Dziugaite et al. [2015] on MMDs, Sriperumbudur et al. [2012], Liang [2019] on integral probability metrics, and Mroueh et al. [2017], Arora et al. [2017], Liang [2021], Singh and Póczos [2018], Bai et al. [2018], Weed and Berthet [2019], Lei et al. [2019], Chen et al. [2020] on generative adversarial networks. Last but not least, we emphasize that an alternative implicit distribution estimation approach using the simulated method of moments has been formulated in the econometrics literature since McFadden [1989], Pakes and Pollard [1989] and Gouriéroux and Monfort [1997].

Originally introduced as a tool for comparing objects in computer graphics, analytic properties of the Gromov-Wasserstein distance have been studied extensively [Mémoli, 2011, Sturm, 2012]; the most important one is that it defines a distance between metric measure spaces, namely, metric spaces endowed with probability measures. Since many real-world data sets can be modeled as metric measure spaces, the GW distance has been utilized in various problems such as shape correspondence [Solomon et al., 2016], graph matching [Xu et al., 2019], and protein comparison [Gellert et al., 2019]. Certain statistical aspects of comparing metric measure spaces have been studied in Brécheteau [2019], Weitkamp et al.

[2024].

Computation of the GW distance amounts to a relaxation of the quadratic assignment problem [Koopmans and Beckmann, 1957]; both are known to be NP-hard [Çela, 1998] in the worst case. Several approaches have been proposed for the approximate computation of the GW distance. Mémoli [2011] studies lower bounds on the GW distance that are easier to compute, which are further studied by Mémoli and Needham [2022], Hur and Khoo [2023] Peyré et al. [2016] adds an entropic regularization term to the GW distance, which leads to a fast iterative algorithm; Scetbon et al. [2021] further modifies this by imposing a low-rank constraint on couplings. Titouan et al. [2019] proposes the sliced Gromov-Wasserstein distance defined by integrating GW distances over one-dimensional projections. Last but not least, recent papers [Xu et al., 2019, Blumberg et al., 2020, Chowdhury et al., 2021] study scalable partitioning schemes to approximately compute GW distances.

GW distances have been utilized as a discrepancy measure in the generative modeling [Bunne et al., 2019]; roughly speaking, they use GW distances as $\mathcal{L}$ in (4.1.2), which is significantly different from the method developed in this chapter.

4.2 Gromov-Wasserstein and Gromov-Monge Distances

Recall from Section 1.1 that OT problems can be defined between arbitrary Polish probability spaces. In practice, however, it is unclear how to design a function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ to represent meaningful cost associated with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ in two heterogeneous spaces. For instance, if $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}^q$ with $p \neq q$, there is no simple choice for a cost function c over $\mathbb{R}^p \times \mathbb{R}^q$. As a result, classic OT theory (including Brenier's result) cannot be directly used for comparing heterogeneous Polish probability spaces.

Mémoli's pioneering work [Mémoli, 2011] resolved this issue by considering a quadratic

objective function of γ:

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\gamma\,(x,y) \Rightarrow \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y'))^2\,\mathrm{d}\gamma\,(x,y)\,\mathrm{d}\gamma\,(x',y')\ ,$$

where $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are defined over $\mathcal{X}\times\mathcal{X}$ and $\mathcal{Y}\times\mathcal{Y}$, respectively. For instance, one can specify $c_{\mathcal{X}} = d_{\mathcal{X}}$ and $c_{\mathcal{Y}} = d_{\mathcal{Y}}$. Rather than considering a unit cost corresponding to each pair $(x,y)\in\mathcal{X}\times\mathcal{Y}$, we associate two pairs $(x,y),(x',y')\in\mathcal{X}\times\mathcal{Y}$ with the discrepancy of intra-space quantities $c_{\mathcal{X}}(x,x')$ and $c_{\mathcal{Y}}(y,y')$. In summary, by switching from the integration $\mathrm{d}\gamma$ to the double integration $\mathrm{d}\gamma\,\mathrm{d}\gamma$, we no longer need an otherwise inter-space quantity $c\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+$. Therefore, we can always define this objective function whenever we have proper $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ in each individual space, leading to the following definition.

Definition 4.2.1. A triple $(\mathcal{X},\mu,c_{\mathcal{X}})$ is called a network space if $(\mathcal{X},\mu)$ is a Polish probability space such that $\mathrm{supp}(\mu)=\mathcal{X}$ and $c_{\mathcal{X}}\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is measurable. The Gromov-Wasserstein distance between network spaces $(\mathcal{X},\mu,c_{\mathcal{X}})$ and $(\mathcal{Y},\nu,c_{\mathcal{Y}})$ is defined as

$$\mathrm{GW}(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y'))^2\,\mathrm{d}\gamma\,(x,y)\,\mathrm{d}\gamma\,(x',y')\right)^{1/2}\ .$$

Remark 4.2.2. We adopt the network space definition introduced in Chowdhury and Mémoli [2019]. A network space $(\mathcal{X},\mu,c_{\mathcal{X}})$ is called a metric measure space if $c_{\mathcal{X}} = d_{\mathcal{X}}$ as introduced in Mémoli [2011] and Sturm [2012]. In short, a network space is a generalization of a metric measure space.

Like the Wasserstein distance, the GW distance has metric properties; it satisfies symmetry and the triangle inequality, and $\mathrm{GW}(\mu,\nu)=0$ if $(\mathcal{X},\mu,c_{\mathcal{X}})=(\mathcal{Y},\nu,c_{\mathcal{Y}})$. However, the converse of this last statement does not hold in general: for its validity, a suitable equivalence relation needs to be defined on the collection of network spaces.

Definition 4.2.3. Network spaces $(\mathcal{X},\mu,c_{\mathcal{X}})$ and $(\mathcal{Y},\nu,c_{\mathcal{Y}})$ are strongly isomorphic if there

exists $T \in \mathcal{T}(\mu, \nu)$ such that $T \colon \mathcal{X} \to \mathcal{Y}$ is bijective and $c_{\mathcal{X}}(x, x') = c_{\mathcal{Y}}(T(x), T(x'))$ for all $x, x' \in \mathcal{X}$. In this case, we write $(\mathcal{X}, \mu, c_{\mathcal{X}}) \cong (\mathcal{Y}, \nu, c_{\mathcal{Y}})$ and such a transport map T is called a strong isomorphism.

One can check that $\cong$ is indeed an equivalence relation on the collection of network spaces. The following theorem states that the GW distance satisfies all metric axioms on the quotient space—under the equivalence relation $\cong$—of metric measure spaces.

Theorem 4.2.4 (Lemma 1.10 of Sturm [2012])**.** *Let $\mathcal{M}$ be the collection of all metric measure spaces. Then, GW satisfies the three metric axioms on $\mathcal{M}/_{\cong}$, the collection of all equivalence classes of $\mathcal{M}$ induced by $\cong$.*

Recall that the Monge problem is a restricted version of the Kantorovich problem with an additional constraint that couplings are given by a transport map; replacing $\Pi(\mu, \nu)$ in the Kantorovich problem with $\Pi_{\mathcal{T}}$ yields the Monge problem. Imposing the same constraint on the definition of GW leads to the Gromov-Monge distance.

Definition 4.2.5. The Gromov-Monge distance between spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ is defined as

$$\mathrm{GM}(\mu, \nu) = \inf_{T \in \mathcal{T}(\mu, \nu)} \left(\int_{\mathcal{X}} \int_{\mathcal{X}} (c_{\mathcal{X}}(x, x') - c_{\mathcal{Y}}(T(x), T(x')))^2 \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(x') \right)^{1/2}.$$

Loosely speaking, computing GM amounts to finding a transport map T such that $c_{\mathcal{X}}(x, x')$ best matches $c_{\mathcal{Y}}(T(x), T(x'))$ on average; we can view such a map T as a surrogate for an isomorphism. See Section 2.4 of Mémoli and Needham [2022] for more details of GM. Observe that the objective function in the definition of GM is exactly (4.1.6) mentioned in Section 4.1. As briefly hinted in Section 4.1, one natural way to design a practical transform sampler with inductive biases toward isomorphisms is to replace the transport cost C in (4.1.4) with (4.1.6); this can also be viewed as the Lagrangian form of GM, which replaces the constraint $T \in \mathcal{T}(\mu, \nu)$ with a penalty term $\lambda \cdot \mathcal{L}(T_{\#}\mu, \nu)$. It turns out, however, such

a formulation has some subtle issues which might be unfavorable in practice (see Remark 4.3.3). To circumvent such issues, we develop a transform sampler with slight modifications which we introduce in the next section.

4.3 Transform Sampling via Reversible Gromov-Monge

Our formulation is based on the following observation: for a coupling γ such that $\gamma = (\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$, which presents a binding constraint, we can simplify the objective function of GW as

$$\int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y))^2 \, \mathrm{d}\mu \otimes \nu \ ,$$

where $\mathrm{d}\mu \otimes \nu := \mathrm{d}\mu\,(x)\,\mathrm{d}\nu\,(y)$ denotes the product measure of μ and ν. Minimizing the above objective function under the binding constraint leads to the following definition.

Definition 4.3.1. For network spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$, we write $(F, B) \in \mathcal{I}(\mu, \nu)$ if measurable maps $F\colon \mathcal{X} \to \mathcal{Y}$ and $B\colon \mathcal{Y} \to \mathcal{X}$ satisfy the binding constraint $(\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$. We define the reversible Gromov-Monge (RGM) distance between $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ as

$$\mathrm{RGM}(\mu, \nu) := \inf_{(F,B)\in\mathcal{I}(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y))^2 \, \mathrm{d}\mu \otimes \nu \right)^{1/2} . \tag{4.3.1}$$

Remark 4.3.2. A few remarks are in place for the binding constraint. If $(\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$, then $F_{\#}\mu = \nu$ and $B_{\#}\nu = \mu$ follow due to marginal conditions. However, the converse is not true in general. To see this, let $\mu = \nu = \mathrm{Unif}([0, 1])$, then $F_{\#}\mu = \nu$ and $B_{\#}\nu = \mu$ hold for $F(x) = B(x) = |2x - 1|$. However, $(\mathrm{Id}, F)_{\#}\mu \neq (B, \mathrm{Id})_{\#}\nu$ because $(\mathrm{Id}, F)_{\#}\mu$ is a uniform measure on $\{(x, |2x-1|) : x \in [0, 1]\}$, whereas $(B, \mathrm{Id})_{\#}\nu$ is a uniform measure on $\{(|2y-1|, y) : y \in [0, 1]\}$. Lastly, note that $\mathcal{I}(\mu, \nu)$ might be empty, for instance, if μ and ν are discrete and their supports have different cardinality; say, $\mu = \delta_x$ and $\nu = (\delta_{y_1} + \delta_{y_2})/2$, namely, Dirac measures supported on $x \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$; in such a case,

$\mathrm{RGM}(\mu, \nu) = \infty$.

Roughly speaking, computing RGM consists in finding a pair $(F, B) \in \mathcal{I}(\mu, \nu)$ such that $c_{\mathcal{X}}(x, B(y))$ best matches $c_{\mathcal{Y}}(F(x), y)$ on average. Like a strong isomorphism, we can view such a pair as jointly capturing an isomorphic relation of $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$; details on such analytic properties will be discussed in Section 4.5.1.

RGM Sampler We design a transform sampling method based on finding a minimizing pair (F, B) of RGM to capture isomorphic relations between network spaces. To implement this method, we utilize the form similar to (4.1.4) mentioned in Section 4.1, which leads to the Lagrangian form that allows efficient estimation of (F, B) using i.i.d. samples from μ and ν. The idea is to consider the Lagrangian of the minimization problem in the definition of RGM. First, we rewrite the minimization problem with the binding constraint as follows,

$$\begin{aligned} \min_{\substack{F\colon \mathcal{X}\to\mathcal{Y} \\ B\colon \mathcal{Y}\to\mathcal{X}}} \quad & \int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y))^2 \, \mathrm{d}\mu \otimes \nu \\ \text{s.t.} \quad & \mathcal{L}_{\mathcal{X}\times\mathcal{Y}}((\mathrm{Id}, F)_{\#}\mu, (B, \mathrm{Id})_{\#}\nu) = 0 \, . \end{aligned} \tag{4.3.2}$$

Here, $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$ is a suitable discrepancy measure on $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ so that the constraint of (4.3.2) is a surrogate for the original constraint $(\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$. In practice, we do not require that $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}} = 0$ implies $(\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$; in fact, the former constraint can be a relaxation of the latter. The choice of $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$ will be specified later. Now, we turn (4.3.2) into the Lagrangian:

$$\min_{\substack{F\colon \mathcal{X}\to\mathcal{Y} \\ B\colon \mathcal{Y}\to\mathcal{X}}} \int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y))^2 \, \mathrm{d}\mu \otimes \nu + \lambda \cdot \mathcal{L}_{\mathcal{X}\times\mathcal{Y}}((\mathrm{Id}, F)_{\#}\mu, (B, \mathrm{Id})_{\#}\nu) \, .$$

Given i.i.d. samples $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$ from μ and ν, respectively, we replace the population objective with its empirical estimates:

$$\min_{\substack{F\colon \mathcal{X}\to\mathcal{Y}\\ B\colon \mathcal{Y}\to\mathcal{X}}} \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n (c_{\mathcal{X}}(x_i, B(y_j)) - c_{\mathcal{Y}}(F(x_i), y_j))^2 + \lambda \cdot \mathcal{L}_{\mathcal{X}\times\mathcal{Y}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n)\ ,$$

where $\widehat{\mu}_m$ and $\widehat{\nu}_n$ are the empirical measures based on $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$, respectively. Empirically, we find that adding the following extra terms often enhances empirical results:

$$\begin{aligned}\min_{\substack{F\colon \mathcal{X}\to\mathcal{Y}\\ B\colon \mathcal{Y}\to\mathcal{X}}} \quad & \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n (c_{\mathcal{X}}(x_i, B(y_j)) - c_{\mathcal{Y}}(F(x_i), y_j))^2 + \lambda_1 \cdot \mathcal{L}_{\mathcal{X}\times\mathcal{Y}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n) \\ & + \lambda_2 \cdot \mathcal{L}_{\mathcal{X}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) + \lambda_3 \cdot \mathcal{L}_{\mathcal{Y}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n)\ .\end{aligned}$$

Like $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$, we utilize suitable discrepancy measures $\mathcal{L}_{\mathcal{X}}$ and $\mathcal{L}_{\mathcal{Y}}$ on $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$, respectively, so that the additional terms help matching the marginals of $(\mathrm{Id}, F)_{\#}\widehat{\mu}_m$ and $(B, \mathrm{Id})_{\#}\widehat{\nu}_n$.

Lastly, we discuss the choice of $\mathcal{L}_{\mathcal{X}}, \mathcal{L}_{\mathcal{Y}}$, and $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$. We use the square of Maximum Mean Discrepancy (MMD) as the leading example.[3] MMD between two measures is a distance between their embeddings in some Reproducing Kernel Hilbert Space (RKHS), which is indeed a metric under mild conditions [Muandet et al., 2017]. Also, MMD is representable via the reproducing kernel of the RKHS, hence one may simply choose a kernel function to define it. Concretely, for any kernel $K_{\mathcal{X}}$ on $\mathcal{X}$, the square of MMD between $\widehat{\mu}_m$ and $B_{\#}\widehat{\nu}_n$ is

$$\frac{1}{m^2}\sum_{i,i'} K_{\mathcal{X}}(x_i, x_{i'}) + \frac{1}{n^2}\sum_{j,j'} K_{\mathcal{X}}(B(y_j), B(y_{j'})) - \frac{2}{mn}\sum_{i,j} K_{\mathcal{X}}(x_i, B(y_j))\ .$$

To utilize such a convenient closed form, we specify $\mathcal{L}_{\mathcal{X}}, \mathcal{L}_{\mathcal{Y}}, \mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$ as the squares of corresponding MMDs by choosing kernels $K_{\mathcal{X}}, K_{\mathcal{Y}}, K_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X}, \mathcal{Y}, \mathcal{X}\times\mathcal{Y}$, respectively. For the

3. This is merely a proof of concept. One may use other quantities in practice as described in Section 4.6.

kernel $K_{\mathcal{X}\times\mathcal{Y}}$ on the product space, we use the tensor product kernel $K_{\mathcal{X}} \otimes K_{\mathcal{Y}}$ given as

$$K_{\mathcal{X}} \otimes K_{\mathcal{Y}}((x,y),(x',y')) = K_{\mathcal{X}}(x,x')K_{\mathcal{Y}}(y,y') \,.$$

The tensor product notation is employed since the kernel on the product space inherits the feature map as the tensor product of two individual feature maps w.r.t. $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$. Denoting the MMD associated with a kernel K as MMD_K, we obtain the following minimization problem:

$$\begin{aligned} \min_{\substack{F\colon \mathcal{X}\to\mathcal{Y} \\ B\colon \mathcal{Y}\to\mathcal{X}}} \quad & \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(c_{\mathcal{X}}(x_i, B(y_j)) - c_{\mathcal{Y}}(F(x_i), y_j))^2 \\ & + \lambda_1 \cdot \mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n) \\ & + \lambda_2 \cdot \mathrm{MMD}^2_{K_{\mathcal{X}}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) + \lambda_3 \cdot \mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n) \,. \end{aligned} \tag{4.3.3}$$

Once we solve the problem above, the solution $\widehat{F}\colon \mathcal{X} \to \mathcal{Y}$ will serve as an approximate isomorphism and facilitate transform sampling of the target ν from a known distribution μ. The map $\widehat{B}$ possesses similar properties as $\widehat{F}$, whereas the map $\widehat{F}$ is of our primary interest for sampling purposes. The reverse map $\widehat{B}\colon \mathcal{Y} \to \mathcal{X}$ also embeds point clouds in $\mathcal{Y}$ into $\mathcal{X}$, with approximate isomorphism properties in the sense of Gromov-Monge.

Like other transform sampling approaches in generative modeling, a practical way to solve (4.3.3) is to restrict the maps F and B to vector-valued function classes $\mathcal{F}$ and $\mathcal{B}$ parametrized by neural networks, respectively, and then optimize using a gradient descent algorithm. We note the following difference between this minimization problem and adversarial formulations as in GANs: variational problems of GANs consist of minimization over a class of generators and maximization over a class of discriminators, which requires complex saddle-point dynamics [Daskalakis et al., 2017, Liang and Stokes, 2019]. In contrast, the proposed RGM sampler only solves a single minimization problem in network parameters.

Although generally non-convex in nature, the parameter minimization problem in neural networks can often be efficiently optimized by stochastic gradient descent, and can even provably achieve the global optima if the loss satisfies certain Polyak-Łojasiewicz conditions [Bassily et al., 2018].

Remark 4.3.3. We conclude this section by explaining a subtle difference between the proposed RGM sampler and using the form (4.1.4) with C replaced by Q defined in (4.1.6). This subtlety also explains why we employ the RGM formulation rather than GM. The latter approach—which can be viewed as the Lagrangian of GM as mentioned at the end of Section 4.2—aims to find an isomorphism $T\colon \mathcal{X} \to \mathcal{Y}$ such that $c_{\mathcal{X}}(x, x')$ best matches $c_{\mathcal{Y}}(T(x), T(x'))$ on average, whose plug-in version with empirical data is

$$\frac{1}{m^2} \sum_{i,i'=1}^{m} (c_{\mathcal{X}}(x_i, x_{i'}) - c_{\mathcal{Y}}(T(x_i), T(x_{i'})))^2. \tag{4.3.4}$$

It is desirable to use information from both samples $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$ to capture any isomorphic relation between two spaces. However, the GM objective (4.3.4) only uses information—the samples $\{x_i\}_{i=1}^m$—from $\mathcal{X}$, not from the target space $\mathcal{Y}$. In contrast, our RGM objective uses both information—samples from both spaces—to capture isomorphic relation, which is encoded by the first term in (4.3.3):

$$\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (c_{\mathcal{X}}(x_i, B(y_j)) - c_{\mathcal{Y}}(F(x_i), y_j))^2.$$

As the most valuable information (given that our goal is to learn the target distribution ν) is the available samples $\{y_j\}_{j=1}^n$ from ν, it is favorable to utilize them to learn isomorphisms.

4.4 Statistical Analysis of the RGM Sampler

This section provides the main theoretical analysis of this chapter: we study the statistical rate of convergence for the empirical problem (4.3.3), assuming it can be solved accurately. First, define

$$\begin{aligned} C(\mu,\nu,F,B) := &\int (c_{\mathcal{X}}(x,B(y)) - c_{\mathcal{Y}}(F(x),y))^2 \,\mathrm{d}\mu\otimes\nu \\ &+ \lambda_1 \cdot \mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}((\mathrm{Id},F)_{\#}\mu,(B,\mathrm{Id})_{\#}\nu) \\ &+ \lambda_2 \cdot \mathrm{MMD}^2_{K_{\mathcal{X}}}(\mu,B_{\#}\nu) + \lambda_3 \cdot \mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\mu,\nu)\ . \end{aligned} \tag{4.4.1}$$

Then, the objective function of (4.3.3) is a plug-in estimator $C(\widehat{\mu}_m,\widehat{\nu}_n,F,B)$. We consider solving (4.3.3) over the transformation class $\mathcal{F}\times\mathcal{B}$ given as follows, for which we will state our non-asymptotic results in full generality. From now on, let $\mathcal{X}$ and $\mathcal{Y}$ be subsets of Euclidean spaces of dimensions $\dim(\mathcal{X})$ and $\dim(\mathcal{Y})$, respectively. $\mathcal{F}$ (resp. $\mathcal{B}$) is a collection of vector-valued measurable functions from $\mathcal{X}$ to $\mathcal{Y}$ (resp. from $\mathcal{Y}$ to $\mathcal{X}$). For each $F\in\mathcal{F}$ and $k\in[\dim(\mathcal{Y})]$, we write $F_k(x)$ to denote the k-th coordinate of $F(x)$. Accordingly, we define $\mathcal{F}_k=\{F_k:\mathcal{X}\to\mathbb{R}\mid F\in\mathcal{F}\}$, namely, a collection of real-valued measurable functions defined on $\mathcal{X}$ that are given as the k-th coordinate of $F\in\mathcal{F}$. For $\ell\in[\dim(\mathcal{X})]$, we define B_ℓ and $\mathcal{B}_\ell=\{B_\ell:\mathcal{Y}\to\mathbb{R}\mid B\in\mathcal{B}\}$ analogously. Then, solving (4.3.3) over $\mathcal{F}\times\mathcal{B}$ is written as $\min_{(F,B)\in\mathcal{F}\times\mathcal{B}} C(\widehat{\mu}_m,\widehat{\nu}_n,F,B)$. We prove that the empirical solution leads to an approximate infimum of $(F,B)\mapsto C(\mu,\nu,F,B)$ evaluated with the population measures μ,ν, with sufficiently large sample sizes m and n.

Overview of Assumptions Before stating the main theorem, we briefly outline technical assumptions; the complete statement of the assumptions and definitions will be provided shortly. First, we assume that the cost functions $c_{\mathcal{X}},c_{\mathcal{Y}}$ are bounded and Lipschitz (Assumptions 4.4.3, 4.4.9). Similarly, we assume boundedness and Lipschitzness of the kernel functions $K_{\mathcal{X}},K_{\mathcal{Y}}$ corresponding to the MMD term (Assumptions 4.4.5, 4.4.16). Lastly,

we impose two assumptions on the classes of transformations $F\colon \mathcal{X} \to \mathcal{Y}$ and $B\colon \mathcal{Y} \to \mathcal{X}$: we assume that the transformation classes are uniformly bounded (Assumption 4.4.8) and should contain non-trivial maps (Assumption 4.4.17). We will also employ a notion of combinatorial dimension to measure the complexity of real-valued function classes, which is called the pseudo-dimension (Definition 4.4.11).

Theorem 4.4.1. *Let $(\widehat{F}, \widehat{B})$ be a solution to the empirical problem*

$$(\widehat{F}, \widehat{B}) \in \operatorname*{arg\,min}_{(F,B)\in\mathcal{F}\times\mathcal{B}} C(\widehat{\mu}_m, \widehat{\nu}_n, F, B) \ ,$$

with $C\colon \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \times \mathcal{F} \times \mathcal{B} \to \mathbb{R}$ defined in (4.4.1). Under Assumptions 4.4.3-4.4.17, the following inequality holds with probability at least $1-\delta$ on $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$

$$C(\mu, \nu, \widehat{F}, \widehat{B}) - \inf_{(F,B)\in\mathcal{F}\times\mathcal{B}} C(\mu, \nu, F, B) \precsim \mathcal{M}(\mathcal{F}, \mathcal{B}, m, n, \delta) \ . \tag{4.4.2}$$

Here, $\mathcal{M}(\mathcal{F}, \mathcal{B}, m, n, \delta)$ denotes a complexity measure of $(\mathcal{F}, \mathcal{B})$ given in terms of pseudo-dimensions (Pdim) of $\mathcal{F}_k$ and $\mathcal{B}_\ell$ defined in Definition 4.4.11:

$$\mathcal{M}(\mathcal{F}, \mathcal{B}, m, n, \delta) := \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m \wedge n}} + \sqrt{\frac{\log(m \vee n)}{m \wedge n}\left(\sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k) + \sum_{\ell=1}^{\dim(\mathcal{X})} \mathrm{Pdim}(\mathcal{B}_\ell)\right)} \ .$$

Remark 4.4.2. When $\mathcal{F}$ and $\mathcal{B}$ are parametrized by neural network classes (the ones we will use for numerical demonstrations in Section 4.6), tight pseudo-dimension bounds established in Anthony and Bartlett [1999], Harvey et al. [2017] can be plugged in Theorem 4.4.1 for concrete non-asymptotic rates.

Overview of the Proof of Theorem 4.4.1 The rest of this section presents the main ideas of the proof of Theorem 4.4.1. To derive an upper bound (4.4.2), we will decompose the left-hand side of (4.4.2) into several terms, derive upper bounds on them separately, and

then combine those upper bounds together to obtain the right-hand side of (4.4.2); details of the proofs are provided in the supplementary material.

Without loss of generality, we assume $\lambda_1 = \lambda_2 = \lambda_3 = 1$ in $C(\mu, \nu, F, B)$ since the proof is essentially identical with any constants $\lambda_1, \lambda_2, \lambda_3 > 0$. For convenience, we denote

$$C_0(F, B) = \int \left(c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y)\right)^2 \mathrm{d}\mu \otimes \nu \ ,$$
$$M(F, B) = \mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\mu, \nu) + \mathrm{MMD}^2_{K_{\mathcal{X}}}(\mu, B_{\#}\nu) + \mathrm{MMD}^2_{K_{\mathcal{X}} \otimes K_{\mathcal{Y}}}((\mathrm{Id}, F)_{\#}\mu, (B, \mathrm{Id})_{\#}\nu)$$

and therefore $C(\mu, \nu, F, B) = C_0(F, B) + M(F, B)$. Similarly, define the empirical counterparts as

$$\widehat{C}_0(F, B) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (c_{\mathcal{X}}(x_i, B(y_j)) - c_{\mathcal{Y}}(F(x_i), y_j))^2 \ ,$$
$$\widehat{M}(F, B) = \mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n) + \mathrm{MMD}^2_{K_{\mathcal{X}}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) + \mathrm{MMD}^2_{K_{\mathcal{X}} \otimes K_{\mathcal{Y}}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n)$$

and thus $C(\widehat{\mu}_m, \widehat{\nu}_n, F, B) = \widehat{C}_0(F, B) + \widehat{M}(F, B)$.

Our goal is to give an upper bound on $C(\mu, \nu, \widehat{F}, \widehat{B}) - \inf_{(F,B) \in \mathcal{F} \times \mathcal{B}} C(\mu, \nu, F, B)$. To this end, first recall that

$$C(\widehat{\mu}_m, \widehat{\nu}_n, \widehat{F}, \widehat{B}) \leq C(\widehat{\mu}_m, \widehat{\nu}_n, F, B)$$

holds for any $F \in \mathcal{F}$ and $B \in \mathcal{B}$ by definition of $\widehat{F}$ and $\widehat{B}$ given in Theorem 4.4.1. Therefore,

$$C(\mu, \nu, \widehat{F}, \widehat{B}) - C(\mu, \nu, F, B) \leq C(\mu, \nu, \widehat{F}, \widehat{B}) - C(\widehat{\mu}_m, \widehat{\nu}_n, \widehat{F}, \widehat{B}) + C(\widehat{\mu}_m, \widehat{\nu}_n, F, B) - C(\mu, \nu, F, B) \ ,$$

where the right-hand side can be decomposed as

$$C_0(\widehat{F}, \widehat{B}) - \widehat{C}_0(\widehat{F}, \widehat{B}) + M(\widehat{F}, \widehat{B}) - \widehat{M}(\widehat{F}, \widehat{B}) + \widehat{C}_0(F, B) - C_0(F, B) + \widehat{M}(F, B) - M(F, B) \ .$$

To further control the expression, we first derive probabilistic bounds on $|\widehat{C}_0(F,B) - C_0(F,B)|$ and $|\widehat{M}(F,B) - M(F,B)|$ that hold for a fixed $(F,B) \in \mathcal{F} \times \mathcal{B}$ via standard concentration inequalities. We derive uniform probabilistic bounds on $\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F,B) - C_0(F,B)|$ and $\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{M}(F,B) - M(F,B)|$, using tools from empirical process theory.

4.4.1 Concentration Inequalities and Uniform Deviations

We utilize the McDiarmid's inequality to derive bounds on $|\widehat{C}_0(F,B) - C_0(F,B)|$ and $|\widehat{M}(F,B) - M(F,B)|$ for fixed F, B. To give a bound on the former, we make the following boundedness assumption.

Assumption 4.4.3. $c_{\mathcal{X}}(\cdot,\cdot), c_{\mathcal{Y}}(\cdot,\cdot)$ is uniformly bounded, that is, there exists a constant $H > 0$ such that

$$\sup_{(x,x')\in\mathcal{X}\times\mathcal{X}} c_{\mathcal{X}}(x,x'), \sup_{(y,y')\in\mathcal{Y}\times\mathcal{Y}} c_{\mathcal{Y}}(y,y') \leq \sqrt{\frac{H}{4}} \, .$$

Proposition 4.4.4. *Under Assumption 4.4.3, for any pair $(F,B) \in \mathcal{F} \times \mathcal{B}$ and $\delta > 0$,*

$$|\widehat{C}_0(F,B) - C_0(F,B)| \precsim \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m \wedge n}}$$

holds with probability at least $1 - 4\delta$.

To derive a similar bound on $|\widehat{M}(F,B) - M(F,B)|$, we assume that kernels are bounded.

Assumption 4.4.5. There exists $K > 0$ such that

$$\sup_{x\in\mathcal{X}} |K_{\mathcal{X}}(x,x)| \, , \, \sup_{y\in\mathcal{Y}} |K_{\mathcal{Y}}(y,y)| \leq K \, .$$

Proposition 4.4.6. *Under Assumption 4.4.5, for any pair* $(F, B) \in \mathcal{F} \times \mathcal{B}$ *and* $\delta > 0$,

$$|\widehat{M}(F, B) - M(F, B)| \precsim \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}}$$

holds with probability at least $1 - 6\delta$.

We now derive uniform deviation bounds for

$$\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F, B) - C_0(F, B)| \,, \quad \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{M}(F, B) - M(F, B)| \,.$$

For the former, we use the notion of uniform covering numbers defined below.

Definition 4.4.7 (Uniform Covering Number)**.** Let $\mathcal{G}$ be a collection of real-valued functions defined on a set $\mathcal{Z}$. Given m points $z_1, \ldots, z_m \in \mathcal{Z}$ and any $\delta > 0$, we define $N_\infty(\delta, \mathcal{G}, \{z_i\}_{i=1}^m)$ to be the δ-covering number of $\mathcal{G}$ under the pseudometric d induced by points $z_1, \ldots, z_m$:

$$d(g, g') := \max_{i\in[m]} |g(z_i) - g'(z_i)| \,.$$

Also, we define the uniform δ-covering number of $\mathcal{G}$ as follows:

$$N_\infty(\delta, \mathcal{G}, m) := \sup \{N_\infty(\delta, \mathcal{G}, \{z_i\}_{i=1}^m) : z_1, \ldots, z_m \in \mathcal{Z}\} \,.$$

Here, the supremum is taken over all possible combinations of m points in $\mathcal{Z}$.

Also, we make the following assumptions.

Assumption 4.4.8. $\mathcal{F}_k$ and $\mathcal{B}_\ell$ consist of uniformly bounded functions, that is, there exists a constant $b > 0$ such that

$$\max_{k\in[\dim(\mathcal{Y})]} \sup_{F_k\in\mathcal{F}_k} \|F_k\|_\infty \,, \quad \max_{\ell\in[\dim(\mathcal{X})]} \sup_{B_\ell\in\mathcal{B}_\ell} \|B_\ell\|_\infty \le b \,.$$

Assumption 4.4.9. There exists a constant $L > 0$ such that

$$|c_{\mathcal{X}}(x, x_1) - c_{\mathcal{X}}(x, x_2)| \leq L\|x_1 - x_2\| \ , \ |c_{\mathcal{Y}}(y_1, y) - c_{\mathcal{Y}}(y_2, y)| \leq L\|y_1 - y_2\| \ .$$

The above assumption ensures smoothness of the map $(F, B) \mapsto |\widehat{C}_0(F, B) - C_0(F, B)|$ over $\mathcal{F} \times \mathcal{B}$, which allows us to utilize the uniform covering numbers.

Proposition 4.4.10. *Under Assumptions 4.4.3, 4.4.8, 4.4.9, for any $\varepsilon > 0$ and $\delta > 0$,*

$$\begin{aligned}&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F, B) - C_0(F, B)| \\ &\precsim \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m \wedge n}} + \varepsilon + \sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\varepsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\varepsilon, \mathcal{B}_\ell, n)}{m \wedge n}}\end{aligned}$$

holds with probability at least $1 - 2\delta$.

Now, the remaining task is to choose ε carefully in Proposition 4.4.10 for a concrete upper bound. To this end, we utilize the pseudo-dimension defined below.

Definition 4.4.11 (Pseudo-Dimension)**.** Let $\mathcal{G}$ be a collection of real-valued functions defined on a set $\mathcal{Z}$. Given a subset $S := \{z_1, \ldots, z_m\} \subset \mathcal{Z}$, we say S is pseudo-shattered by $\mathcal{G}$ if there are $r_1, \ldots, r_m \in \mathbb{R}$ such that for each $b \in \{0, 1\}^m$ we can find $g_b \in \mathcal{G}$ satisfying $\operatorname{sign}(g_b(z_i) - r_i) = b_i$ for all $i \in [m]$. We define the pseudo-dimension of $\mathcal{G}$, denoted as $\operatorname{Pdim}(\mathcal{G})$, as the maximum cardinality of a subset $S \subset \mathcal{Z}$ that is pseudo-shattered by $\mathcal{G}$.

Using a well-established relation of the uniform covering number and the pseudo-dimension (Lemma 4.8.2), we can simplify Proposition 4.4.10 as follows.

Corollary 4.4.12. *Under Assumptions 4.4.3, 4.4.8, 4.4.9, for any $\delta > 0$,*

$$\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F,B) - C_0(F,B)|$$
$$\precsim \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}} + \sqrt{\frac{\log(m\vee n)}{m\wedge n}\left(\sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k) + \sum_{\ell=1}^{\dim(\mathcal{X})} \mathrm{Pdim}(\mathcal{B}_\ell)\right)}$$

holds with probability at least $1-2\delta$.

To derive an upper bound on $\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{M}(F,B) - M(F,B)|$, we first introduce the Rademacher complexities defined below.

Definition 4.4.13 (Rademacher Complexity)**.** Let $(\mathcal{Z},\rho)$ be a probability space and $\mathcal{G}$ be a collection of measurable functions defined on $\mathcal{Z}$. We define the Rademacher complexity of $\mathcal{G}$ with respect to m samples from ρ as follows:

$$R_m(\mathcal{G},\rho) = \mathbb{E}_{z_i \overset{\text{iid}}{\sim}\rho} \mathbb{E}_{\varepsilon_i} \sup_{g\in\mathcal{G}} \left| \frac{1}{m}\sum_{i=1}^m \varepsilon_i g(z_i) \right| ,$$

Here, $z_1,\ldots,z_m$ are i.i.d. samples from ρ and $\varepsilon_1,\ldots,\varepsilon_m$ are i.i.d. Rademacher random variables such that $(z_1,\ldots,z_m)$ and $(\varepsilon_1,\ldots,\varepsilon_m)$ are independent.

Proposition 4.4.14. *Denote a closed unit ball of any RKHS $\mathcal{H}$ as $\mathcal{H}(1)$. Also, let $(\mathrm{Id},\mathcal{F}) := \{(\mathrm{Id},F) : F\in\mathcal{F}\}$ and $(\mathcal{B},\mathrm{Id}) := \{(B,\mathrm{Id}) : B\in\mathcal{B}\}$; hence, they are classes of maps from $\mathcal{X}$ to $\mathcal{X}\times\mathcal{Y}$ and from $\mathcal{Y}$ to $\mathcal{X}\times\mathcal{Y}$, respectively. Under Assumption 4.4.5, for any $\delta>0$,*

$$\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{M}(F,B) - M(F,B)| \precsim \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}} + R_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F},\mu) + R_n(\mathcal{H}_\mathcal{X}(1)\circ\mathcal{B},\nu)$$
$$+ R_m(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathrm{Id},\mathcal{F}),\mu) + R_n(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathcal{B},\mathrm{Id}),\nu)$$

holds with probability at least $1-6\delta$. Here, $\mathcal{F}\circ\mathcal{G} = \{f\circ g : f\in\mathcal{F}, g\in\mathcal{G}\}$ for any function classes $\mathcal{F}$ and $\mathcal{G}$ with matching input and output space.

4.4.2 Bounding Rademacher Complexities via Chaining

Now, the only remaining task is to bound the four Rademacher complexities that appear in the upper bound of Proposition 4.4.14. We will derive upper bounds for the compositional function classes using the chaining technique. To illustrate the main idea, let us consider $\mathcal{H}_{\mathcal{Y}}(1) \circ \mathcal{F}$. Recall that

$$R_m(\mathcal{H}_{\mathcal{Y}}(1) \circ \mathcal{F}, \mu) = \mathbb{E}_{x_i \overset{\text{iid}}{\sim} \mu} R_m(\mathcal{H}_{\mathcal{Y}}(1) \circ \mathcal{F}, \{x_i\}_{i=1}^m) \ ,$$

where $R_m(\mathcal{H}_{\mathcal{Y}}(1) \circ \mathcal{F}, \{x_i\}_{i=1}^m)$ is the empirical Rademacher complexity of $\mathcal{H}_{\mathcal{Y}}(1) \circ F$ associated with $\{x_i\}_{i=1}^m$:

$$R_m(\mathcal{H}_{\mathcal{Y}}(1) \circ \mathcal{F}, \{x_i\}_{i=1}^m) = \mathbb{E}_{\varepsilon_i} \sup_{h \in \mathcal{H}_{\mathcal{Y}}(1), F \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i h(F(x_i)) \right| = \mathbb{E}_{\varepsilon_i} \sup_{h \in \mathcal{H}_{\mathcal{Y}}(1), F \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \varepsilon_i h(F(x_i)) \ .$$

Notice that we may remove the absolute value since $\mathcal{H}_{\mathcal{Y}}(1) = -\mathcal{H}_{\mathcal{Y}}(1)$. Now, considering $\{x_i\}_{i=1}^m$ as fixed, we will first bound the empirical Rademacher complexity by replacing the Rademacher random variables with Gaussian random variables. Let g_i be i.i.d. standard Gaussian random variables, then it is well known that

$$R_m(\mathcal{H}_{\mathcal{Y}}(1) \circ \mathcal{F}, \{x_i\}_{i=1}^m) \leq \sqrt{\frac{\pi}{2}} \mathbb{E}_{g_i} \sup_{h \in \mathcal{H}_{\mathcal{Y}}(1), F \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m g_i h(F(x_i)) =: \sqrt{\frac{\pi}{2}} \mathcal{G}_m(\mathcal{H}_{\mathcal{Y}}(1) \circ \mathcal{F}, \{x_i\}_{i=1}^m) \ .$$

Also, under the assumption that $K_{\mathcal{Y}}$ is bounded by K, the reproducing property and the Cauchy-Schwarz inequality imply

$$\begin{aligned}\sup_{h\in\mathcal{H}_{\mathcal{Y}}(1),F\in\mathcal{F}}\sum_{i=1}^{m}g_i h(F(x_i)) &= \sup_{h\in\mathcal{H}_{\mathcal{Y}}(1),F\in\mathcal{F}}\left\langle h,\sum_{i=1}^{m}g_i K_{\mathcal{Y}}(\cdot,F(x_i))\right\rangle_{\mathcal{H}_{\mathcal{Y}}}\\ &\leq \sup_{F\in\mathcal{F}}\left[\sum_{i=1}^{m}g_i^2 K+\sum_{i\neq j}g_i g_j K_{\mathcal{Y}}(F(x_i),F(x_j))\right]^{1/2}\\ &\leq \left[\sum_{i=1}^{m}g_i^2 K+\sup_{F\in\mathcal{F}}\sum_{i\neq j}g_i g_j K_{\mathcal{Y}}(F(x_i),F(x_j))\right]^{1/2}.\end{aligned}$$

Here, $\langle\cdot,\cdot\rangle_{\mathcal{H}_{\mathcal{Y}}}$ denotes the inner product on $\mathcal{H}_{\mathcal{Y}}$. Hence,

$$\begin{aligned}\mathcal{G}_m(\mathcal{H}_{\mathcal{Y}}(1)\circ\mathcal{F},\{x_i\}_{i=1}^m) &\leq \frac{1}{m}\mathbb{E}_{g_i}\left[\sum_{i=1}^{m}g_i^2 K+\sup_{F\in\mathcal{F}}\sum_{i\neq j}g_i g_j K_{\mathcal{Y}}(F(x_i),F(x_j))\right]^{1/2}\\ &\leq \frac{1}{m}\left[mK+\mathbb{E}_{g_i}\sup_{F\in\mathcal{F}}\sum_{i\neq j}g_i g_j K_{\mathcal{Y}}(F(x_i),F(x_j))\right]^{1/2},\end{aligned}$$

where the second inequality follows from the Jensen's inequality and $\mathbb{E}g_i^2=1$.

For any $F\colon\mathcal{X}\to\mathcal{Y}$, let $A_F\in\mathbb{R}^{m\times m}$ be a matrix whose diagonal elements are zero and (i,j)-th element is $K_{\mathcal{Y}}(F(x_i),F(x_j))$ for $i\neq j$. Then, the last term amounts to the supremum of a quadratic process

$$\mathbb{E}_g\sup_{F\in\mathcal{F}}g^\top A_F g\,,$$

where $g:=[g_1,\ldots,g_m]^\top\sim N(0,I_m)$. We rely on the following chaining bound for the quadratic processes.

Lemma 4.4.15 (Chaining Bound)**.** *Let $\mathbb{S}_0^{m\times m}$ be the collection of all symmetric matrices A whose diagonal elements are zero. Endow $\mathbb{S}_0^{m\times m}$ with a metric d given by $d(A,A'):=$*

$\|A - A'\|$. *Given* $\mathcal{T} \subset \mathbb{S}_0^{m\times m}$ *and a fixed* $A_0 \in \mathcal{T}$, *define* $\Delta = \sup_{A\in\mathcal{T}} d(A, A_0)$. *Let* $N(\delta, \mathcal{T})$ *be the covering number of* $\mathcal{T}$ *under the metric* $d(\cdot,\cdot)$, *then*

$$\mathbb{E}_g \sup_{A\in\mathcal{T}} g^\top A g \le \inf_{J\in\mathbb{N}} \left\{ m\delta_J + 12\int_{\delta_J/2}^{\Delta/2} \sqrt{2\log N(\delta,\mathcal{T})}\,\mathrm{d}\delta + 24\int_{\delta_J/2}^{\Delta/2} \log N(\delta,\mathcal{T})\,\mathrm{d}\delta \right\}, \tag{4.4.3}$$

where for any integer $J \ge 0$, *we define* $\delta_J = 2^{-J}\Delta$.

With the above chaining bound, we can directly upper bound the Rademacher complexities of the compositional classes such as $R_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F}, \mu)$ and $R_m(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathrm{Id},\mathcal{F}), \mu)$. More specifically, for the former class, we will apply this chaining bound to $\mathcal{T} := \{A_F : F \in \mathcal{F}\}$. Then, to further bound the right-hand side of (4.4.3), we make the following assumptions.

Assumption 4.4.16. Suppose $K_\mathcal{X}$ and $K_\mathcal{Y}$ are Lipschitz: there exists $L > 0$ such that

$$|K_\mathcal{X}(x_1, x') - K_\mathcal{X}(x_2, x')| \le L\|x_1 - x_2\|\,, \quad |K_\mathcal{Y}(y_1, y') - K_\mathcal{Y}(y_2, y')| \le L\|y_1 - y_2\|\,.$$

This plays a similar role as Assumption 4.4.9: we can derive an upper bound on $d(A_F, A_{F'})$ via closeness of F and F' in $\mathcal{F}$. As a result, we will see that the covering number $N(\delta, \mathcal{T})$ can be bounded by the complexity of $\mathcal{F}$.

Assumption 4.4.17. There exist y_0 and y_0' in $\mathcal{Y}$ with $K_\mathcal{Y}(y_0, y_0') \ne K_\mathcal{Y}(y_0, y_0)$ such that

- $\mathcal{F}$ contains a constant map F satisfying $F(x) = y_0$ for all $x \in \mathcal{X}$,
- whenever we have $x \ne x' \in \mathcal{X}$, we can find a non-constant map $F \in \mathcal{F}$ such that $F(x) = y_0$ and $F(x') = y_0'$.

Similarly, there exist x_0 and x_0' in $\mathcal{X}$ with $K_\mathcal{X}(x_0, x_0') \ne K_\mathcal{X}(x_0, x_0)$ such that

- $\mathcal{B}$ contains a constant map B such that $B(y) = x_0$ for all $y \in \mathcal{Y}$,

- whenever we have $y \neq y' \in \mathcal{Y}$, we can find a non-constant map $B \in \mathcal{B}$ such that $B(y) = x_0$ and $B(y') = x_0'$.

The main purpose of this assumption is to exclude overly restrictive $\mathcal{F}$ and $\mathcal{B}$, and is minimal: $\mathcal{F}$ and $\mathcal{B}$ should contain constant maps, as well as non-constant maps. With these assumptions, we can derive the following result.

Proposition 4.4.18. *Under Assumptions 4.4.5, 4.4.8, 4.4.16, 4.4.17,*

$$R_m(\mathcal{H}_{\mathcal{Y}}(1) \circ \mathcal{F}, \mu) \ , \ R_m(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1) \circ (\mathrm{Id}, \mathcal{F}), \mu) \precsim \sqrt{\frac{\log m}{m} \sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k)} \ ,$$

$$R_n(\mathcal{H}_{\mathcal{X}}(1) \circ \mathcal{B}, \mu) \ , \ R_n(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1) \circ (\mathcal{B}, \mathrm{Id}), \nu) \precsim \sqrt{\frac{\log n}{n} \sum_{k=1}^{\dim(\mathcal{X})} \mathrm{Pdim}(\mathcal{B}_k)} \ .$$

In summary, Propositions 4.4.4, 4.4.6, 4.4.14, 4.4.18 and Corollary 4.4.12 directly imply Theorem 4.4.1. Omitted proofs and details can be found in the supplementary material.

4.5 Further Studies: Metric Properties and Representer Theorem

This section discusses some further perspectives on the RGM sampler. First, we focus on the metric side of the new notion RGM; we formally state its metric properties and connections to the Gromov-Wasserstein and Gromov-Monge distance. Next, by utilizing an operator viewpoint, we develop an infinite-dimensional convex relaxation of (4.3.3), where global optima can be found efficiently; we analyze the new formulation by establishing a representer theorem on a suitable reproducing kernel Hilbert space.

4.5.1 Metric Properties of RGM

It turns out that RGM possesses metric properties similar to those of the Gromov-Wasserstein distance; the following result is an analogue to Theorem 4.2.4.

Theorem 4.5.1. *Let $\mathcal{M}$ be the collection of all metric measure spaces. Then, RGM satisfies the three metric axioms on $\mathcal{M}/\cong$, the collection of all equivalence classes of $\mathcal{M}$ induced by $\cong$.*

Remark 4.5.2. In practice, $\mathcal{X}, \mathcal{Y}$ are usually Euclidean spaces and $d_{\mathcal{X}}, d_{\mathcal{Y}}$ are the standard Euclidean distances. In many applications, instead of comparing the metric measures spaces $(\mathcal{X}, \mu, d_{\mathcal{X}}), (\mathcal{Y}, \nu, d_{\mathcal{Y}})$, it is often more desirable to work with network spaces $(\mathcal{X}, \mu, c_{\mathcal{X}}), (\mathcal{Y}, \nu, c_{\mathcal{Y}})$ with cost functions given by $c_{\mathcal{X}} = h(d_{\mathcal{X}})$ and $c_{\mathcal{Y}} = h(d_{\mathcal{Y}})$ for some transformation $h\colon \mathbb{R}_+ \to \mathbb{R}$; one of the most common choices is $h(x) = \exp\bigl(-\alpha x^2\bigr)$ with $\alpha > 0$, which makes $h(d_{\mathcal{X}}), h(d_{\mathcal{Y}})$ Radial Basis Function (RBF) kernels on $\mathcal{X}, \mathcal{Y}$, respectively (we will use these in numerical experiments). As a result, it is desirable to consider a distance on the collection of network spaces equipped with cost functions given as a composition of a transformation h and the base metric. The next result provides a modification of Theorem 4.5.1 tailored to such a situation.

Theorem 4.5.3. *Let $h\colon \mathbb{R}_+ \to \mathbb{R}$ be a continuous and strictly monotone function and $\mathcal{N}^h$ be a collection of all network spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ such that $c_{\mathcal{X}} = h(d_{\mathcal{X}})$. Then RGM satisfies the three metric axioms on $\mathcal{N}^h/\cong$, the collection of all equivalence classes of $\mathcal{N}^h$ induced by $\cong$.*

Now that we have three distances—GW, GM,[4] and RGM—between network spaces, one may wonder about the generic relations among three distances GW, GM, and RGM, which will be established in the next proposition.

Proposition 4.5.4. *For network spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ as in Definition 4.2.1,*

$$\mathrm{GW}(\mu, \nu) \leq \mathrm{GM}(\mu, \nu) \leq \mathrm{RGM}(\mu, \nu)\ . \tag{4.5.1}$$

4. Technically speaking, GM is not a distance as it is not symmetric.

Interestingly, under mild conditions, the above inequalities in Proposition 4.5.4 hold as equality, thus showing that RGM provides the exact metric as GW.

Theorem 4.5.5. *Let* $(\mathcal{X}, \mu, c_{\mathcal{X}})$ *and* $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ *be two network spaces. Assume that* $c_{\mathcal{X}}$ *and* $c_{\mathcal{Y}}$ *are bounded and* $\mu(\{x\}) = \nu(\{y\}) = 0$ *for any* $(x, y) \in \mathcal{X} \times \mathcal{Y}$. *Then,* $\mathrm{GW}(\mu, \nu) = \mathrm{GM}(\mu, \nu) = \mathrm{RGM}(\mu, \nu)$.

The proof details of Theorem 4.5.3, Proposition 4.5.4, and Theorem 4.5.5 can be found in the supplementary material.

Remark 4.5.6. The proof of Theorem 4.5.5 is inspired by Mémoli and Needham [2024], Brenier and Gangbo [2003], recent developments in the GW analysis literature that study conditions under which GW = GM holds. Another important topic in the GW analysis literature is to study the type of cost functions under which the optimal coupling of GW is induced by a transport map [Dumont et al., 2024]. Finally, we clarify that this section aims to derive metric properties of RGM in connection with GW and isomorphisms, as opposed to establishing new analytical results for GW and GM.

Finally, we conclude this section by pointing out a connection between inductive biases in RGM motivated by Brenier [1991]. Given two Polish probability spaces $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$, there exist cost functions $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ (that depend on μ, ν) such that the resulting network spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ are strongly isomorphic. Then, more importantly, among (possibly) infinitely many pairs (F, B)'s in $\mathcal{I}(\mu, \nu)$, which are all valid for transform sampling, any optimal pair $(F^\star, B^\star)$ minimizing the RGM term (4.3.1) achieves the strong isomorphism

$$\mathrm{RGM}(\mu, \nu) = \int_{\mathcal{X} \times \mathcal{Y}} (c_{\mathcal{X}}(x, B^\star(y)) - c_{\mathcal{Y}}(F^\star(x), y))^2 \, \mathrm{d}\mu \otimes \nu = 0 \,, \tag{4.5.2}$$

and thus $c_{\mathcal{X}}(x, B^\star(y)) = c_{\mathcal{Y}}(F^\star(x), y)$ almost surely. In plain language, the RGM introduces an inductive bias favoring strong isomorphisms, in the same spirit as the Wasserstein-2 metric

favors the transport map with the optimal cost seen in the introduction. Please refer to the supplementary material for a detailed discussion in 4.12.

4.5.2 Convex Relaxation and Representer Theorem

As the last bit of the study, we utilize an operator viewpoint and develop a convex relaxation of (4.3.3) by relaxing and lifting it to an infinite-dimensional space. There are two reasons behind our convex relaxation: first, as a computational alternative to the possibly non-convex optimization; second, to point out a connection with the Nadaraya-Watson estimator in classic nonparametric statistics. The crux lies in relaxing optimizing over the map $F\colon \mathcal{X} \to \mathcal{Y}$ to optimizing over its induced (dual) linear operator $\mathbf{F}\colon L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$ that maps functions on $\mathcal{Y}$ to functions on $\mathcal{X}$, where $L^2_{\mathcal{X}}$ is the collection of real-valued measurable functions f defined on $\mathcal{X}$ such that $\int_{\mathcal{X}} f^2 \,\mathrm{d}\pi_{\mathcal{X}} < \infty$ given a Borel measure $\pi_{\mathcal{X}}$ on $\mathcal{X}$; similarly, define $L^2_{\mathcal{Y}}$ given a Borel measure $\pi_{\mathcal{Y}}$ on $\mathcal{Y}$. Then, for a measurable map $F\colon \mathcal{X} \to \mathcal{Y}$, we can define $\mathbf{F}\colon L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$ by letting $\mathbf{F}(g) = g \circ F$ for all $g \in L^2_{\mathcal{Y}}$. Similarly, we define $\mathbf{B}\colon L^2_{\mathcal{X}} \to L^2_{\mathcal{Y}}$ for each measurable map $B\colon \mathcal{Y} \to \mathcal{X}$. Then, one can verify that $\mathbf{F}$ and $\mathbf{B}$ are well-defined bounded linear operators under a mild assumption (detailed proofs are provided in the supplementary material).

We derive the representer theorem for the convex relaxation of (4.3.3) under $c_{\mathcal{X}} = K_{\mathcal{X}}$ and $c_{\mathcal{Y}} = K_{\mathcal{Y}}$, namely, the cost functions are the kernel functions specified in MMD terms. We show that this problem can be reduced to a finite-dimensional convex optimization by proving a representer theorem. Moreover, since finite-dimensional convex optimization can be optimized globally with provable guarantees, such a formulation can be solved numerically efficiently.

Let us lay out more details to state the result. Due to Mercer's theorem, let $\{\phi_k \in L^2_{\mathcal{X}}\}_{k\in\mathbb{N}}$ and $\{\psi_\ell \in L^2_{\mathcal{Y}}\}_{\ell\in\mathbb{N}}$ be countable orthonormal bases of $L^2_{\mathcal{X}}$ and $L^2_{\mathcal{Y}}$ where the

kernels admit the following spectral decompositions:

$$K_{\mathcal{X}}(x,x') = \sum_k \lambda_k \phi_k(x)\phi_k(x') \ , \quad K_{\mathcal{Y}}(y,y') = \sum_\ell \gamma_\ell \psi_\ell(y)\psi_\ell(y') \ , \tag{4.5.3}$$

with positive eigenvalues $\lambda_k, \gamma_\ell > 0$. Since $\mathbf{F}\colon L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$ defines a bounded linear operator, one can represent $\mathbf{F}$ (correspondingly $\mathbf{B}$) under the orthonormal bases

$$\mathbf{F}[\psi_\ell] = \sum_{k=1}^{\infty} \mathbf{F}_{k\ell}\phi_k \ , \quad \mathbf{B}[\phi_k] = \sum_{\ell=1}^{\infty} \mathbf{B}_{\ell k}\psi_\ell \ . \tag{4.5.4}$$

Here, $[\mathbf{F}_{k\ell}]$ is a semi-infinite matrix with each column describing the $L^2_{\mathcal{X}}$ representation of $\mathbf{F}[\psi_\ell]$ under the basis $\{\phi_k \in L^2_{\mathcal{X}}\}_{k\in\mathbb{N}}$. With a slight abuse of notation, we will write $\mathbf{F}$ and $\mathbf{B}$ to denote these matrices $[\mathbf{F}_{k\ell}]$ and $[\mathbf{B}_{\ell k}]$.[5] Then, we can prove that the objective function in (4.3.3) with $c_{\mathcal{X}} = K_{\mathcal{X}}$ and $c_{\mathcal{Y}} = K_{\mathcal{Y}}$ is

$$\begin{aligned}
\Omega(\mathbf{F},\mathbf{B}) := \frac{1}{mn}\sum_{i,j}(\Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} - \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j})^2 \\
+ \lambda_1 \cdot \left(\frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \Lambda \Phi_{x_i'} \Phi_{x_i}^\top \mathbf{F}\Gamma\mathbf{F}^\top \Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \Gamma \Psi_{y_{j'}} \Psi_{y_j}^\top \mathbf{B}\Lambda\mathbf{B}^\top \Psi_{y_{j'}} \right. \\
\left. - \frac{2}{mn}\sum_{i,j} \Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i}\Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j} \right) \\
+ \lambda_2 \cdot \left(\frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \Lambda \Phi_{x_i'} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \mathbf{B}\Lambda\mathbf{B}^\top \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} \right) \\
+ \lambda_3 \cdot \left(\frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \mathbf{F}\Gamma\mathbf{F}^\top \Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \Gamma \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j} \right) \ .
\end{aligned}$$

Here, $\mathbf{F}$ and $\mathbf{B}$ are the matrices denoting the operators induced by F and B, respectively, $\Phi_x = [\cdots, \phi_k(x), \cdots]^\top \in \mathbb{R}^\infty$ and $\Psi_y = [\cdots, \psi_\ell(y), \cdots]^\top \in \mathbb{R}^\infty$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

5. In other words, $\mathbf{F}$ and $\mathbf{B}$ are infinite-dimensional matrices, in which the number of rows and the number of columns can be infinite.

and $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \dots)$ and $\Gamma = \operatorname{diag}(\gamma_1, \gamma_2, \dots)$ are diagonal matrices; detailed proofs are provided in the supplementary material. Hence, (4.3.3) can be lifted to an infinite-dimensional optimization problem

$$\min_{(\mathbf{F},\mathbf{B})\in\mathcal{C}} \ \Omega(\mathbf{F},\mathbf{B}) \ , \tag{4.5.5}$$

where $\mathcal{C}$ denotes the constraint set implying that $\mathbf{F}$ and $\mathbf{B}$ are matrices corresponding to bounded linear operators induced by some maps $F\colon \mathcal{X} \to \mathcal{Y}$ and $B\colon \mathcal{Y} \to \mathcal{X}$.

We will relax this problem by removing the constraint set $\mathcal{C}$, namely, by considering all matrices in $\mathbb{R}^{\infty\times\infty}$ as the decision variables,

$$\min_{\mathbf{F},\mathbf{B}\in\mathbb{R}^{\infty\times\infty}} \ \Omega(\mathbf{F},\mathbf{B}) \ . \tag{4.5.6}$$

In other words, this relaxed problem minimizes Ω over any pair of infinite-dimensional matrices. The next result, which we refer to as the representer theorem, shows that (4.5.6) boils down to a finite-dimensional convex program; the proof and relevant details can be found in the supplementary material.

Theorem 4.5.7. *Consider* (4.5.5) *under the assumptions in Proposition 4.10.2. Then, for any minimizer* $(\mathbf{F}^\star, \mathbf{B}^\star)$ *to the relaxed problem* (4.5.6)*, we can find finite-dimensional matrices* $\mathsf{F}^\star_{m,n} \in \mathbb{R}^{m\times n}$ *and* $\mathsf{B}^\star_{n,m} \in \mathbb{R}^{n\times m}$ *such that*

$$\mathbf{F}^\star = \Lambda\Phi_m \mathsf{F}^\star_{m,n} \Psi_n^\top \ , \quad \mathbf{B}^\star = \Gamma\Psi_n \mathsf{B}^\star_{n,m} \Phi_m^\top \ ,$$

where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \dots)$*,* $\Gamma = \operatorname{diag}(\gamma_1, \gamma_2, \dots)$*, and* $\Phi_m \in \mathbb{R}^{\infty\times m}$ *and* $\Psi_n \in \mathbb{R}^{\infty\times n}$ *are matrices whose elements are* $\phi_k(x_i)$ *and* $\psi_\ell(y_j)$*, as defined in* (4.5.3)*. In this case,* $\Omega(\mathbf{F}^\star, \mathbf{B}^\star)$ *can be rewritten as* $\omega(\mathsf{F}^\star_{m,n}, \mathsf{B}^\star_{n,m})$ *for some convex function* ω *defined over* $\mathbb{R}^{m\times n} \times \mathbb{R}^{n\times m}$*.*

Hence, by minimizing ω over $\mathbb{R}^{m\times n}\times\mathbb{R}^{n\times m}$, we obtain a relaxation of (4.5.6)*, that is,*

$$\min_{\mathbf{F},\mathbf{B}\in\mathbb{R}^{\infty\times\infty}}\Omega(\mathbf{F},\mathbf{B})\geq\min_{\substack{\mathsf{F}_{m,n}\in\mathbb{R}^{m\times n}\\ \mathsf{B}_{n,m}\in\mathbb{R}^{n\times m}}}\omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m})\ .$$

In particular, the right-hand side is a finite-dimensional convex optimization. Lastly, this relaxation is tight, that is,

$$\min_{\mathbf{F},\mathbf{B}\in\mathbb{R}^{\infty\times\infty}}\Omega(\mathbf{F},\mathbf{B})=\min_{\substack{\mathsf{F}_{m,n}\in\mathbb{R}^{m\times n}\\ \mathsf{B}_{n,m}\in\mathbb{R}^{n\times m}}}\omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m})\ ,$$

if kernel matrices $\mathbf{K}_{\mathcal{X}}$ and $\mathbf{K}_{\mathcal{Y}}$ whose elements are $K_{\mathcal{X}}(x_i,x_{i'})$ and $K_{\mathcal{Y}}(y_j,y_{j'})$, are positive definite.

Remark 4.5.8. Looking inside the proof of Theorem 4.5.7, we know the solution to the infinite-dimensional optimization is an operator taking form of $\mathbf{F}^\star=\Lambda\Phi_m\mathsf{F}^\star_{m,n}\Psi_n^\top$, with a finite-dimensional matrix $\mathsf{F}^\star_{m,n}\in\mathbb{R}^{m\times n}$. Therefore, for any $g\in L^2_{\mathcal{Y}}$, we can deduce

$$\mathbf{F}^\star[g](x)=\underbrace{K_{\mathcal{X}}(x,X_m)}_{1\times m}\underbrace{\mathsf{F}^\star_{m,n}}_{m\times n}\underbrace{g(Y_n)}_{n\times 1}\ , \tag{4.5.7}$$

where $K_{\mathcal{X}}(x,X_m)$ maps each $x\in\mathcal{X}$ to a row vector whose i-th element is $K_{\mathcal{X}}(x,x_i)$ and $g(Y_n)$ denotes a column vector whose j-th element is $g(y_j)$.

Now let's draw a connection between the classic Nadaraya-Watson estimator and (4.5.7). For now consider a special case: (x_i,y_i)'s are paired with $m=n$. In such a case, the Nadaraya-Watson estimator takes the form

$$\sum_{i,j}K_{\mathcal{X}}(x,x_i)\cdot\tfrac{1}{m}\delta_{i=j}\cdot g(y_j)\ ; \tag{4.5.8}$$

Namely, for a new point x, the corresponding function value $g(y)$ evaluated on its coupled

$y = F(x)$ is a weighted average of $g(y_j)$'s according to the affinity $K_{\mathcal{X}}(x, x_i)$. Our solution (4.5.7) extends the above nonparametric smoothing idea to the decoupled data case, where the coupling weights $\mathsf{F}^{\star}_{m,n}$ is based on a solution to a convex program, with

$$(4.5.7) = \sum_{i,j} K_{\mathcal{X}}(x, x_i) \cdot \mathsf{F}^{\star}_{m,n}[i, j] \cdot g(y_j) \ . \tag{4.5.9}$$

Lastly, we draw another connection to the Monte-Carlo integration. One downstream task after learning the distribution ν is to perform numerical integration of $g \in L^2_{\mathcal{Y}}$ under the measure $\nu \in \mathcal{P}(\mathcal{Y})$. In our transform sampling framework, this amounts to evaluating $\mathbb{E}_{y \sim F^{\star}_{\#}\mu}[g(y)] = \mathbb{E}_{x \sim \mu}[g \circ F^{\star}(x)]$. The integration, casted in the induced operator form, has the expression

$$\mathbb{E}_{x \sim \mu}\big[\mathbf{F}^{\star}[g](x)\big] = \mathbb{E}_{x \sim \mu}\big[\underbrace{K_{\mathcal{X}}(x, X_m)\mathsf{F}^{\star}_{m,n}}_{=:W(x) \in \mathbb{R}^n} g(Y_n)\big] = \mathbb{E}_{x \sim \mu}\big[\sum_{j=1}^{n} W_j(x) g(y_j)\big] \tag{4.5.10}$$

where $W(x)$ can be interpreted as the importance weights in the Monte-Carlo integration. We conclude with one more remark as a sanity check: if plug in instead $x \sim \widehat{\mu}_m$ in (4.5.10), one can verify that under mild conditions,

$$\mathbb{E}_{x \sim \widehat{\mu}_m}\big[\mathbf{F}^{\star}[g](x)\big] = \frac{1}{n} \sum_{j=1}^{n} g(y_j) \ . \tag{4.5.11}$$

That is, with the empirical measure as input, (4.5.10) outputs the simple sample average.

4.6 Experiments

This section studies the empirical performance of the reversible Gromov-Monge sampler as a proof of concept. Following Section 4.3, we find a minimum $(\widehat{F}, \widehat{B})$ of (4.3.3) over a suitable class $\mathcal{F} \times \mathcal{B}$ via gradient descent; then, we inspect the quality of transform sampling.

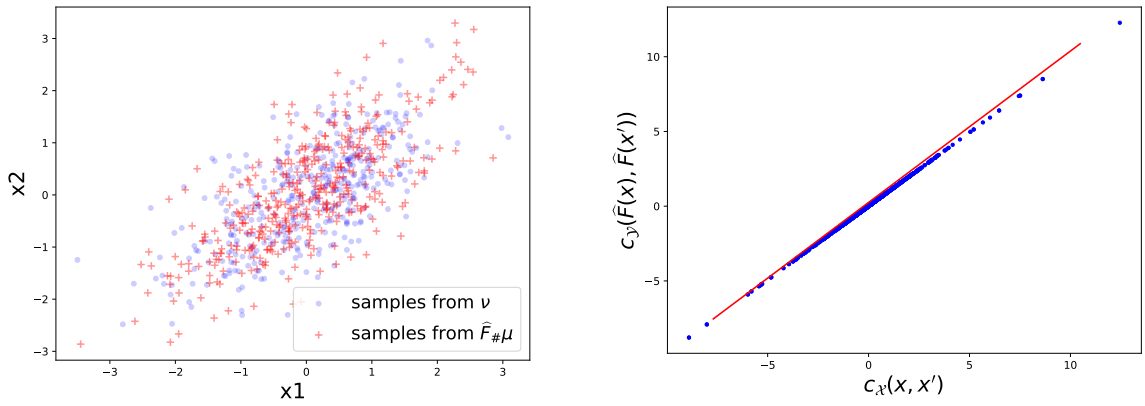
Complete implementation details of the experiments are deferred to 4.11.

4.6.1 Gaussian distributions

Consider two strongly isomorphic Gaussian distributions on $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$: the base measure $\mu = N(0, I_2)$ and the target distribution $\nu = N(0, \Sigma)$, where I_2 is the identity matrix and the entries of Σ are $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = 0.7$. We let $c_{\mathcal{X}}(x, x') = x^\top x'$ and $c_{\mathcal{Y}}(y, y') = y^\top \Sigma^{-1} y'$, then two network spaces are strongly isomorphic by design; indeed, any pair (F, B) given by $F(x) = \Sigma^{1/2} Q x$ and $B(y) = Q^\top \Sigma^{-1/2} y$ for $Q \in O(2)$, where $O(2)$ is the orthogonal group, yields $c_{\mathcal{X}}(x, B(y)) = c_{\mathcal{Y}}(F(x), y)$ for all $x, y \in \mathbb{R}^2$, hence F and B are strong isomorphisms. We aim at obtaining such a pair of (linear) isomorphisms by letting $\mathcal{F} = \mathcal{B} = \{x \mapsto Wx : W \in \mathbb{R}^{2\times 2}\}$, that is, the collection of all linear maps from $\mathbb{R}^2$ to $\mathbb{R}^2$. We set $K_{\mathcal{X}} = K_{\mathcal{Y}}$ as a degree-2 polynomial kernel that maps (x, y) to $(x^\top y + 1)^2$; the resulting MMD compares distributions by matching the first two moments, which is sufficient to distinguish Gaussian distributions. The linear maps found by solving the optimization problem (4.3.3) are given by $\widehat{F}(x) = \mathbf{F}x$ and $\widehat{B}(y) = \mathbf{B}y$ for some $\mathbf{F}, \mathbf{B} \in \mathbb{R}^{2\times 2}$ satisfying

$$\mathbf{F}\mathbf{F}^\top - \Sigma \approx \begin{pmatrix} -0.015 & -0.009 \\ -0.009 & -0.007 \end{pmatrix}, \quad \mathbf{B}\Sigma\mathbf{B}^\top - I_2 \approx \begin{pmatrix} 0.023 & -0.011 \\ -0.011 & 0.013 \end{pmatrix},$$
$$\mathbf{F}\mathbf{B} - I_2 \approx \begin{pmatrix} 0.006 & 0.002 \\ 0.004 & 0.002 \end{pmatrix}.$$

Since $\mathbf{F}\mathbf{F}^\top \approx \Sigma$, $\mathbf{B}\Sigma\mathbf{B}^\top \approx I_2$, and $\mathbf{F}\mathbf{B} \approx I_2$, the pair $(\widehat{F}, \widehat{B})$ can be seen as an instance of the pair of strong isomorphisms described above. Figure 4.1 illustrates that $\widehat{F}$ is a strong isomorphism (Definition 4.2.3): (a) shows that $\widehat{F}_{\#}\mu \approx \nu$, that is, $\widehat{F}$ is roughly a transport map, and (b) implies that $c_{\mathcal{X}}(x, x') \approx c_{\mathcal{Y}}(\widehat{F}(x), \widehat{F}(x'))$ holds.



(a) (b)

Figure 4.1: Gaussian experiment: $m = n = 50000$ and $\lambda_1 = \lambda_2 = \lambda_3 = 1$. (a) shows $\{\tilde{y}_j\}_{j=1}^{400}$ versus $\{\widehat{F}(\tilde{x}_i)\}_{i=1}^{400}$, where $\{\tilde{y}_j\}_{j=1}^{400}$ and $\{\tilde{x}_i\}_{i=1}^{400}$ are i.i.d. from $\nu = N(0, \Sigma)$ and $\mu = N(0, I_2)$, respectively; they are new samples independent from $\{y_j\}_{j=1}^{1000}$ and $\{x_i\}_{i=1}^{1000}$ used in (4.3.3). (b) shows the points $\{(c_{\mathcal{X}}(\tilde{x}_i, \tilde{x}_{i'}), c_{\mathcal{Y}}(\widehat{F}(\tilde{x}_i), \widehat{F}(\tilde{x}_{i'})))\}_{i,i'=1}^{40}$ and a straight line $y = x$.

4.6.2 MNIST

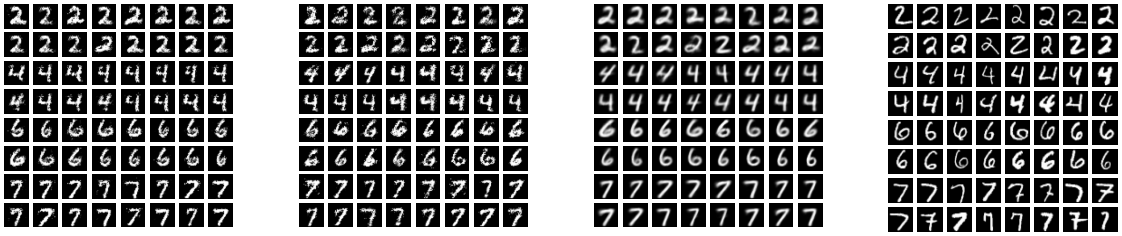
Let ν be the distribution of images corresponding to four digits (2, 4, 6, 7) from the MNIST data set, which is supported on $\mathbb{R}^{784}$. Recall from Section 4.1 that the support $\mathcal{Y}$ of ν is low-dimensional [Facco et al., 2017]; hence, choosing $\mathcal{X} = \mathbb{R}^d$ with $d \ll 784$ is reasonable. Here, for visualization, we first try an extreme embedding task with $d = 2$ and $\mu = N(0, I_2)$, that is, generate MNIST images by transforming two-dimensional Gaussian samples.

Model Specifications Unlike the Gaussian example, where we design the cost functions in advance to make the two spaces strongly isomorphic, specifying them can be more complicated in general cases, which might affect the quality of the RGM sampler. One common choice is the Radial Basis Function (RBF) kernel, also called the heat kernel, widely used in the object matching literature [Solomon et al., 2016]. In this MNIST example, we have found that using RBF kernels as cost functions provides reasonable performance once they are scaled properly. Concretely, first define the RBF kernel $K_d(x, y) = \exp\big(-\|x - y\|^2/d\big)$

for $d \in \mathbb{N}$ and $x, y \in \mathbb{R}^d$; here, the constant $(1/d)$ serves as a scaling factor. Then, we define the cost functions as $c_{\mathcal{X}} = (K_2 - m_{\mathcal{X}})/\mathrm{sd}_{\mathcal{X}}$ and $c_{\mathcal{Y}} = (K_{784} - m_{\mathcal{Y}})/\mathrm{sd}_{\mathcal{Y}}$, where $m_{\mathcal{X}}$ and $\mathrm{sd}_{\mathcal{X}}$ are the median and the standard error of $\{K_{\mathcal{X}}(x_i, x_{i'})\}_{i,i'=1}^m$, respectively; $m_{\mathcal{Y}}$ and $\mathrm{sd}_{\mathcal{Y}}$ are defined analogously. This additional standardization process helps to align the cost functions. Similarly, $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$ must be properly specified; comparing the first two moments using the degree-2 polynomial kernel is no longer sufficient as the target distribution is non-Gaussian. We also suggest using RBF kernels for the MMD terms; let $K_{\mathcal{X}} = K_2$ and $K_{\mathcal{Y}} = K_{784}$. The MMD induced by the RBF kernel indeed defines a metric between distributions under mild assumptions [Muandet et al., 2017], which allows the resulting MMD terms to represent the binding constraint mentioned in Section 4.3. We need richer classes than the linear maps used in the Gaussian case for the function classes $\mathcal{F}$ and $\mathcal{B}$. To this end, we will use the Fully Connected Neural Network (FCNN) with three hidden layers, each consisting of 50 neurons. Lastly, we let $m = n = 20000$.

Key Features of the RGM Sampler Figure 4.2(a) shows the images generated by applying the resulting map $\widehat{F}$ to new i.i.d. samples from $\mu = N(0, I_2)$, which are independent of the samples used for training $\widehat{F}, \widehat{B}$. Though not perfect, we see that recognizable images can be generated by transforming two-dimensional Gaussian samples, efficient in computation.[6] Meanwhile, the map $\widehat{B}$ shows how the MNIST images can be embedded in $\mathbb{R}^2$. Figure 4.3(a) shows $\{\widehat{B}(\tilde{y}_j)\}_{j=1}^{500}$, where $\{\tilde{y}_j\}_{j=1}^{500}$ are drawn from ν (125 images for each digit), which are independent of the samples used for training $\widehat{F}, \widehat{B}$. We see that each digit forms a local cluster in $\mathbb{R}^2$, each of which is roughly representable according to the range of the angular coordinate. Lastly, though not perfect as in Figure 4.1(b) (strongly isomorphic case), Figure 4.3(b) shows that $\widehat{B}$ leads to a reasonable alignment of $c_{\mathcal{X}}(\widehat{B}(y), \widehat{B}(y'))$ versus $c_{\mathcal{Y}}(y, y')$.

6. Computational cost for obtaining $\widehat{F}\colon \mathbb{R}^2 \to \mathbb{R}^{784}$ and computing $\widehat{F}(X)$ from $X \sim \mu$ is far less than that of the OT-based sampler as explained in Section 4.1.

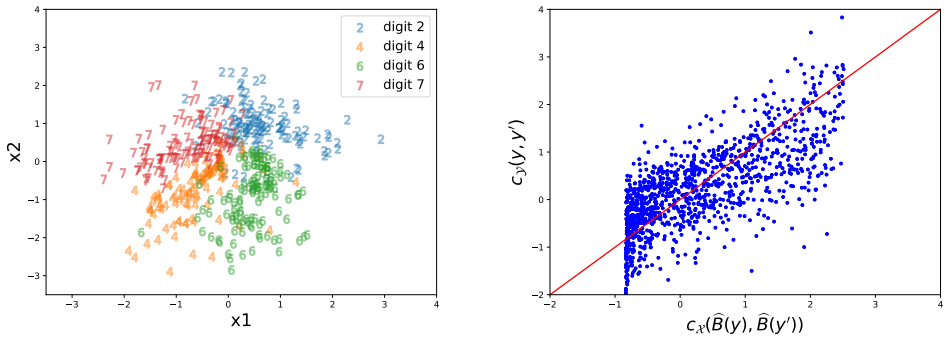


(a) MMD ($\mathbb{R}^2$) (b) MMD ($\mathbb{R}^4$) (c) Sinkhorn ($\mathbb{R}^4$) (d) Original

Figure 4.2: (a) is generated by transforming new i.i.d. samples from $\mu = N(0, I_2)$ using $\widehat{F}$, trained under the aforementioned model specifications. (b) is generated analogously, but with replacing the source dimension $d = 2$ with $d = 4$, namely, $\mu = N(0, I_4)$; for (c), we further replace the three MMD terms in (4.3.3) with Sinkhorn divergences. (d) shows real MNIST images.

Quantitative Evaluation and Ablation Analysis We clarify that the above experiment with $\mu = N(0, I_2)$ is a proof of concept to demonstrate the key features of the RGM sampler. To obtain high-fidelity images comparable to those generated by dedicated MNIST generators, exhaustive tests should be done for tuning every component of the RGM sampler and optimization strategies, which is beyond the scope of this chapter. Instead, we conclude this section by quantitatively evaluating some key components—discrepancy measures, the source dimension, and the neural network architecture—to inspect the performance of the RGM sampler more systematically.

The first component is choosing the discrepancies $\mathcal{L}_{\mathcal{X}}, \mathcal{L}_{\mathcal{Y}}, \mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$. Though we have been using MMD as the leading example, one may use other discrepancies. Here, we consider a modification of (4.3.3) by replacing the MMD terms in (4.3.3) with Sinkhorn divergences, which has shown good performance for generative modeling tasks [Genevay et al., 2018]. Next, we also try higher source dimensions, namely, $\mu = N(0, I_d)$ with $d = 4, 10, 20, 50$. Lastly, we test a different neural network architecture called the Deep Convolutional Neural Network (DCNN), which has shown good empirical performance in representing image-type



(a) (b)

Figure 4.3: (a) is generated by applying $\widehat{B}$ to 500 out-of-sample MNIST images, i.i.d. $\{\tilde{y}_j\}_{j=1}^{500}$ from ν. (b) shows the points $\{(c_{\mathcal{X}}(\widehat{B}(\tilde{y}_j), \widehat{B}(\tilde{y}_{j'})), c_{\mathcal{Y}}(\tilde{y}_j, \tilde{y}_{j'}))\}_{j,j'=1}^{50}$ and a straight line $y = x$.

data [Radford et al., 2015].[7] In summary, we test the combinations of two different discrepancies (MMD and Sinkhorn), five choices for the source dimension $d \in \{2, 4, 10, 20, 50\}$, and two choices for the neural network architecture (FCNN and DCNN). To quantitatively evaluate the performance of $\widehat{F}$ trained under each combination, we should quantify the closeness of $\widehat{F}_{\#}\mu$ and the target distribution ν; though there are many suggestions in the literature for such a task, we take a straightforward approach by comparing the moments and evaluating log likelihood scores.

Table 4.1 summarizes the evaluation scores defined as $2\|\sum_i \widehat{F}(\tilde{x}_i)/\tilde{m} - \sum_j \tilde{y}_j/\tilde{n}\|^2$,[8] where $\{\tilde{x}_i\}_{i=1}^{\tilde{m}}$ and $\{\tilde{y}_j\}_{j=1}^{\tilde{n}}$ are i.i.d. samples from μ and ν, respectively; in words, we estimate the squared difference between the first moments of $\widehat{F}_{\#}\mu$ and ν by using new samples that are independent of the samples used to train $\widehat{F}, \widehat{B}$. By comparing the first two rows, we can observe that replacing the MMD terms with the Sinkhorn divergence leads to smaller (= better) evaluation scores (exception: FCNN with $d = 2$). By comparing Figure 4.2(b) and Figure 4.2(c), we can see that images generated by Sinkhorn divergences

7. Roughly speaking, we adapt the structures of the generator and discriminator in Radford et al. [2015] to construct $\mathcal{F}$ and $\mathcal{B}$, respectively.

8. This formulation is motivated by the energy distance [Székely and Rizzo, 2013], which is obtained by replacing the distance in the energy distance formulation with the squared distance.

d	2	4	10	20	50
FCNN					
Sinkhorn	0.852(0.017)	0.743(0.011)	0.627(0.013)	0.587(0.012)	0.672(0.015)
MMD	0.784(0.015)	0.875(0.014)	0.941(0.020)	1.004(0.018)	1.131(0.021)
MMD-GAN	0.804(0.018)	1.016(0.024)	0.777(0.018)	1.121(0.024)	0.999(0.021)
DCNN					
Sinkhorn	0.841(0.014)	0.729(0.015)	0.671(0.014)	0.695(0.016)	0.647(0.015)
MMD	0.900(0.018)	0.913(0.017)	1.031(0.020)	1.249(0.025)	1.155(0.021)
MMD-GAN	0.927(0.022)	0.825(0.019)	1.044(0.020)	0.841(0.018)	1.069(0.023)
baseline	0.432				

Table 4.1: Evaluation scores using the squared differences between the first moments of the generated distribution $\widehat{F}_{\#}\mu$ and the target distribution ν; these are estimated by new samples $\{\tilde{x}_i\}_{i=1}^{2000}$ from $\mu = N(0, I_d)$ and $\{\tilde{y}_j\}_{j=1}^{2000}$ from the MNIST test set. We repeat the evaluation process for 100 times in each setting, and present the average evaluation score with the corresponding standard deviation in the parenthesis. The baseline is computed by randomly dividing the MNIST test set of 4000 samples into two subsets of 2000 samples and calculating the squared difference between the first moments of the two subsets analogously to the evaluation score. Here smaller evaluation scores imply better empirical results.

tend to have smoother boundaries, which are visually more natural.[9] Next, we can see that the RGM sampler with the MMD terms often (but not always) has smaller evaluation scores than the MMD-GAN [Dziugaite et al., 2015, Li et al., 2015], which aims to solve $\min_F \mathrm{MMD}^2_{K_\mathcal{Y}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n)$, that is, only minimizing the last MMD term associated with $\mathcal{Y}$ in (4.3.3); note that the RGM sampler is better when d is very small, especially when $d = 2$.[10] Increasing the source dimension d or using a more sophisticated neural network architecture DCNN shows mixed results. For instance, for the Sinkhorn one, increasing d often (but not always) leads to smaller (= better) evaluation scores; however, for the MMD one, larger d results in the opposite. Similarly, DCNN often (but not always) leads to better scores for the Sinkhorn one, but the other way around for the MMD one. The main reason for such mixed results is likely related to the optimization; larger d and DCNN usually require

9. Of course, given that many images in MNIST have uneven boundaries, one cannot conclude that Sinkhorn outperforms MMD.

10. One should, however, be reminded the following difference when comparing the RGM sampler and MMD-GAN: the former requires training of both maps F, B, whereas the latter is a minimization over F only.

more sophisticated and tailored optimization schemes. We can similarly analyze different evaluation scores defined by comparing the second moments or computing the log likelihood (estimated using a kernel density estimator [Goodfellow et al., 2014]) to complement Table 4.1, which we present in the supplementary material. Omitted details of this section can also be found therein.

4.7 Discussion

In this chapter, we proposed the Reversible Gromov-Monge (RGM) sampler, a new variant of transform sampling based on the RGM distance operating between distributions on heterogeneous spaces. The RGM sampler fuses the Gromov-Wasserstein (GW) idea with the transform sampling task, aiming at introducing inductive biases towards isomorphisms (see 4.12). We have established statistical results as the main theoretical contribution, along with several supporting results on analytic properties of the RGM distance and the convex relaxation based on the operator viewpoint and the representer theorem.

Lastly, we mention two directions for future research. First, as briefly mentioned in Section 4.3, the RGM sampler is easily implementable by solving a minimization problem via the gradient descent algorithm; however, it is unclear whether one can derive global convergence results from this minimization, which we leave as future research. Next, another interesting direction is to study analytic properties of the RGM distance; investigating deeper connections with the GW distance and strong isomorphisms would be an interesting topic in the GW analysis literature.

4.8 Details of Section 4.4

4.8.1 Proofs in Section 4.4

Proof of Proposition 4.4.4. Let $h_{F,B}(x,y) := (c_{\mathcal{X}}(x,B(y)) - c_{\mathcal{Y}}(F(x),y))^2$, then

$$\begin{aligned}\widehat{C}_0(F,B) - C_0(F,B) &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} h_{F,B}(x_i,y_j) - \mathbb{E}_{(x,y)\sim\mu\otimes\nu} h_{F,B}(x,y) \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n} h_{F,B}(x_i,y_j) - \mathbb{E}_{y\sim\nu} h_{F,B}(x_i,y)\right) \\ &\quad + \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{y\sim\nu} h_{F,B}(x_i,y) - \mathbb{E}_{(x,y)\sim\mu\otimes\nu} h_{F,B}(x,y)\ .\end{aligned}$$

Assumption 4.4.3 implies that a function $x \mapsto \mathbb{E}_{y\sim\nu} h_{F,B}(x,y)$ is bounded in $[0,H]$. Thus, by the McDiarmid's inequality,

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{y\sim\nu} h_{F,B}(x_i,y) - \mathbb{E}_{(x,y)\sim\mu\otimes\nu} h_{F,B}(x,y) \le \sqrt{\frac{H^2\log(1/\delta)}{2m}}$$

holds with probability at least $1-\delta$. By the same logic, for fixed x_i,

$$\frac{1}{n}\sum_{j=1}^{n} h_{F,B}(x_i,y_j) - \mathbb{E}_{y\sim\nu} h_{F,B}(x_i,y) \le \sqrt{\frac{H^2\log(1/\delta)}{2n}}$$

holds with probability at least $1-\delta$, where the probability is the conditional probability of $y_1,\ldots,y_n$ given $x_1,\ldots,x_m$. Since this is true for all x_i, the union bound implies

$$\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n} h_{F,B}(x_i,y_j) - \mathbb{E}_{y\sim\nu} h_{F,B}(x_i,y)\right) \le \sqrt{\frac{H^2\log(m/\delta)}{2n}}$$

holds with probability at least $1-\delta$. Hence,

$$\widehat{C}_0(F,B) - C_0(F,B) \precsim \sqrt{\frac{\log(m/\delta)}{m}} \leq \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}}$$

holds with probability at least $1-2\delta$. The same result holds for $C_0(F,B) - \widehat{C}_0(F,B)$. Hence we complete the proof. □

Proof of Proposition 4.4.6. By the triangle inequality, $|\widehat{M}(F,B) - M(F,B)|$ is bounded above by the sum of the following three terms:

$$\begin{aligned}
&|\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\mu, \nu)| \,, \\
&|\mathrm{MMD}^2_{K_\mathcal{X}}(\widehat{\mu}_m, B_\#\widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{X}}(\mu, B_\#\nu)| \,, \\
&|\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\widehat{\mu}_m, (B,\mathrm{Id})_\#\widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\mu, (B,\mathrm{Id})_\#\nu)| \,.
\end{aligned}$$

First, we give an upper bound on the first term. Boundedness of kernels (Assumption 4.4.5) implies

$$\mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) \,,\ \mathrm{MMD}_{K_\mathcal{Y}}(F_\#\mu, \nu) \leq 2\sqrt{K} \,.$$

Hence,

$$|\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\mu, \nu)| \leq 4\sqrt{K}|\mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}_{K_\mathcal{Y}}(F_\#\mu, \nu)| \,.$$

Due to the triangle inequality of MMD, we have

$$|\mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}_{K_\mathcal{Y}}(F_\#\mu, \nu)| \leq \mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, F_\#\mu) + \mathrm{MMD}_{K_\mathcal{Y}}(\widehat{\nu}_n, \nu) \,.$$

By Theorem 3.4 of Muandet et al. [2017],

$$\mathrm{MMD}_{K_\mathcal{Y}}(\widehat{\nu}_n, \nu) \leq \sqrt{\frac{K}{n}} + \sqrt{\frac{2K\log(1/\delta)}{n}}$$

holds with probability at least $1-\delta$. Next, note that $F_{\#}\widehat{\mu}_m = \frac{1}{m}\sum_i \delta_{F(x_i)}$ is the empirical measure constructed from $\{F(x_i)\}_{i=1}^m$. Since they are m many i.i.d. samples from $F_{\#}\mu$, by the same theorem,

$$\mathrm{MMD}_{K_{\mathcal{Y}}}(F_{\#}\widehat{\mu}_m, F_{\#}\mu) \leq \sqrt{\frac{K}{m}} + \sqrt{\frac{2K\log(1/\delta)}{m}}$$

holds with probability at least $1-\delta$. Hence,

$$|\mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\mu, \nu)| \precsim \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}}$$

holds with probability at least $1-2\delta$. Similarly, we have

$$\begin{aligned}
&|\mathrm{MMD}^2_{K_{\mathcal{X}}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) - \mathrm{MMD}^2_{K_{\mathcal{X}}}(\mu, B_{\#}\nu)| \precsim \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}}\,, \\
&|\mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n) - \mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}((\mathrm{Id}, F)_{\#}\mu, (B, \mathrm{Id})_{\#}\nu)| \\
&\precsim \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}}\,,
\end{aligned}$$

each holds with probability at least $1-2\delta$. Combining these three probabilistic bounds, we obtain a bound for $|\widehat{M}(F,B) - M(F,B)|$. □

Proof of Proposition 4.4.10. Without loss of generality, assume $n \geq m$. From the proof of Proposition 4.4.4,

$$\begin{aligned}
\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F,B) - C_0(F,B)| \leq \frac{1}{m}\sum_{i=1}^m \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left|\frac{1}{n}\sum_{j=1}^n h_{F,B}(x_i, y_j) - \mathbb{E}_{y\sim\nu} h_{F,B}(x_i, y)\right| \\
+ \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left|\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{y\sim\nu} h_{F,B}(x_i, y) - \mathbb{E}_{x\sim\mu}\mathbb{E}_{y\sim\nu} h_{F,B}(x, y)\right|\,.
\end{aligned}$$

Since $x \mapsto \mathbb{E}_{y\sim\nu} h_{F,B}(x,y)$ is bounded in $[0,H]$, Lemma 4.8.1 implies

$$\begin{aligned}
&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{y\sim\nu} h_{F,B}(x_i,y) - \mathbb{E}_{x\sim\mu}\mathbb{E}_{y\sim\nu} h_{F,B}(x,y) \right| \\
&\precsim \sqrt{\frac{\log(1/\delta)}{m}} + \mathbb{E}_{x_i}\mathbb{E}_{\varepsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i \mathbb{E}_{y\sim\nu} h_{F,B}(x_i,y) \right|
\end{aligned}$$

holds with probability at least $1-\delta$. Since $n \geq m$,

$$\begin{aligned}
\mathbb{E}_{x_i}\mathbb{E}_{\varepsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i \mathbb{E}_{y\sim\nu} h_{F,B}(x_i,y) \right| &= \mathbb{E}_{x_i}\mathbb{E}_{\varepsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \mathbb{E}_{y_1,\ldots,y_m} \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i h_{F,B}(x_i,y_i) \right| \\
&\leq \mathbb{E}_{x_i}\mathbb{E}_{\varepsilon_i}\mathbb{E}_{y_1,\ldots,y_m} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i h_{F,B}(x_i,y_i) \right| \\
&= \mathbb{E}_{x_i}\mathbb{E}_{y_i}\mathbb{E}_{\varepsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i h_{F,B}(x_i,y_i) \right| .
\end{aligned}$$

We first give an upper bound on

$$\mathbb{E}_{\varepsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \underbrace{\left| \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i h_{F,B}(x_i,y_i) \right|}_{=:X_{F,B}} .$$

First, observe that Assumption 4.4.3 and Assumption 4.4.9 imply

$$\begin{aligned}
|h_{F,B}(x,y) - h_{F',B'}(x,y)| &\le \left|\sqrt{h_{F,B}(x,y)} + \sqrt{h_{F',B'}(x,y)}\right| \left|\sqrt{h_{F,B}(x,y)} - \sqrt{h_{F',B'}(x,y)}\right| \\
&\le 2\sqrt{H}\left(|c_{\mathcal{X}}(x,B(y)) - c_{\mathcal{X}}(x,B'(y))| + |c_{\mathcal{Y}}(F(x),y) - c_{\mathcal{Y}}(F'(x),y)|\right) \\
&\le 2\sqrt{H}L\left(\|F(x) - F'(x)\| + \|B(y) - B'(y)\|\right) \\
&= 2\sqrt{H}L\left[\sqrt{\sum_{k=1}^{\dim(\mathcal{Y})} |F_k(x) - F'_k(x)|^2} + \sqrt{\sum_{\ell=1}^{\dim(\mathcal{X})} |B_\ell(y) - B'_\ell(y)|^2}\right] \\
&\le 2\sqrt{H}L\left(\sum_{k=1}^{\dim(\mathcal{Y})} |F_k(x) - F'_k(x)| + \sum_{\ell=1}^{\dim(\mathcal{X})} |B_\ell(y) - B'_\ell(y)|\right) .
\end{aligned}$$

Therefore,

$$\begin{aligned}
|X_{F,B} - X_{F',B'}| &\le \frac{1}{m}\sum_{i=1}^{m} |h_{F,B}(x_i,y_i) - h_{F',B'}(x_i,y_i)| \\
&\le 2\sqrt{H}L\left(\sum_{k=1}^{\dim(\mathcal{Y})} \frac{1}{m}\sum_{i=1}^{m} |F_k(x_i) - F'_k(x_i)| + \sum_{\ell=1}^{\dim(\mathcal{X})} \frac{1}{m}\sum_{i=1}^{m} |B_\ell(y_i) - B'_\ell(y_i)|\right) \\
&\le 2\sqrt{H}L\left(\sum_{k=1}^{\dim(\mathcal{Y})} \max_{i\in[m]} |F_k(x_i) - F'_k(x_i)| + \sum_{\ell=1}^{\dim(\mathcal{X})} \max_{i\in[m]} |B_\ell(y_i) - B'_\ell(y_i)|\right) \\
&=: \rho((F,B),(F',B')) .
\end{aligned}$$

For $\varepsilon > 0$, let $\mathcal{N}_\infty(\varepsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m)$ be the minimal ε-covering net of $\mathcal{F}_k$ under the pseudo-metric d induced by $x_1, \ldots, x_m$:

$$d(F_k, F'_k) := \max_{i\in[m]} |F_k(x_i) - F_k(x'_i)| .$$

In other words, for any $F_k \in \mathcal{F}_k$, we can find $F'_k \in \mathcal{N}_\infty(\varepsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m)$ such that $d(F_k, F'_k) \le \varepsilon$. Also, $|\mathcal{N}_\infty(\varepsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m)| = N_\infty(\varepsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m)$. We define $\mathcal{N}_\infty(\varepsilon, \mathcal{B}_\ell, \{y_i\}_{i=1}^m)$ in a similar fashion.

Given $\varepsilon > 0$, let $T_\varepsilon = \otimes_{k=1}^{\dim(\mathcal{Y})} \mathcal{N}_\infty(\varepsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m) \times \otimes_{\ell=1}^{\dim(\mathcal{X})} \mathcal{N}_\infty(\varepsilon, \mathcal{B}_\ell, \{y_i\}_{i=1}^m)$. Then, for any $(F, B) \in \mathcal{F} \times \mathcal{B}$, we can find $(F', B') \in T_\varepsilon$ such that

$$\rho((F, B), (F', B')) \leq \eta\varepsilon \, ,$$

where $\eta = 2\sqrt{H}L(\dim(\mathcal{X}) + \dim(\mathcal{Y}))$. As a result, one can easily check

$$\begin{aligned} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} X_{F,B} &\leq \sup_{\rho((F,B),(F',B'))\leq\eta\varepsilon} |X_{F,B} - X_{F',B'}| + \sup_{(F,B)\in T_\varepsilon} X_{F,B} \\ &\leq \eta\varepsilon + \sup_{(F,B)\in T_\varepsilon} X_{F,B} \, . \end{aligned}$$

Note that $X_{F,B}$ is the absolute value of a sub-Gaussian random variable with parameter $H^2/(4m)$. Hence, the maximal inequality yields

$$\mathbb{E}_{\varepsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} X_{F,B} \leq \eta\varepsilon + \mathbb{E}_{\varepsilon_i} \sup_{(F,B)\in T_\varepsilon} X_{F,B} \leq \eta\varepsilon + \sqrt{\frac{H^2 \log(|T_\varepsilon|)}{m}} \, .$$

Using $|T_\varepsilon| = \prod_{k=1}^{\dim(\mathcal{Y})} N_\infty(\varepsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m) \times \prod_{\ell=1}^{\dim(\mathcal{X})} N_\infty(\varepsilon, \mathcal{B}_\ell, \{y_i\}_{i=1}^m)$, we have

$$\begin{aligned} &\mathbb{E}_{\varepsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i h_{F,B}(x_i, y_i) \right| \\ &\leq \eta\varepsilon + H\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\varepsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\varepsilon, \mathcal{B}_\ell, \{y_i\}_{i=1}^m)}{m}} \\ &\leq \eta\varepsilon + H\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\varepsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\varepsilon, \mathcal{B}_\ell, m)}{m}} \, . \end{aligned}$$

The second inequality is obvious from the definition of the uniform covering number. Since

the last equation is independent of x_i and y_i, we have

$$\begin{aligned}&\mathbb{E}_{x_i}\mathbb{E}_{y_i}\mathbb{E}_{\varepsilon_i}\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}}\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i h_{F,B}(x_i,y_i)\right| \\ &\leq \eta\varepsilon + H\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})}\log N_\infty(\varepsilon,\mathcal{F}_k,m)+\sum_{\ell=1}^{\dim(\mathcal{X})}\log N_\infty(\varepsilon,\mathcal{B}_\ell,m)}{m}}\,.\end{aligned}$$

As a result,

$$\begin{aligned}&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{y\sim\nu}h_{F,B}(x_i,y)-\mathbb{E}_{x\sim\mu}\mathbb{E}_{y\sim\nu}h_{F,B}(x,y)\right| \\ &\precsim \sqrt{\frac{\log(1/\delta)}{m}}+\varepsilon+\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})}\log N_\infty(\varepsilon,\mathcal{F}_k,m)+\sum_{\ell=1}^{\dim(\mathcal{X})}\log N_\infty(\varepsilon,\mathcal{B}_\ell,m)}{m}} \qquad (4.8.1)\\ &\leq \sqrt{\frac{\log(1/\delta)}{m}}+\varepsilon+\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})}\log N_\infty(\varepsilon,\mathcal{F}_k,m)+\sum_{\ell=1}^{\dim(\mathcal{X})}\log N_\infty(\varepsilon,\mathcal{B}_\ell,n)}{m}}\end{aligned}$$

holds with probability at least $1-\delta$. Here, $\log N_\infty(\varepsilon,\mathcal{B}_\ell,m)\leq\log N_\infty(\varepsilon,\mathcal{B}_\ell,n)$ holds since $n\geq m$, which is obvious from the definition of the uniform covering number.

Next, we give a bound on

$$\frac{1}{m}\sum_{i=1}^{m}\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}h_{F,B}(x_i,y_j)-\mathbb{E}_{y\sim\nu}h_{F,B}(x_i,y)\right|\,.$$

Considering $x_1,\ldots,x_m$ are fixed, Lemma 4.8.1 implies

$$\begin{aligned}&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}h_{F,B}(x_i,y_j)-\mathbb{E}_{y\sim\nu}h_{F,B}(x_i,y)\right| \\ &\precsim \sqrt{\frac{\log(1/\delta)}{n}}+\mathbb{E}_{y_j}\mathbb{E}_{\varepsilon_j}\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}}\underbrace{\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j h_{F,B}(x_i,y_j)\right|}_{Y_{F,B}}\end{aligned}$$

holds with probability at least $1-\delta$. Here, the probability should be understood as a

conditional probability of $y_1, \ldots, y_n$ given $x_1, \ldots, x_m$. Again, we have

$$|Y_{F,B} - Y_{F',B'}| \leq 2\sqrt{H}L \left(\sum_{k=1}^{\dim(\mathcal{Y})} |F_k(x_i) - F'_k(x_i)| + \sum_{\ell=1}^{\dim(\mathcal{X})} \frac{1}{n} \sum_{j=1}^{n} |B_\ell(y_j) - B'_\ell(y_j)| \right)$$
$$\leq 2\sqrt{H}L \left(\sum_{k=1}^{\dim(\mathcal{Y})} \max_{i \in [m]} |F_k(x_i) - F'_k(x_i)| + \sum_{\ell=1}^{\dim(\mathcal{X})} \max_{j \in [n]} |B_\ell(y_j) - B'_\ell(y_j)| \right) .$$

Also, $Y_{F,B}$ is the absolute value of a sub-Gaussian random variable with parameter $H^2/(4n)$. By the same argument as before,

$$\mathbb{E}_{y_j} \mathbb{E}_{\varepsilon_j} \sup_{(F,B) \in \mathcal{F} \times \mathcal{B}} \left| \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j h_{F,B}(x_i, y_j) \right|$$
$$\leq \eta\varepsilon + H \sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\varepsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\varepsilon, \mathcal{B}_\ell, n)}{n}} .$$

Hence,

$$\sup_{(F,B) \in \mathcal{F} \times \mathcal{B}} \left| \frac{1}{n} \sum_{j=1}^{n} h_{F,B}(x_i, y_j) - \mathbb{E}_{y \sim \nu} h_{F,B}(x_i, y) \right|$$
$$\precsim \sqrt{\frac{\log(1/\delta)}{n}} + \varepsilon + \sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\varepsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\varepsilon, \mathcal{B}_\ell, n)}{n}}$$

holds with probability (conditional probability as explained earlier) at least $1 - \delta$. Since this holds for all x_i, the union bound implies

$$\frac{1}{m} \sum_{i=1}^{m} \sup_{(F,B) \in \mathcal{F} \times \mathcal{B}} \left| \frac{1}{n} \sum_{j=1}^{n} h_{F,B}(x_i, y_j) - \mathbb{E}_{y \sim \nu} h_{F,B}(x_i, y) \right|$$
$$\precsim \sqrt{\frac{\log(m/\delta)}{n}} + \varepsilon + \sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\varepsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\varepsilon, \mathcal{B}_\ell, n)}{n}} \qquad (4.8.2)$$
$$\leq \sqrt{\frac{\log(m/\delta)}{m}} + \varepsilon + \sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\varepsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\varepsilon, \mathcal{B}_\ell, n)}{m}}$$

holds with probability at least $1-\delta$. Combining (4.8.1) and (4.8.2), for any $\varepsilon > 0$, we have

$$\begin{aligned}
&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F,B) - C_0(F,B)| \\
&\lesssim \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}} + \varepsilon + \sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\varepsilon,\mathcal{F}_k,m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\varepsilon,\mathcal{B}_\ell,n)}{m\wedge n}}
\end{aligned}$$

holds with probability at least $1-2\delta$. □

Proof of Corollary 4.4.12. Combining Assumption 4.4.8 and Lemma 4.8.2, we have

$$N_\infty(\varepsilon,\mathcal{F}_k,m) \le \left(\frac{2emb}{\varepsilon\cdot \mathrm{Pdim}(\mathcal{F}_k)}\right)^{\mathrm{Pdim}(\mathcal{F}_k)}.$$

Hence,

$$\begin{aligned}
&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F,B) - C_0(F,B)| \\
&\lesssim \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}} + \varepsilon + \sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\varepsilon,\mathcal{F}_k,m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\varepsilon,\mathcal{B}_\ell,n)}{m\wedge n}} \\
&\le \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}} + \varepsilon + \sqrt{\frac{\log\left(\frac{2eb(m\vee n)}{\varepsilon}\right)}{m\wedge n}\left(\sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k) + \sum_{\ell=1}^{\dim(\mathcal{X})} \mathrm{Pdim}(\mathcal{B}_\ell)\right)} \\
&\lesssim \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}} + \sqrt{\frac{\log(m\vee n)}{m\wedge n}\left(\sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k) + \sum_{\ell=1}^{\dim(\mathcal{X})} \mathrm{Pdim}(\mathcal{B}_\ell)\right)}
\end{aligned}$$

holds with probability at least $1-2\delta$, where the last bound comes from choosing $\varepsilon = (m\wedge n)^{-1/2}$. □

Proof of Proposition 4.4.14. Using triangle inequality, we bound $\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{M}(F,B) -$

$M(F, B)|$ by the sum of the following three terms:

$$\sup_{F\in\mathcal{F}} |\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\mu, \nu)| \,,$$

$$\sup_{B\in\mathcal{B}} |\mathrm{MMD}^2_{K_\mathcal{X}}(\widehat{\mu}_m, B_\#\widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{X}}(\mu, B_\#\nu)| \,,$$

$$\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\widehat{\mu}_m, (B, \mathrm{Id})_\#\widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\mu, (B, \mathrm{Id})_\#\nu)| \,.$$

As in the proof of Proposition 4.4.6, we have

$$\begin{aligned}
&\sup_{F\in\mathcal{F}} |\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\mu, \nu)| \\
&\leq 4\sqrt{K}\left[\sup_{F\in\mathcal{F}} \mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, F_\#\mu) + \mathrm{MMD}_{K_\mathcal{Y}}(\widehat{\nu}_n, \nu)\right] \,.
\end{aligned}$$

$\mathrm{MMD}_{K_\mathcal{Y}}(\widehat{\nu}_n, \nu)$ has already been bounded in Proposition 4.4.6. For the first term on the RHS, observe that

$$\begin{aligned}
\sup_{F\in\mathcal{F}} \mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, F_\#\mu) &= \sup_{F\in\mathcal{F}} \sup_{f\in\mathcal{H}_\mathcal{Y}(1)} \left|\int f \,\mathrm{d}F_\#\widehat{\mu}_m - \int f \,\mathrm{d}F_\#\mu\right| \\
&= \sup_{F\in\mathcal{F}} \sup_{f\in\mathcal{H}_\mathcal{Y}(1)} \left|\int f\circ F \,\mathrm{d}\widehat{\mu}_m - \int f\circ F \,\mathrm{d}\mu\right| \\
&= \sup_{f\in\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F}} \left|\int f \,\mathrm{d}\widehat{\mu}_m - \int f \,\mathrm{d}\mu\right| \,,
\end{aligned}$$

where the second equality follows from change-of-variables.

First, we show $\mathcal{H}_\mathcal{Y}(1)$ consists of $\sqrt{K}$-uniformly bounded functions. Let $\|\cdot\|_{\mathcal{H}_\mathcal{Y}}$ be the norm of $\mathcal{H}_\mathcal{Y}$ so that $f \in \mathcal{H}_\mathcal{Y}(1)$ is equivalent to $\|f\|_{\mathcal{H}_\mathcal{Y}} \leq 1$. Then, the reproducing property implies

$$|f(y)| \leq \|f\|_{\mathcal{H}_\mathcal{Y}}\sqrt{K_\mathcal{Y}(y, y)} \leq \sqrt{K}$$

for any $f \in \mathcal{H}_\mathcal{Y}(1)$. Accordingly, $\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F}$ also consists of $\sqrt{K}$-uniformly bounded functions.

Hence, Lemma 4.8.1 implies that

$$\begin{aligned}\sup_{F\in\mathcal{F}} \mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, F_\#\mu) &= \sup_{f\in\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F}} \left|\int f \,\mathrm{d}\widehat{\mu}_m - \int f \,\mathrm{d}\mu\right| \\ &\leq 2R_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F}, \mu) + \sqrt{\frac{2K\log(1/\delta)}{m}}\end{aligned}$$

holds with probability at least $1-\delta$. Therefore, combining this with the upper bound on $\mathrm{MMD}_{K_\mathcal{Y}}(\widehat{\nu}_n, \nu)$ derived in Proposition 4.4.6,

$$\sup_{F\in\mathcal{F}} |\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\mu, \nu)| \precsim R_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F}, \mu) + \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}} \tag{4.8.3}$$

holds with probability at least $1-2\delta$. Similarly, we can prove that

$$\sup_{B\in\mathcal{B}} |\mathrm{MMD}^2_{K_\mathcal{X}}(\widehat{\mu}_m, B_\#\widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{X}}(\mu, B_\#\nu)| \precsim R_n(\mathcal{H}_\mathcal{X}(1)\circ\mathcal{B}, \nu) + \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}} \tag{4.8.4}$$

holds with probability at least $1-2\delta$.

Lastly, since $K_\mathcal{X} \otimes K_\mathcal{Y}$ is bounded by K^2, that is,

$$\sup_{(x,y),(x',y')\in\mathcal{X}\times\mathcal{Y}} K_\mathcal{X} \otimes K_\mathcal{Y}((x,y),(x',y')) \leq K^2,$$

by the same argument, we have

$$\begin{aligned}&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\widehat{\mu}_m, (B, \mathrm{Id})_\#\widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\mu, (B, \mathrm{Id})_\#\nu)| \\ &\quad\leq 4K\left[\sup_{F\in\mathcal{F}} \mathrm{MMD}_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\widehat{\mu}_m, (\mathrm{Id}, F)_\#\mu) + \sup_{B\in\mathcal{B}} \mathrm{MMD}_{K_\mathcal{X}\otimes K_\mathcal{Y}}((B, \mathrm{Id})_\#\widehat{\nu}_n, (B, \mathrm{Id})_\#\nu)\right].\end{aligned}$$

Analogously, $\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)$ consists of K-uniformly bounded functions, hence

$$\begin{aligned}
\sup_{F\in\mathcal{F}} \mathrm{MMD}_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\widehat{\mu}_m,(\mathrm{Id},F)_\#\mu) &= \sup_{f\in\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathrm{Id},\mathcal{F})}\left|\int f\,\mathrm{d}\widehat{\mu}_m - \int f\,\mathrm{d}\mu\right| \\
&\le 2R_m(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathrm{Id},\mathcal{F}),\mu) + \sqrt{\frac{2K^2\log(1/\delta)}{m}}\,, \\
\sup_{B\in\mathcal{B}} \mathrm{MMD}_{K_\mathcal{X}\otimes K_\mathcal{Y}}((B,\mathrm{Id})_\#\widehat{\nu}_n,(B,\mathrm{Id})_\#\nu) &= \sup_{f\in\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathcal{B},\mathrm{Id})}\left|\int f\,\mathrm{d}\widehat{\nu}_n - \int f\,\mathrm{d}\nu\right| \\
&\le 2R_n(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathcal{B},\mathrm{Id}),\nu) + \sqrt{\frac{2K^2\log(1/\delta)}{n}}\,,
\end{aligned}$$

each of which holds with probability at least $1-\delta$. Therefore,

$$\begin{aligned}
&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\widehat{\mu}_m,(B,\mathrm{Id})_\#\widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\mu,(B,\mathrm{Id})_\#\nu)| \\
&\precsim R_m(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathrm{Id},\mathcal{F}),\mu) + R_n(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathcal{B},\mathrm{Id}),\nu) + \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}}
\end{aligned} \tag{4.8.5}$$

holds with probability at least $1-2\delta$. We complete the proof by combining (4.8.3), (4.8.4), (4.8.5). □

Proof of Lemma 4.4.15. For any integer $j\in\mathbb{N}\cup\{0\}$, define $\delta_j = 2^{-j}\Delta$, and let $\mathcal{T}_j\subset\mathbb{S}_0^{m\times m}$ be a minimal δ_j-covering net of $\mathcal{T}$; clearly, $|\mathcal{T}_j| = N(\delta_j,\mathcal{T})$. For each j, the covering set induces a mapping $\Pi_j:\mathcal{T}\to\mathcal{T}_j$ such that

$$\sup_{A\in\mathcal{T}} d(A,\Pi_j(A)) \le \delta_j\,.$$

By definition of Δ, we may assume $\mathcal{T}_0 = \{A_0\}$ so that $\Pi_0(A) = A_0$ for all $A\in\mathcal{T}$.

Note that $\mathbb{E}g^\top A_0 g = 0$ by definition. Using this, we write $\mathbb{E}\sup_{A\in\mathcal{T}} g^\top A g$ as a chaining

sum:

$$\begin{aligned}
\mathbb{E}\sup_{A\in\mathcal{T}} g^\top Ag &= \mathbb{E}_g \sup_{A\in\mathcal{T}} g^\top Ag - \mathbb{E}g^\top A_0 g \\
&= \mathbb{E}\sup_{A\in\mathcal{T}} \left(g^\top Ag - g^\top A_0 g\right) \\
&= \mathbb{E}\sup_{A\in\mathcal{T}} \left(g^\top Ag - g^\top \Pi_J(A)g + \sum_{j=0}^{J-1} g^\top \Pi_{j+1}(A)g - g^\top \Pi_j(A)g\right) \\
&\le \mathbb{E}\sup_{A\in\mathcal{T}} \left(g^\top Ag - g^\top \Pi_J(A)g\right) + \sum_{j=0}^{J-1} \mathbb{E}\sup_{A\in\mathcal{T}} \left(g^\top \Pi_{j+1}(A)g - g^\top \Pi_j(A)g\right) \ .
\end{aligned}$$

For the first term on RHS, using the Cauchy-Schwarz inequality and Jensen's inequality, we have

$$\mathbb{E}\sup_{A\in\mathcal{T}} \left(g^\top Ag - g^\top \Pi_J(A)g\right) \le \mathbb{E}\left[\left(\sum_{i\neq j} g_i^2 g_j^2\right)^{1/2} \cdot \delta_J\right] \le m\delta_J \ .$$

For each summand in the second term on RHS, use Lemma 4.8.4. Note that for any j, the maximal cardinality of $\{(\Pi_{j+1}(A), \Pi_j(A)) \ : \ A \in \mathcal{T}\}$ satisfies

$$|\{(\Pi_{j+1}(A), \Pi_j(A)) \ : \ A \in \mathcal{T}\}| \le N(\delta_{j+1}, \mathcal{T}) \times N(\delta_j, \mathcal{T}) \le N(\delta_{j+1}, \mathcal{T})^2$$

and that

$$d(\Pi_{j+1}(A), \Pi_j(A)) \le d(\Pi_{j+1}(A), A) + d(A, \Pi_j(A)) \le 3\delta_{j+1} \ .$$

Since $\Pi_{j+1}(A) - \Pi_j(A) \in \mathbb{S}_0^{m\times m}$ and $\|\Pi_{j+1}(A) - \Pi_j(A)\| \le 3\delta_{j+1}$, Lemma 4.8.4 asserts that for any j

$$\mathbb{E}\sup_{A\in\mathcal{T}} \left(g^\top \Pi_{j+1}(A)g - g^\top \Pi_j(A)g\right) \le 6\delta_{j+1}\sqrt{2\log N(\delta_{j+1}, \mathcal{T})} + 12\delta_{j+1}\log N(\delta_{j+1}, \mathcal{T}) \ .$$

Summing over j, we have for any J, the following inequality,

$$\mathbb{E} \sup_{A \in \mathcal{T}} \left(g^\top A g - g^\top A_0 g \right) \leq m\delta_J + 12 \int_{\delta_J/2}^{\Delta/2} \sqrt{2 \log N(\delta, \mathcal{T})} \, \mathrm{d}\delta + 24 \int_{\delta_J/2}^{\Delta/2} \log N(\delta, \mathcal{T}) \, \mathrm{d}\delta \ .$$

□

Proof of Proposition 4.4.18. To make use of the chaining inequality, we are only left to bound the covering number $N(\delta, \mathcal{T})$ with

$$\mathcal{T} := \{A_F : F \in \mathcal{F}\} \subset \mathbb{S}_0^{m \times m} \ .$$

Lipschitzness of $K_{\mathcal{Y}}$ (Assumption 4.4.16) implies

$$\left| K_{\mathcal{Y}}(F(x_i), F(x_j)) - K_{\mathcal{Y}}(F'(x_i), F'(x_j)) \right| \leq \frac{L}{2} \|F(x_i) - F'(x_i)\| + \frac{L}{2} \|F(x_j) - F'(x_j)\| \ .$$

Hence,

$$\begin{aligned}
d(A_F, A_{F'}) &\leq m \max_{i \neq j} \left| K_{\mathcal{Y}}(F(x_i), F(x_j)) - K_{\mathcal{Y}}(F'(x_i), F'(x_j)) \right| \\
&\leq \frac{mL}{2} \left(\max_{i \in [m]} \|F(x_i) - F'(x_i)\| + \max_{j \in [m]} \|F(x_j) - F'(x_j)\| \right) \\
&= mL \max_{i \in [m]} \|F(x_i) - F'(x_i)\| \\
&\leq mL \max_{i \in [m]} \left(\sum_{k=1}^{\dim(\mathcal{Y})} |F_k(x_i) - F_k'(x_i)|^2 \right)^{1/2} \\
&\leq mL \left[\sum_{k=1}^{\dim(\mathcal{Y})} \left(\max_{i \in [m]} |F_k(x_i) - F_k'(x_i)| \right)^2 \right]^{1/2} \ .
\end{aligned}$$

As in the proof of Proposition 4.4.10, for $\varepsilon > 0$, let $\mathcal{N}_\infty(\varepsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m)$ be a minimal ε-

covering net of $\mathcal{F}_k$. Then, one can easily see that

$$\left\{ A_F : F \in \otimes_{k=1}^{\dim(\mathcal{Y})} \mathcal{N}_\infty \left(\tfrac{\delta}{mL\sqrt{\dim(\mathcal{Y})}}, \mathcal{F}_k, \{x_i\}_{i=1}^m \right) \right\}$$

is a δ-covering of $\mathcal{T}$. Therefore, we conclude

$$N(\delta, \mathcal{T}) \leq \prod_{k=1}^{\dim(\mathcal{Y})} N_\infty \left(\tfrac{\delta}{mL\sqrt{\dim(\mathcal{Y})}}, \mathcal{F}_k, \{x_i\}_{i=1}^m \right) \leq \prod_{k=1}^{\dim(\mathcal{Y})} N_\infty \left(\tfrac{\delta}{mL\sqrt{\dim(\mathcal{Y})}}, \mathcal{F}_k, m \right) .$$

Lastly, we bound $N_\infty(\delta, \mathcal{F}_k, m)$ by the pseudo-dimension of $\mathcal{F}_k$ via Lemma 4.8.2. Combining Assumption 4.4.8 and Lemma 4.8.2, we have

$$N(\delta, \mathcal{T}) \leq \prod_{k=1}^{\dim(\mathcal{Y})} \left(\frac{2emb}{\frac{\delta}{mL\sqrt{\dim(\mathcal{Y})}} \cdot \mathrm{Pdim}(\mathcal{F}_k)} \right)^{\mathrm{Pdim}(\mathcal{F}_k)} .$$

Now, to apply Lemma 4.4.15, fix $F_0 \in \mathcal{F}$ and let $A_0 = A_{F_0}$. Then,

$$\int_{\delta_J/2}^{\Delta/2} \log N(\delta, \mathcal{T}) \, \mathrm{d}\delta \leq \Delta/2 \cdot \sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k) \cdot \log \left(\frac{2emb}{\frac{1}{mL\sqrt{\dim(\mathcal{Y})}} \Delta 2^{-J}/2 \cdot \mathrm{Pdim}(\mathcal{F}_k)} \right) .$$

First, we obtain an upper bound on Δ:

$$\Delta = \sup_{F \in \mathcal{F}} \|A_F - A_0\| \leq m \max_{i \neq j} \left| K_{\mathcal{Y}}(F(x_i), F(x_j)) - K_{\mathcal{Y}}(F_0(x_i), F_0(x_j)) \right| \leq 2mK .$$

Next, we claim that Δ is bounded below by a universal constant; this is to upper bound Δ in the denominator. Consider y_0 and y_0' given in Assumption 4.4.17. We may assume that F_0 is the constant map explained in Assumption 4.4.17: $F_0(x) = y_0$ for all $x \in \mathcal{X}$. Without loss of generality, we assume $x_1 \neq x_2$. Then, we can find $F \in \mathcal{F}$ such that $F(x_1) = y_0$ and

$F(x_2) = y_0'$ according to Assumption 4.4.17. Hence,

$$\Delta \geq |K_{\mathcal{Y}}(F(x_1), F(x_2)) - K_{\mathcal{Y}}(F_0(x_1), F_0(x_2))| = |K_{\mathcal{Y}}(y_0, y_0') - K_{\mathcal{Y}}(y_0, y_0)| > 0 \,.$$

Therefore, with the choice of J such that $m2^{-J} \asymp \sum_k \operatorname{Pdim}(\mathcal{F}_k)$,

$$\begin{aligned}\int_{\delta_J/2}^{\Delta/2} \log N(\delta, \mathcal{T}) \,\mathrm{d}\delta &\precsim m \left[\sum_{k=1}^{\dim(\mathcal{Y})} \operatorname{Pdim}(\mathcal{F}_k)\right] \cdot \log\left(\frac{m}{\min_{k\in[\dim(\mathcal{Y})]} \operatorname{Pdim}(\mathcal{F}_k)}\right) \\ &\leq m \log(m) \left[\sum_{k=1}^{\dim(\mathcal{Y})} \operatorname{Pdim}(\mathcal{F}_k)\right] \,.\end{aligned}$$

Analogously,

$$\int_{\delta_J/2}^{\Delta/2} \sqrt{2 \log N(\delta, \mathcal{T})} \,\mathrm{d}\delta \precsim m \log(m) \left[\sum_{k=1}^{\dim(\mathcal{Y})} \operatorname{Pdim}(\mathcal{F}_k)\right] \,.$$

Thus,

$$R_m(\mathcal{H}_y(1) \circ \mathcal{F}, \{x_i\}_{i=1}^m) \precsim \frac{1}{m}\left[mK + \mathbb{E}_g \sup_{F\in\mathcal{F}} g^\top A_F g\right]^{1/2} \precsim \sqrt{\frac{\log m}{m} \sum_{k=1}^{\dim(\mathcal{Y})} \operatorname{Pdim}(\mathcal{F}_k)}$$

The same argument can be applied to the other three Rademacher complexities. Hence, we have proved the proposition. □

4.8.2 Auxiliary Lemmas

Lemma 4.8.1 (Theorem 4.10 of Wainwright [2019])**.** *Let $(\mathcal{Z}, \rho)$ be a probability space and $\mathcal{G}$ be a class of b-uniformly bounded measurable functions defined on $\mathcal{Z}$, that is, $\sup_{g\in\mathcal{G}} \|g\|_\infty \leq b$. Let $z_1, \ldots, z_m$ are i.i.d. samples from ρ and let $\widehat{\rho}_m$ be the empirical measure constructed*

from them. Then, for any $\delta > 0$*,*

$$\sup_{g \in \mathcal{G}} \left| \int g \, \mathrm{d}\widehat{\rho}_m - \int g \, \mathrm{d}\rho \right| \leq 2R_m(\mathcal{G}, \rho) + \sqrt{\frac{2b^2 \log(1/\delta)}{m}}$$

holds with probability at least $1 - \delta$*.*

Lemma 4.8.2 (Theorem 12.2 of Anthony and Bartlett [1999])**.** *Let* $\mathcal{G}$ *be a collection of real-valued functions defined on a set* $\mathcal{Z}$*. Suppose* $\sup_{g \in \mathcal{G}} ||g||_\infty = b < \infty$*. For* $\varepsilon > 0$ *and* $m \geq \mathrm{Pdim}(\mathcal{G})$*,*

$$N_\infty(\varepsilon, \mathcal{G}, m) \leq \left(\frac{2emb}{\varepsilon \cdot \mathrm{Pdim}(\mathcal{G})} \right)^{\mathrm{Pdim}(\mathcal{G})} .$$

Lemma 4.8.3 (Example 2.12 of Boucheron et al. [2013])**.** *For any* $A \in \mathbb{S}_0^{m \times m}$ *and* $0 \leq \lambda < 1/(2\|A\|_{\mathrm{op}})$*,*

$$\log \mathbb{E}_g e^{\lambda g^\top A g} \leq \frac{\lambda^2 \|A\|}{1 - 2\lambda \|A\|_{\mathrm{op}}} , \tag{4.8.6}$$

where $g \sim N(0, I_m)$*. Here,* $\|\cdot\|_{\mathrm{op}}$ *denotes the operator norm of* A*.*

This lemma tells that $g^\top A g$ is a sub-Gamma random variable with variance factor $2\|A\|^2$ and scale parameter $2\|A\|_{\mathrm{op}}$ (see Chapter 2.4 of Boucheron et al. [2013] for the definition). Using Corollary 2.6 of the same text, we can derive the following maximal inequality.

Lemma 4.8.4 (Maximal inequality)**.** *For* $A_1, \ldots, A_N \in \mathbb{S}_0^{m \times m}$*, suppose* $\max_{i=1,\ldots,N} \|A_i\| \leq \delta$*. Then,*

$$\mathbb{E}_g \max_{i=1,\ldots,N} g^\top A_i g \leq 2\delta \left(\sqrt{\log N} + \log N \right) , \tag{4.8.7}$$

where $g \sim N(0, I_m)$*.*

4.9 Details of Section 4.5.1

4.9.1 Metric Properties and Some Basic Properties

In this section, we derive the metric properties of the proposed RGM distance. First, observe that our RGM is symmetric while the original GM is not. Next, we prove a triangle inequality using a gluing argument.

Proposition 4.9.1. *RGM satisfies the triangle inequality, that is,*

$$\mathrm{RGM}(\mu_{\mathcal{X}}, \mu_{\mathcal{Z}}) \leq \mathrm{RGM}(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}}) + \mathrm{RGM}(\mu_{\mathcal{Y}}, \mu_{\mathcal{Z}})$$

holds for three network spaces $(\mathcal{X}, \mu_{\mathcal{X}}, c_{\mathcal{X}})$, $(\mathcal{Y}, \mu_{\mathcal{Y}}, c_{\mathcal{Y}})$, *and* $(\mathcal{Z}, \mu_{\mathcal{Z}}, c_{\mathcal{Z}})$.

Proof of Proposition 4.9.1. Recall that

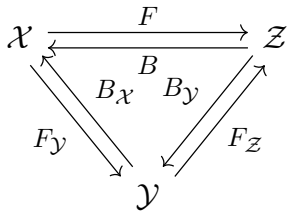
$$\mathrm{RGM}(\mu_{\mathcal{X}}, \mu_{\mathcal{Z}}) = \inf_{(F,B)\in\mathcal{I}(\mu_{\mathcal{X}},\mu_{\mathcal{Z}})} C_{\mathcal{X}\mathcal{Z}}(F, B) \ ,$$

where

$$C_{\mathcal{X}\mathcal{Z}}(F, B) = \left(\int (c_{\mathcal{X}}(x, B(z)) - c_{\mathcal{Z}}(F(x), z))^2 \, \mathrm{d}\mu_{\mathcal{X}} \otimes \mu_{\mathcal{Z}} \, (x, z) \right)^{1/2} .$$

First, $(F_{\mathcal{Z}} \circ F_{\mathcal{Y}}, B_{\mathcal{X}} \circ B_{\mathcal{Y}}) \in \mathcal{I}(\mu_{\mathcal{X}}, \mu_{\mathcal{Z}})$ holds for $(F_{\mathcal{Y}}, B_{\mathcal{X}}) \in \mathcal{I}(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})$ and $(F_{\mathcal{Z}}, B_{\mathcal{Y}}) \in \mathcal{I}(\mu_{\mathcal{Y}}, \mu_{\mathcal{Z}})$ since

$$\begin{aligned}
(\mathrm{Id}, F_{\mathcal{Z}} \circ F_{\mathcal{Y}})_{\#}\mu_{\mathcal{X}} &= (\mathrm{Id}, F_{\mathcal{Z}})_{\#}(\mathrm{Id}, F_{\mathcal{Y}})_{\#}\mu_{\mathcal{X}} \\
&= (\mathrm{Id}, F_{\mathcal{Z}})_{\#}(B_{\mathcal{X}}, \mathrm{Id})_{\#}\mu_{\mathcal{Y}} \quad (\because (F_{\mathcal{Y}}, B_{\mathcal{X}}) \in \mathcal{I}(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})) \\
&= (B_{\mathcal{X}}, \mathrm{Id})_{\#}(\mathrm{Id}, F_{\mathcal{Z}})_{\#}\mu_{\mathcal{Y}} \\
&= (B_{\mathcal{X}}, \mathrm{Id})_{\#}(B_{\mathcal{Y}}, \mathrm{Id})_{\#}\mu_{\mathcal{Z}} \quad (\because (F_{\mathcal{Z}}, B_{\mathcal{Y}}) \in \mathcal{I}(\mu_{\mathcal{Y}}, \mu_{\mathcal{Z}})) \\
&= (B_{\mathcal{X}} \circ B_{\mathcal{Y}}, \mathrm{Id})_{\#}\mu_{\mathcal{Z}} \ .
\end{aligned}$$



Moreover, in this case, we have

$$C_{\mathcal{X}\mathcal{Z}}(F_{\mathcal{Z}} \circ F_{\mathcal{Y}}, B_{\mathcal{X}} \circ B_{\mathcal{Y}}) \leq C_{\mathcal{X}\mathcal{Y}}(F_{\mathcal{Y}}, B_{\mathcal{X}}) + C_{\mathcal{Y}\mathcal{Z}}(F_{\mathcal{Z}}, B_{\mathcal{Y}})$$

since

$$\begin{aligned}
&C_{\mathcal{X}\mathcal{Z}}(F_{\mathcal{Z}} \circ F_{\mathcal{Y}}, B_{\mathcal{X}} \circ B_{\mathcal{Y}}) \\
&= \left(\int \left[c_{\mathcal{X}}(x, B_{\mathcal{X}} \circ B_{\mathcal{Y}}(z)) - c_{\mathcal{Z}}(F_{\mathcal{Z}} \circ F_{\mathcal{Y}}(x), z) \right]^2 \mathrm{d}\mu_{\mathcal{X}} \otimes \mu_{\mathcal{Z}}\,(x, z) \right)^{1/2} \\
&\leq \left(\int \int \left[c_{\mathcal{Y}}(x, B_{\mathcal{X}} \circ B_{\mathcal{Y}}(z)) - c_{\mathcal{Y}}(F_{\mathcal{Y}}(x), B_{\mathcal{Y}}(z)) \right]^2 \mathrm{d}\mu_{\mathcal{Z}}\,(z)\, \mathrm{d}\mu_{\mathcal{X}}\,(x) \right)^{1/2} \\
&\quad + \left(\int \int \left[c_{\mathcal{Y}}(F_{\mathcal{Y}}(x), B_{\mathcal{Y}}(z)) - c_{\mathcal{Z}}(F_{\mathcal{Z}} \circ F_{\mathcal{Y}}(x), z) \right]^2 \mathrm{d}\mu_{\mathcal{X}}\,(x)\, \mathrm{d}\mu_{\mathcal{Z}}\,(z) \right)^{1/2} \\
&= \left(\int \int \left[c_{\mathcal{Y}}(x, B_{\mathcal{X}}(y)) - c_{\mathcal{Y}}(F_{\mathcal{Y}}(x), y) \right]^2 \mathrm{d}\mu_{\mathcal{Y}}\,(y)\, \mathrm{d}\mu_{\mathcal{X}}\,(x) \right)^{1/2} \quad (\because (B_{\mathcal{Y}})_{\#}\mu_{\mathcal{Z}} = \mu_{\mathcal{Y}}) \\
&\quad + \left(\int \int \left[c_{\mathcal{Y}}(y, B_{\mathcal{Y}}(z)) - c_{\mathcal{Z}}(F_{\mathcal{Z}}(y), z) \right]^2 \mathrm{d}\mu_{\mathcal{Y}}\,(y)\, \mathrm{d}\mu_{\mathcal{Z}}\,(z) \right)^{1/2} \quad (\because (F_{\mathcal{Y}})_{\#}\mu_{\mathcal{X}} = \mu_{\mathcal{Y}}) \\
&= C_{\mathcal{X}\mathcal{Y}}(F_{\mathcal{Y}}, B_{\mathcal{X}}) + C_{\mathcal{Y}\mathcal{Z}}(F_{\mathcal{Z}}, B_{\mathcal{Y}})\,.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathrm{RGM}(\mu_{\mathcal{X}}, \mu_{\mathcal{Z}}) &= \inf_{(F,B) \in \mathcal{I}(\mu_{\mathcal{X}}, \mu_{\mathcal{Z}})} C_{\mathcal{X}\mathcal{Z}}(F, B) \\
&\leq \inf_{\substack{(F_{\mathcal{Y}}, B_{\mathcal{X}}) \in \mathcal{I}(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}}) \\ (F_{\mathcal{Z}}, B_{\mathcal{Y}}) \in \mathcal{I}(\mu_{\mathcal{Y}}, \mu_{\mathcal{Z}})}} C_{\mathcal{X}\mathcal{Z}}(F_{\mathcal{Z}} \circ F_{\mathcal{Y}}, B_{\mathcal{X}} \circ B_{\mathcal{Y}}) \\
&\leq \inf_{(F_{\mathcal{Y}}, B_{\mathcal{X}}) \in \mathcal{I}(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})} C_{\mathcal{X}\mathcal{Y}}(F_{\mathcal{Y}}, B_{\mathcal{X}}) + \inf_{(F_{\mathcal{Z}}, B_{\mathcal{Y}}) \in \mathcal{I}(\mu_{\mathcal{Y}}, \mu_{\mathcal{Z}})} C_{\mathcal{Y}\mathcal{Z}}(F_{\mathcal{Z}}, B_{\mathcal{Y}}) \\
&= \mathrm{RGM}(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}}) + \mathrm{RGM}(\mu_{\mathcal{Y}}, \mu_{\mathcal{Z}})\,.
\end{aligned}$$

□

Next, we study whether $\mathrm{RGM}(\mu,\nu)=0$ holds if and only if $(\mathcal{X},\mu,c_{\mathcal{X}})\cong(\mathcal{Y},\nu,c_{\mathcal{Y}})$. Here the equivalence relation induced by $\cong$ can be read from Definition 4.2.3. As in the Gromov-Wasserstein distance, in general, we can only assert the if part without further conditions. The following proposition states that $\mathrm{RGM}(\mu,\nu)=0$ if and only if $(\mathcal{X},\mu,c_{\mathcal{X}})\cong(\mathcal{Y},\nu,c_{\mathcal{Y}})$ under some additional conditions on $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$, thereby implying Theorem 4.5.3.

Proposition 4.9.2. *Let $(\mathcal{X},\mu,c_{\mathcal{X}})$ and $(\mathcal{Y},\nu,c_{\mathcal{Y}})$ be two network spaces. If $(\mathcal{X},\mu,c_{\mathcal{X}})\cong(\mathcal{Y},\nu,c_{\mathcal{Y}})$, then $\mathrm{RGM}(\mu,\nu)=0$. The converse is true if there exists a continuous and strictly monotone function $h\colon\mathbb{R}_+\to\mathbb{R}$ such that $c_{\mathcal{X}}=h(d_{\mathcal{X}})$ and $c_{\mathcal{Y}}=h(d_{\mathcal{Y}})$.*

Proof of Proposition 4.9.2. Suppose $\mathrm{RGM}(\mu,\nu)=0$. Due to the inequality $\mathrm{GW}(\mu,\nu)\leq\mathrm{RGM}(\mu,\nu)$, we have $\mathrm{GW}(\mu,\nu)=0$, that is,

$$\inf_{\gamma\in\Pi(\mu,\nu)}\left(\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}}(h(d_{\mathcal{X}}(x,x'))-h(d_{\mathcal{Y}}(y,y')))^2\,\mathrm{d}\gamma\,(x,y)\,\mathrm{d}\gamma\,(x',y')\right)^{1/2}=0\,.$$

Since there exists a coupling $\gamma^\star$ that achieves the minimum of GW due to Theorem 2.2 of Chowdhury and Mémoli [2019], we conclude

$$h(d_{\mathcal{X}}(x,x'))=h(d_{\mathcal{Y}}(y,y'))$$

holds $\gamma^\star\otimes\gamma^\star$ almost surely on $(\mathcal{X}\times\mathcal{Y})^2$. Since h is strictly monotone, this means

$$d_{\mathcal{X}}(x,x')=d_{\mathcal{Y}}(y,y')$$

holds $\gamma^\star \otimes \gamma^\star$ almost surely on $(\mathcal{X} \times \mathcal{Y})^2$. Therefore,

$$\begin{aligned}
&\inf_{\gamma \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} (d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}(y,y'))^2 \, \mathrm{d}\gamma\,(x,y)\, \mathrm{d}\gamma\,(x',y') \right)^{1/2} \\
&\geq \left(\int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} (d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}(y,y'))^2 \, \mathrm{d}\gamma^\star\,(x,y)\, \mathrm{d}\gamma^\star\,(x',y') \right)^{1/2} \\
&= 0 \, .
\end{aligned}$$

Theorem 4.2.4 implies that metric measure spaces $(\mathcal{X}, \mu, d_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, d_{\mathcal{Y}})$ are strongly isomorphic. Since $c_{\mathcal{X}} = h(d_{\mathcal{X}})$ and $c_{\mathcal{Y}} = h(d_{\mathcal{Y}})$, it follows easily that $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ are strongly isomorphic as well.

To prove the if part, suppose $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ are strongly isomorphic and consider a strong isomorphism T. Then, $(T, T^{-1}) \in \mathcal{I}(\mu, \nu)$ holds since $(\mathrm{Id}, T)_{\#}\mu = (T^{-1}, \mathrm{Id})_{\#} T_{\#}\mu = (T^{-1}, \mathrm{Id})_{\#}\nu$. Also, by definition of T, we have $c_{\mathcal{X}}(x, T^{-1}(y)) = c_{\mathcal{Y}}(T(x), T \circ T^{-1}(y)) = c_{\mathcal{Y}}(T(x), y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, thus

$$\mathrm{RGM}(\mu,\nu) \leq \int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x, T^{-1}(y)) - c_{\mathcal{Y}}(T(x), y))^2 \, \mathrm{d}\mu \otimes \nu = 0 \, .$$

□

We conclude this section with a few more properties and examples. We first complete the proof of Proposition 4.5.4, which characterizes the relations among three distances: GW, GM, and RGM.

Proof of Proposition 4.5.4. Define

$$Q(\gamma) = \int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y'))^2 \, \mathrm{d}\gamma\,(x,y)\, \mathrm{d}\gamma\,(x',y')$$

for all $\gamma \in \Pi(\mu, \nu)$ so that

$$\mathrm{GW}(\mu,\nu)^2 = \inf_{\gamma \in \Pi(\mu,\nu)} Q(\gamma) \, .$$

Recall that $\Pi_{\mathcal{T}} = \{(\mathrm{Id}, T)_{\#}\mu : T \in \mathcal{T}(\mu, \nu)\} \subset \Pi(\mu, \nu)$. Hence, as noted in Section 4.2,

$$\mathrm{GM}(\mu, \nu)^2 = \inf_{\gamma \in \Pi_{\mathcal{T}}} Q(\gamma) \ .$$

Define $\Pi' = \{\gamma \in \Pi(\mu, \nu) : \gamma = (\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu \ \exists (F, B) \in \mathcal{I}(\nu, \mu)\}$, then one can check

$$\mathrm{RGM}(\mu, \nu)^2 = \inf_{\gamma \in \Pi'} Q(\gamma)$$

using change-of-variables. Note that Π' may be rewritten as $\{\gamma \in \Pi_{\mathcal{T}} : \gamma = (B, \mathrm{Id})_{\#}\nu \ \exists B \in \mathcal{T}(\nu, \mu)\}$. Hence, $\Pi' \subseteq \Pi_{\mathcal{T}} \subseteq \Pi(\mu, \nu)$, thus we conclude $\mathrm{GW}(\mu, \nu) \leq \mathrm{GM}(\mu, \nu) \leq \mathrm{RGM}(\mu, \nu)$. □

In short, Proposition 4.5.4 shows that RGM, GM, and GW are minimization problems of a common objective function Q over different constraint sets of couplings, namely, $\Pi' \subseteq \Pi_{\mathcal{T}} \subseteq \Pi(\mu, \nu)$, respectively.

Next, we discuss the binding constraint $(\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$, or equivalently, the feasible set $\mathcal{I}(\mu, \nu)$. One should notice that $\mathcal{I}(\mu, \nu)$ might be empty, for instance, if μ and ν are discrete and their supports have different cardinality; say, $\mu = \delta_x$ and $\nu = (\delta_{y_1} + \delta_{y_2})/2$, namely, Dirac measures supported on $x \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$, then even $\mathcal{T}(\mu, \nu)$ is empty, meaning that the Monge problem is infeasible and so is RGM. On the flip side, the following lemma gives a sufficient condition for $(F, B) \in \mathcal{I}(\mu, \nu)$ which can be useful in practice.

Lemma 4.9.3. *Let $(F, B) \in \mathcal{T}(\mu, \nu) \times \mathcal{T}(\nu, \mu)$. If $F \circ B = \mathrm{Id}$ or $B \circ F = \mathrm{Id}$ holds, then $(F, B) \in \mathcal{I}(\mu, \nu)$.*

Proof. Without loss of generality, assume $B \circ F = \mathrm{Id}$. Then,

$$(\mathrm{Id}, F)_{\#}\mu = (B \circ F, F)_{\#}\mu = (B, \mathrm{Id})_{\#}(F_{\#}\mu) = (B, \mathrm{Id})_{\#}\nu \ .$$

Hence, $(F, B) \in \mathcal{I}(\mu, \nu)$. □

The following example illustrates that this condition can be used to find a pair $(F, B) \in \mathcal{I}(\mu, \nu)$ when μ and ν are Gaussian distributions.

Example 4.9.4. Given $p < q$, suppose $\mu = N(0, I_p)$ and $\nu = N(0, \Sigma)$, where $I_p \in \mathbb{R}^{p\times p}$ is the identity matrix and $\Sigma \in \mathbb{R}^{q\times q}$ is of rank p. Then, we can find a rank-p matrix $A \in \mathbb{R}^{q\times p}$ such that $\Sigma = AA^\top$. Let $F(x) = Ax$ and $B(y) = A^\dagger y$, then one can easily check $F_\# \mu = \nu$, $B_\# \nu = \mu$, and $B \circ F = \mathrm{Id}$. Hence, $(F, B) \in \mathcal{I}(\mu, \nu)$.

Lastly, we provide a simple example that shows that properly chosen cost functions give a strong isomorphism between two Gaussian distributions.

Example 4.9.5. Consider two Gaussians on $\mathbb{R}^d$, say $\mu = N(0, \Sigma_1)$ and $\nu = N(0, \Sigma_2)$. Assume Σ_1 and Σ_2 are invertible. Then two network spaces $(\mathbb{R}^d, \mu, c_\mathcal{X})$ and $(\mathbb{R}^d, \nu, c_\mathcal{Y})$ are strongly isomorphic if $c_\mathcal{X}$ and $c_\mathcal{Y}$ are Mahalanobis distances, that is,

$$c_\mathcal{X}(x, x') = \sqrt{(x - x')^\top \Sigma_1^{-1}(x - x')}\,, \quad c_\mathcal{Y}(y, y') = \sqrt{(y - y')^\top \Sigma_2^{-1}(y - y')}\,.$$

To see this, let $T = \Sigma_2^{1/2}\Sigma_1^{-1/2}$, where $\Sigma_1^{1/2}$ and $\Sigma_2^{1/2}$ are the square roots of Σ_1 and Σ_2, respectively. Obviously, a linear map T satisfies $T \in \mathcal{T}(\mu, \nu)$ and $c_\mathcal{X}(x, x') = c_\mathcal{Y}(Tx, Tx')$ for all $x, x' \in \mathbb{R}^d$. According to Definition 4.2.3, a linear map T is a strong isomorphism. Proposition 4.9.2 implies $\mathrm{RGM}(\mu, \nu) = 0$. Notice that the same results hold for $c_\mathcal{X}(x, x') = x^\top \Sigma_1^{-1} x'$ and $c_\mathcal{Y}(y, y') = y^\top \Sigma_2^{-1} y'$ as well.

4.9.2 Analysis of GW = RGM in Non-atomic Cases

We have seen in Proposition 4.5.4 that $\mathrm{GW} \le \mathrm{GM} \le \mathrm{RGM}$ holds in general. This section proves Theorem 4.5.5, which states that these inequalities become equalities under mild conditions. Our techniques are inspired by Mémoli and Needham [2024].

Proof of Theorem 4.5.5. First, we invoke Theorem 16 in Chapter 15 of Royden [1988], which states that any Polish probability space that has no atoms is equivalent to $([0, 1], \lambda)$, where

λ is the Lebesgue measure on $[0,1]$. More precisely, $(\mathcal{X},\mu)$ is equivalent to $([0,1],\lambda)$ in the following sense: there exists a bijection $\phi\colon [0,1]\to\mathcal{X}$ such that ϕ,ϕ^{-1} are measurable, $\phi_{\#}\lambda=\mu$, and $(\phi^{-1})_{\#}\mu=\lambda$. Then, we can define the following network space $([0,1],\lambda,\tilde{c}_{\mathcal{X}})$, where

$$\tilde{c}_{\mathcal{X}}(u,v) := c_{\mathcal{X}}(\phi(u),\phi(v)) \quad \forall u,v\in[0,1]\ .$$

By construction, $([0,1],\lambda,\tilde{c}_{\mathcal{X}})$ is strongly isomorphic to $(\mathcal{X},\mu,c_{\mathcal{X}})$. Similarly, we can find a bijection $\psi\colon [0,1]\to\mathcal{Y}$ such that ψ,ψ^{-1} are measurable, $\psi_{\#}\lambda=\nu$, and $(\psi^{-1})_{\#}\nu=\lambda$. Then, we can also define a network space $([0,1],\lambda,\tilde{c}_{\mathcal{Y}})$ that is strongly isomorphic to $(\mathcal{Y},\nu,c_{\mathcal{Y}})$ by letting

$$\tilde{c}_{\mathcal{Y}}(u,v) := c_{\mathcal{Y}}(\psi(u),\psi(v)) \quad \forall u,v\in[0,1]\ .$$

Next, we show that the RGM distance between $(\mathcal{X},\mu,c_{\mathcal{X}})$ and $(\mathcal{Y},\nu,c_{\mathcal{Y}})$ is the same as the RGM distance between $([0,1],\lambda,\tilde{c}_{\mathcal{X}})$ and $([0,1],\lambda,\tilde{c}_{\mathcal{Y}})$, namely,

$$\mathrm{RGM}(\mu,\nu)=\mathrm{RGM}((\lambda,\tilde{c}_{\mathcal{X}}),(\lambda,\tilde{c}_{\mathcal{Y}}))\ , \tag{4.9.1}$$

where $(\lambda,\tilde{c}_{\mathcal{X}})$ and $(\lambda,\tilde{c}_{\mathcal{Y}})$ are placed to distinguish the two network spaces $([0,1],\lambda,\tilde{c}_{\mathcal{X}})$ and $([0,1],\lambda,\tilde{c}_{\mathcal{Y}})$ defined on the same space $([0,1],\lambda)$ with different cost functions. To see this, we use the triangle inequality of RGM (Proposition 4.9.1):

$$|\mathrm{RGM}(\mu,\nu)-\mathrm{RGM}((\lambda,\tilde{c}_{\mathcal{X}}),(\lambda,\tilde{c}_{\mathcal{Y}}))|\le \mathrm{RGM}(\mu,(\lambda,\tilde{c}_{\mathcal{X}}))+\mathrm{RGM}(\nu,(\lambda,\tilde{c}_{\mathcal{Y}}))\ .$$

By Proposition 4.9.2, we have $\mathrm{RGM}(\mu,(\lambda,\tilde{c}_{\mathcal{X}}))=\mathrm{RGM}(\nu,(\lambda,\tilde{c}_{\mathcal{Y}}))=0$, hence we have (4.9.1). Similarly, one can verify that $\mathrm{GW}(\mu,\nu)=\mathrm{GW}((\lambda,\tilde{c}_{\mathcal{X}}),(\lambda,\tilde{c}_{\mathcal{Y}}))$.

Now, to show $\mathrm{GW}(\mu,\nu)=\mathrm{RGM}(\mu,\nu)$, it suffices to prove

$$\mathrm{GW}((\lambda,\tilde{c}_{\mathcal{X}}),(\lambda,\tilde{c}_{\mathcal{Y}}))=\mathrm{RGM}((\lambda,\tilde{c}_{\mathcal{X}}),(\lambda,\tilde{c}_{\mathcal{Y}}))\ .$$

Equivalently, as in the proof of Proposition 4.5.4, we show

$$\inf_{\gamma\in\Pi(\lambda,\lambda)} \tilde{Q}(\gamma) = \inf_{\gamma\in\Pi'} \tilde{Q}(\gamma) \ ,$$

where $\Pi' = \{\gamma \in \Pi(\lambda,\lambda) : \gamma = (\mathrm{Id}, F)_{\#}\lambda = (B, \mathrm{Id})_{\#}\lambda, \ \ \exists (F, B) \in \mathcal{I}(\lambda,\lambda)\}$ and

$$\tilde{Q}(\gamma) = \int_{[0,1]\times[0,1]} \int_{[0,1]\times[0,1]} (\tilde{c}_{\mathcal{X}}(u,u') - \tilde{c}_{\mathcal{Y}}(v,v'))^2 \, \mathrm{d}\gamma\,(u,v)\,\mathrm{d}\gamma\,(u',v') \quad \forall \gamma \in \Pi(\lambda,\lambda) \ .$$

By Theorem 2.2 of Chowdhury and Mémoli [2019], there exists an optimal coupling $\gamma^* \in \Pi(\lambda,\lambda)$ such that

$$\mathrm{GW}((\lambda, \tilde{c}_{\mathcal{X}}), (\lambda, \tilde{c}_{\mathcal{Y}})) = \tilde{Q}(\gamma^*).$$

Next, we use the following fact from Theorem 1.1 of Brenier and Gangbo [2003]: there exists a sequence $(p_n)_{n\in\mathbb{N}}$ of bijective maps from $[0,1]$ to $[0,1]$ such that both p_n, p_n^{-1} are measure-preserving, namely, $(p_n)_{\#}\lambda = (p_n^{-1})_{\#}\lambda = \lambda$ for all $n \in \mathbb{N}$, and $\gamma_n := (\mathrm{Id}, p_n)_{\#}\lambda$ converges weakly to γ^*.[11] One can verify that

$$\gamma_n = (\mathrm{Id}, p_n)_{\#}\lambda = (\mathrm{Id}, p_n)_{\#}(p_n^{-1})_{\#}\lambda = (p_n^{-1}, \mathrm{Id})_{\#}\lambda \ ,$$

which shows that $\gamma_n \in \Pi'$. Accordingly, $(\gamma_n)_{n\in\mathbb{N}}$ is a sequence in Π' converging weakly to γ^*. As Lemma 2.3 of Chowdhury and Mémoli [2019] shows that $\tilde{Q}$ is continuous on $\Pi(\lambda,\lambda)$ with respect to the weak topology,

$$\mathrm{GW}((\lambda, \tilde{c}_{\mathcal{X}}), (\lambda, \tilde{c}_{\mathcal{Y}})) = \tilde{Q}(\gamma^*) = \lim_{n\to\infty} \tilde{Q}(\gamma_n) \geq \inf_{\gamma\in\Pi'} \tilde{Q}(\gamma) = \mathrm{RGM}((\lambda, \tilde{c}_{\mathcal{X}}), (\lambda, \tilde{c}_{\mathcal{Y}})) \ .$$

11. In fact, Theorem 1.1 of Brenier and Gangbo [2003] is stated for the case where λ is the Lebesgue measure restricted on $[0,1]^d$ for $d \geq 2$. However, one can verify that the proof still applies to $d = 1$.

As $\mathrm{GW}((\lambda, \tilde{c}_{\mathcal{X}}) \leq \mathrm{RGM}((\lambda, \tilde{c}_{\mathcal{X}}), (\lambda, \tilde{c}_{\mathcal{Y}}))$ by Proposition 4.5.4, we have

$$\mathrm{GW}((\lambda, \tilde{c}_{\mathcal{X}}) = \mathrm{RGM}((\lambda, \tilde{c}_{\mathcal{X}}), (\lambda, \tilde{c}_{\mathcal{Y}})).$$

Therefore, we have $\mathrm{GW}(\mu, \nu) = \mathrm{RGM}(\mu, \nu)$. □

4.10 Details of Section 4.5.2

This section provides details of the results presented in Section 4.5.2. Again, without loss of generality, we only consider $\lambda_1 = \lambda_2 = \lambda_3 = 1$ in (4.3.3). First, we clarify how measurable maps correspond to bounded linear operators between L^2 spaces.

Proposition 4.10.1. *Let $F\colon \mathcal{X} \to \mathcal{Y}$ be a measurable map such that $\| \mathrm{d}F_{\#}\pi_{\mathcal{X}} / \mathrm{d}\pi_{\mathcal{Y}} \|_{\infty} < \infty$. If we define*

$$\mathbf{F}(g) = g \circ F$$

for all $g \in L^2_{\mathcal{Y}}$, then $\mathbf{F}\colon L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$ is a bounded linear operator. Similarly, a measurable map $B\colon \mathcal{Y} \to \mathcal{X}$ satisfying $\| \mathrm{d}B_{\#}\pi_{\mathcal{Y}} / \mathrm{d}\pi_{\mathcal{X}} \|_{\infty} < \infty$ induces a bounded linear operator $\mathbf{B}\colon L^2_{\mathcal{X}} \to L^2_{\mathcal{Y}}$ such that $\mathbf{B}(g) = g \circ B$ for all $g \in L^2_{\mathcal{X}}$.

Proof. The linearity of $\mathbf{F}$ is obvious. Since

$$\int_{\mathcal{X}} g(F(x))^2 \,\mathrm{d}\pi_{\mathcal{X}} = \int_{\mathcal{Y}} g(y)^2 \,\mathrm{d}F_{\#}\pi_{\mathcal{X}}\,(y) = \int_{\mathcal{Y}} g(y)^2 \frac{\mathrm{d}F_{\#}\pi_{\mathcal{X}}}{\mathrm{d}\pi_{\mathcal{Y}}}(y) \,\mathrm{d}\pi_{\mathcal{Y}}\,(y) \leq \|g\|^2_{L^2(\pi_{\mathcal{Y}})} \left\| \frac{\mathrm{d}F_{\#}\pi_{\mathcal{X}}}{\mathrm{d}\pi_{\mathcal{Y}}} \right\|_{\infty},$$

we can see $\mathbf{F}(g) = g \circ F \in L^2_{\mathcal{X}}$ and thus $\mathbf{F}\colon L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$. From this inequality, the operator norm of $\mathbf{F}$ is bounded by $\| \mathrm{d}F_{\#}\pi_{\mathcal{X}} / \mathrm{d}\pi_{\mathcal{Y}} \|^{1/2}_{\infty}$; hence, boundedness of $\mathbf{F}$ follows. The same argument applies to $\mathbf{B}$. □

Next, we prove that (4.3.3) can be written in terms of $\mathbf{F}$ and $\mathbf{B}$ if $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$ are given

by the Mercer's representation:

$$K_{\mathcal{X}}(x, x') = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k(x') \ , \tag{4.10.1}$$

$$K_{\mathcal{Y}}(y, y') = \sum_{\ell=1}^{\infty} \gamma_\ell \psi_\ell(y) \psi_\ell(y') \ . \tag{4.10.2}$$

Let $\Phi_x = [\cdots, \phi_k(x), \cdots]^\top \in \mathbb{R}^\infty$ and $\Psi_y = [\cdots, \psi_\ell(y), \cdots]^\top \in \mathbb{R}^\infty$. Then, $K_{\mathcal{X}}(x, x') = \Phi_x^\top \Lambda \Phi_{x'}$ and $K_{\mathcal{Y}}(y, y') = \Psi_y^\top \Gamma \Psi_{y'}$. Also,

$$\begin{aligned} K_{\mathcal{X}}(x, B(y)) &= \sum_k \lambda_k \phi_k(x) [\phi_k \circ B](y) \\ &= \sum_k \lambda_k \phi_k(x) \mathbf{B}[\phi_k](y) \\ &= \sum_{k,\ell} \lambda_k \phi_k(x) \mathbf{B}_{\ell k} \psi_\ell(y) \\ &= \Psi_y^\top \mathbf{B} \Lambda \Phi_x \ . \end{aligned}$$

Analogously, we can obtain

$$\begin{aligned} &K_{\mathcal{Y}}(F(x), y) = \Phi_x^\top \mathbf{F} \Gamma \Psi_y \ , \\ &K_{\mathcal{X}}(B(y), B(y')) = \Psi_y^\top \mathbf{B} \Lambda \mathbf{B}^\top \Psi_{y'} \ , \\ &K_{\mathcal{Y}}(F(x), F(x')) = \Phi_x^\top \mathbf{F} \Gamma \mathbf{F}^\top \Phi_{x'} \ . \end{aligned}$$

Using this, we have

$$\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (K_{\mathcal{X}}(x_i, B(y_j)) - K_{\mathcal{Y}}(F(x_i), y_j))^2 = \frac{1}{mn} \sum_{i,j} (\Psi_{y_j}^\top \mathbf{B} \Lambda \Phi_{x_i} - \Phi_{x_i}^\top \mathbf{F} \Gamma \Psi_{y_j})^2 \ . \tag{4.10.3}$$

Also,

$$\begin{aligned}\mathrm{MMD}^2_{K_{\mathcal{X}}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) &= \frac{1}{m^2}\sum_{i,i'} K_{\mathcal{X}}(x_i, x_{i'}) + \frac{1}{n^2}\sum_{j,j'} K_{\mathcal{X}}(B(y_j), B(y_{j'})) - \frac{2}{mn}\sum_{i,j} K_{\mathcal{X}}(x_i, B(y_j)) \\ &= \frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \Lambda \Phi_{x_i'} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \mathbf{B}\Lambda\mathbf{B}^\top \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} \,.\end{aligned} \tag{4.10.4}$$

Similarly, we have

$$\mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n) = \frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \mathbf{F}\Gamma\mathbf{F}^\top \Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \Gamma \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j} \tag{4.10.5}$$

and

$$\begin{aligned}&\mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n) \\ &= \frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \Lambda \Phi_{x_i'} \Phi_{x_i}^\top \mathbf{F}\Gamma\mathbf{F}^\top \Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \Gamma \Psi_{y_{j'}} \Psi_{y_j}^\top \mathbf{B}\Lambda\mathbf{B}^\top \Psi_{y_{j'}} \\ &- \frac{2}{mn}\sum_{i,j} \Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j} \,.\end{aligned} \tag{4.10.6}$$

The following proposition summarizes the discussion so far.

Proposition 4.10.2. *Given Borel measures $\pi_{\mathcal{X}}$ and $\pi_{\mathcal{Y}}$ over $\mathcal{X}$ and $\mathcal{Y}$, respectively, suppose their corresponding L^2 spaces $L^2_{\mathcal{X}}$ and $L^2_{\mathcal{Y}}$ have countable orthonormal bases: $\{\phi_k\}_{k\in\mathbb{N}}$ and $\{\psi_\ell\}_{\ell\in\mathbb{N}}$. Also, assume $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$ are given by the Mercer's representation* (4.10.1) *and* (4.10.2)*. Let $\mathcal{F}_o$ and $\mathcal{B}_o$ be collections of all $F\colon \mathcal{X} \to \mathcal{Y}$ and $B\colon \mathcal{Y} \to \mathcal{X}$ such that $\| \mathrm{d}F_{\#}\pi_{\mathcal{X}} \,/\, \mathrm{d}\pi_{\mathcal{Y}} \|_\infty < \infty$ and $\| \mathrm{d}B_{\#}\pi_{\mathcal{Y}} \,/\, \mathrm{d}\pi_{\mathcal{X}} \|_\infty < \infty$, respectively. Then, solving* (4.3.3) *over $\mathcal{F}_o \times \mathcal{B}_o$ is equivalent to* (4.5.5)*, where $\mathcal{C}$ denotes the collection of all pairs of matrices $(\mathbf{F}, \mathbf{B})$ that correspond to a pair of bounded linear operators induced by $(F, B) \in \mathcal{F}_o \times \mathcal{B}_o$.*

Also, Ω is defined as

$$\begin{aligned}\Omega(\mathbf{F},\mathbf{B}) := \frac{1}{mn}\sum_{i,j}(\Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} - \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j})^2 \\ + \frac{1}{m^2}\sum_{i,i'}\Phi_{x_i}^\top\Lambda\Phi_{x_i'} + \frac{1}{n^2}\sum_{j,j'}\Psi_{y_j}^\top\mathbf{B}\Lambda\mathbf{B}^\top\Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j}\Psi_{y_j}^\top\mathbf{B}\Lambda\Phi_{x_i} \\ + \frac{1}{m^2}\sum_{i,i'}\Phi_{x_i}^\top\mathbf{F}\Gamma\mathbf{F}^\top\Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'}\Psi_{y_j}^\top\Gamma\Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j}\Phi_{x_i}^\top\mathbf{F}\Gamma\Psi_{y_j} \\ + \frac{1}{m^2}\sum_{i,i'}\Phi_{x_i}^\top\Lambda\Phi_{x_i'}\Phi_{x_i}^\top\mathbf{F}\Gamma\mathbf{F}^\top\Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'}\Psi_{y_j}^\top\Gamma\Psi_{y_{j'}}\Psi_{y_j}^\top\mathbf{B}\Lambda\mathbf{B}^\top\Psi_{y_{j'}} \\ - \frac{2}{mn}\sum_{i,j}\Psi_{y_j}^\top\mathbf{B}\Lambda\Phi_{x_i}\Phi_{x_i}^\top\mathbf{F}\Gamma\Psi_{y_j}\ .\end{aligned}$$

Based on this, we now prove Theorem 4.5.7.

Proof of Theorem 4.5.7. Let $\mathbf{x}_i = \Lambda^{1/2}\Phi_{x_i}$ and $\mathbf{y}_j = \Gamma^{1/2}\Psi_{y_j}$. Notice that we can view them as elements of a Hilbert space $\ell^2_{\mathbb{N}}$, that is, the space of square-summable sequences:

$$\ell^2_{\mathbb{N}} = \{(a_k)_{k\in\mathbb{N}} : \sum_k a_k^2 < \infty\}\ .$$

Also, define $\bar{\mathbf{F}} = \Lambda^{-1/2}\mathbf{F}\Gamma^{1/2}$ and $\bar{\mathbf{B}} = \Gamma^{-1/2}\mathbf{B}\Lambda^{1/2}$ where $\Lambda, \Gamma \succ 0$. By rewriting (4.10.3)-

(4.10.6) using $\mathbf{x}_i$, $\mathbf{y}_j$, $\bar{\mathbf{F}}$, and $\bar{\mathbf{B}}$, we have

$$
\begin{aligned}
&\Omega(\mathbf{F}, \mathbf{B}) \\
&= \frac{1}{mn} \sum_{i,j} (\mathbf{y}_j^\top \bar{\mathbf{B}} \mathbf{x}_i - \mathbf{x}_i^\top \bar{\mathbf{F}} \mathbf{y}_j)^2 && \text{(i)} \\
&+ \frac{1}{m^2} \sum_{i,i'} \mathbf{x}_i^\top \mathbf{x}_{i'} + \frac{1}{n^2} \sum_{j,j'} \mathbf{y}_j^\top \bar{\mathbf{B}} \bar{\mathbf{B}}^\top \mathbf{y}_{j'} - \frac{2}{mn} \sum_{i,j} \mathbf{y}_j^\top \bar{\mathbf{B}} \mathbf{x}_i && \text{(ii)} \\
&+ \frac{1}{m^2} \sum_{i,i'} \mathbf{x}_i^\top \bar{\mathbf{F}} \bar{\mathbf{F}}^\top \mathbf{x}_{i'} + \frac{1}{n^2} \sum_{j,j'} \mathbf{y}_j^\top \mathbf{y}_{j'} - \frac{2}{mn} \sum_{i,j} \mathbf{x}_i^\top \bar{\mathbf{F}} \mathbf{y}_j && \text{(iii)} \\
&+ \frac{1}{m^2} \sum_{i,i'} (\mathbf{x}_i^\top \mathbf{x}_{i'})(\mathbf{x}_i^\top \bar{\mathbf{F}} \bar{\mathbf{F}}^\top \mathbf{x}_{i'}) + \frac{1}{n^2} \sum_{j,j'} (\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}} \bar{\mathbf{B}}^\top \mathbf{y}_{j'}) - \frac{2}{mn} \sum_{i,j} (\mathbf{y}_j^\top \bar{\mathbf{B}} \mathbf{x}_i)(\mathbf{x}_i^\top \bar{\mathbf{F}} \mathbf{y}_j) \\
& && \text{(iv)} \\
&=: \bar{\Omega}(\bar{\mathbf{F}}, \bar{\mathbf{B}}) \, .
\end{aligned}
$$

As a result, (4.5.6) reduces to $\min_{\bar{\mathbf{F}}, \bar{\mathbf{B}} \in \mathbb{R}^{\infty \times \infty}} \bar{\Omega}(\bar{\mathbf{F}}, \bar{\mathbf{B}})$. Now, we define two finite-dimensional subspaces of $\ell^2_{\mathbb{N}}$ spanned by $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ and $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$, respectively:

$$
U_m := \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \, , \quad V_n := \operatorname{span}\{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \, .
$$

Also, we define P_{U_m} and P_{V_n} to be matrices that correspond to the orthogonal projection operators from $\ell^2_{\mathbb{N}}$ to U_m and to V_n, respectively. Recall that P_{U_m} and P_{V_n} are symmetric and idempotent by definition.

Our goal is to prove

$$
\bar{\Omega}(\bar{\mathbf{F}}, \bar{\mathbf{B}}) \geq \bar{\Omega}(P_{U_m} \bar{\mathbf{F}} P_{V_n}, P_{V_n} \bar{\mathbf{B}} P_{U_m}) \, .
$$

More precisely, we show that four terms (i)-(iv) decrease if we replace $\bar{\mathbf{F}}$ and $\bar{\mathbf{B}}$ with $P_{U_m} \bar{\mathbf{F}} P_{V_n}$ and $P_{V_n} \bar{\mathbf{B}} P_{U_m}$, respectively. First, observe that (i) remains the same. By definition, $P_{U_m} \mathbf{x}_i = \mathbf{x}_i$ and $P_{V_n} \mathbf{y}_j = \mathbf{y}_j$, thus $\mathbf{y}_j^\top \bar{\mathbf{B}} \mathbf{x}_i = \mathbf{y}_j^\top P_{V_n} \bar{\mathbf{B}} P_{U_m} \mathbf{x}_i$ and $\mathbf{x}_i^\top \bar{\mathbf{F}} \mathbf{y}_j = \mathbf{x}_i^\top P_{U_m} \bar{\mathbf{F}} P_{V_n} \mathbf{y}_j$.

Hence, (i) does not change.

To prove that (ii) decreases, it suffices to prove

$$\sum_{j,j'} \mathbf{y}_j^\top \bar{\mathbf{B}}\bar{\mathbf{B}}^\top \mathbf{y}_{j'} \geq \sum_{j,j'} \mathbf{y}_j^\top (P_{V_n}\bar{\mathbf{B}}P_{U_m})(P_{V_n}\bar{\mathbf{B}}P_{U_m})^\top \mathbf{y}_{j'} = \sum_{j,j'} \mathbf{y}_j^\top \bar{\mathbf{B}}P_{U_m}\bar{\mathbf{B}}^\top \mathbf{y}_{j'} \ .$$

To this end, define $P_{U_m}^\perp$ to be a matrix that corresponds to the orthogonal projection from $\ell^2_{\mathbb{N}}$ to $U_m^\perp$, the orthogonal complement of U_m. By definition, $P_{U_m} + P_{U_m}^\perp$ is the identity matrix and $P_{U_m}P_{U_m}^\perp = 0$. Hence,

$$\sum_{j,j'} \mathbf{y}_j^\top \bar{\mathbf{B}}\bar{\mathbf{B}}^\top \mathbf{y}_{j'} = \left\| \bar{\mathbf{B}}^\top \sum_j \mathbf{y}_j \right\|^2 = \left\| P_{U_m}\bar{\mathbf{B}}^\top \sum_j \mathbf{y}_j \right\|^2 + \left\| P_{U_m}^\perp \bar{\mathbf{B}}^\top \sum_j \mathbf{y}_j \right\|^2 \geq \sum_{j,j'} \mathbf{y}_j^\top \bar{\mathbf{B}}P_{U_m}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}$$

Here, the second equality follows by the Pythagorean theorem. Therefore, we can see (ii) decreases if we replace $\bar{\mathbf{B}}$ with $P_{V_n}\bar{\mathbf{B}}P_{U_m}$. Similarly, (iii) decreases.

For (iv), it suffices to prove

$$\begin{aligned}\sum_{j,j'} (\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}) &\geq \sum_{j,j'} (\mathbf{y}_j^\top \mathbf{y}_{j'}) \left(\mathbf{y}_j^\top (P_{V_n}\bar{\mathbf{B}}P_{U_m})(P_{V_n}\bar{\mathbf{B}}P_{U_m})^\top \mathbf{y}_{j'} \right) \\ &= \sum_{j,j'} (\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}P_{U_m}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}) \ .\end{aligned}$$

To see this,

$$\begin{aligned}\sum_{j,j'} (\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}) &= \sum_{j,j'} (\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}P_{U_m}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}) + \sum_{j,j'} (\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}P_{U_m}^\perp\bar{\mathbf{B}}^\top \mathbf{y}_{j'}) \\ &\geq \sum_{j,j'} (\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}P_{U_m}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}) \ ,\end{aligned}$$

where the inequality holds since

$$\sum_{j,j'}(\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}} P_{U_m}^\perp \bar{\mathbf{B}}^\top \mathbf{y}_{j'}) = \mathrm{Tr}\left[\left(\sum_j P_{U_m}^\perp \bar{\mathbf{B}}^\top \mathbf{y}_j \mathbf{y}_j^\top\right)\left(\sum_j P_{U_m}^\perp \bar{\mathbf{B}}^\top \mathbf{y}_j \mathbf{y}_j^\top\right)^\top\right] \geq 0 \,.$$

Similarly, we can obtain

$$\sum_{i,i'}(\mathbf{x}_i^\top \mathbf{x}_{i'})(\mathbf{x}_i^\top \bar{\mathbf{F}}\bar{\mathbf{F}}^\top \mathbf{x}_{i'}) \geq \sum_{i,i'}(\mathbf{x}_i^\top \mathbf{x}_{i'})(\mathbf{x}_i^\top \bar{\mathbf{F}} P_{V_n} \bar{\mathbf{F}}^\top \mathbf{x}_{i'}).$$

Hence, (iv) decreases.

Consequently, we have

$$(4.5.6) = \min_{\bar{\mathbf{F}},\bar{\mathbf{B}}\in\mathbb{R}^{\infty\times\infty}} \bar{\Omega}(\bar{\mathbf{F}},\bar{\mathbf{B}}) = \min_{\bar{\mathbf{F}},\bar{\mathbf{B}}\in\mathbb{R}^{\infty\times\infty}} \bar{\Omega}(P_{U_m}\bar{\mathbf{F}}P_{V_n}, P_{V_n}\bar{\mathbf{B}}P_{U_m}) \,.$$

By definition of a projection operator, we can find $\mathsf{U}_m \in \mathbb{R}^{m\times\infty}$ and $\mathsf{V}_n \in \mathbb{R}^{n\times\infty}$ such that

$$P_{U_m} = \Lambda^{1/2}\Phi_m \mathsf{U}_m \,, \quad P_{V_n} = \Gamma^{1/2}\Psi_n \mathsf{V}_n.$$

By letting $\mathsf{U}_m\bar{\mathbf{F}}\mathsf{V}_n^\top = \mathsf{F}_{m,n} \in \mathbb{R}^{m\times n}$ and $\mathsf{V}_n\bar{\mathbf{B}}\mathsf{U}_m^\top = \mathsf{B}_{n,m} \in \mathbb{R}^{n\times m}$, we have

$$\begin{aligned} P_{U_m}\bar{\mathbf{F}}P_{V_n} &= \Lambda^{1/2}\Phi_m \mathsf{F}_{m,n}\Psi_n^\top\Gamma^{1/2} \,, \\ P_{V_n}\bar{\mathbf{B}}P_{U_m} &= \Gamma^{1/2}\Psi_n \mathsf{B}_{n,m}\Phi_m^\top\Lambda^{1/2} \,. \end{aligned}$$

Hence,

$$\min_{\bar{\mathbf{F}},\bar{\mathbf{B}}\in\mathbb{R}^{\infty\times\infty}} \bar{\Omega}(P_{U_m}\bar{\mathbf{F}}P_{V_n}, P_{V_n}\bar{\mathbf{B}}P_{U_m}) = \min_{(\mathsf{F}_{m,n},\mathsf{B}_{n,m})\in\mathsf{C}} \omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) \,,$$

where

$$\omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) := \bar{\Omega}(\Lambda^{1/2}\Phi_m\mathsf{F}_{m,n}\Psi_n^\top\Gamma^{1/2}, \Gamma^{1/2}\Psi_n\mathsf{B}_{n,m}\Phi_m^\top\Lambda^{1/2}) \,.$$

Here, C is a constraint set implying that $\mathsf{F}_{m,n}$ and $\mathsf{B}_{n,m}$ are associated with $\bar{\mathbf{F}}$ and $\bar{\mathbf{B}}$, respectively, namely,

$$\mathsf{C} = \{(\mathsf{U}_m \bar{\mathbf{F}} \mathsf{V}_n^\top, \mathsf{V}_n \bar{\mathbf{B}} \mathsf{U}_m^\top) : \bar{\mathbf{F}}, \bar{\mathbf{B}} \in \mathbb{R}^{\infty \times \infty}\} \subset \mathbb{R}^{m \times n} \times \mathbb{R}^{n \times m} \ .$$

Therefore,

$$(4.5.6) = \min_{(\mathsf{F}_{m,n}, \mathsf{B}_{n,m}) \in \mathsf{C}} \omega(\mathsf{F}_{m,n}, \mathsf{B}_{n,m}) \geq \min_{\substack{\mathsf{F}_{m,n} \in \mathbb{R}^{m \times n} \\ \mathsf{B}_{n,m} \in \mathbb{R}^{n \times m}}} \omega(\mathsf{F}_{m,n}, \mathsf{B}_{n,m}) \ .$$

Finally, note that $\mathsf{C} = \mathbb{R}^{m \times n} \times \mathbb{R}^{n \times m}$ if U_m and V_n are full rank, that is, row spaces of U_m and V_n are rank-m and rank-n, respectively. This is true if kernel matrices

$$\mathbf{K}_{\mathcal{X}} = (\Lambda^{1/2} \Phi_m)^\top (\Lambda^{1/2} \Phi_m) \ , \ \mathbf{K}_{\mathcal{Y}} = (\Gamma^{1/2} \Psi_n)^\top (\Gamma^{1/2} \Psi_n)$$

are invertible. This is equivalent to say that they are positive definite. In this case,

$$\mathsf{U}_m = \mathbf{K}_{\mathcal{X}}^{-1} (\Lambda^{1/2} \Phi_m)^\top \ , \ \mathsf{V}_n = \mathbf{K}_{\mathcal{Y}}^{-1} (\Gamma^{1/2} \Psi_n)^\top \ ,$$

which are indeed full rank. Accordingly, we have

$$(4.5.6) = \min_{(\mathsf{F}_{m,n}, \mathsf{B}_{n,m}) \in \mathsf{C}} \omega(\mathsf{F}_{m,n}, \mathsf{B}_{n,m}) = \min_{\substack{\mathsf{F}_{m,n} \in \mathbb{R}^{m \times n} \\ \mathsf{B}_{n,m} \in \mathbb{R}^{n \times m}}} \omega(\mathsf{F}_{m,n}, \mathsf{B}_{n,m}) \ .$$

Finally, we prove ω is convex. To see this, verify

$$\begin{aligned}\omega(\mathsf{F}_{m,n}, \mathsf{B}_{n,m}) = \frac{1}{mn}\|\mathbf{K}_{\mathcal{Y}}\mathsf{B}_{n,m}\mathbf{K}_{\mathcal{X}} - \mathbf{K}_{\mathcal{Y}}\mathsf{F}_{m,n}^{\top}\mathbf{K}_{\mathcal{X}}\|^2 \\ + \left\|\mathbf{K}_{\mathcal{X}}^{1/2} \cdot \left(\frac{1}{m}\mathbf{1}_m - \mathsf{B}_{n,m}^{\top}\mathbf{K}_{\mathcal{Y}}\frac{1}{n}\mathbf{1}_n\right)\right\|^2 + \left\|\mathbf{K}_{\mathcal{Y}}^{1/2} \cdot \left(\frac{1}{n}\mathbf{1}_n - \mathsf{F}_{m,n}^{\top}\mathbf{K}_{\mathcal{X}}\frac{1}{m}\mathbf{1}_m\right)\right\|^2 \\ + \left\|\frac{1}{m}\mathbf{K}_{\mathcal{X}}^{3/2}\mathsf{F}_{m,n}\mathbf{K}_{\mathcal{Y}}^{1/2} - \frac{1}{n}\mathbf{K}_{\mathcal{X}}^{1/2}\mathsf{B}_{n,m}^{\top}\mathbf{K}_{\mathcal{Y}}^{3/2}\right\|^2 ,\end{aligned}$$

where $\mathbf{K}_{\mathcal{X}}^{1/2}$ and $\mathbf{K}_{\mathcal{Y}}^{1/2}$ are the square root matrices of $\mathbf{K}_{\mathcal{X}}$ and $\mathbf{K}_{\mathcal{Y}}$, respectively, and $\mathbf{1}_m \in \mathbb{R}^m$ and $\mathbf{1}_n \in \mathbb{R}^n$ are all-ones vectors. □

4.11 Details of the Experiments in Section 4.6

Here, we provide implementation details of the experiments in Section 4.6.

4.11.1 Gaussian

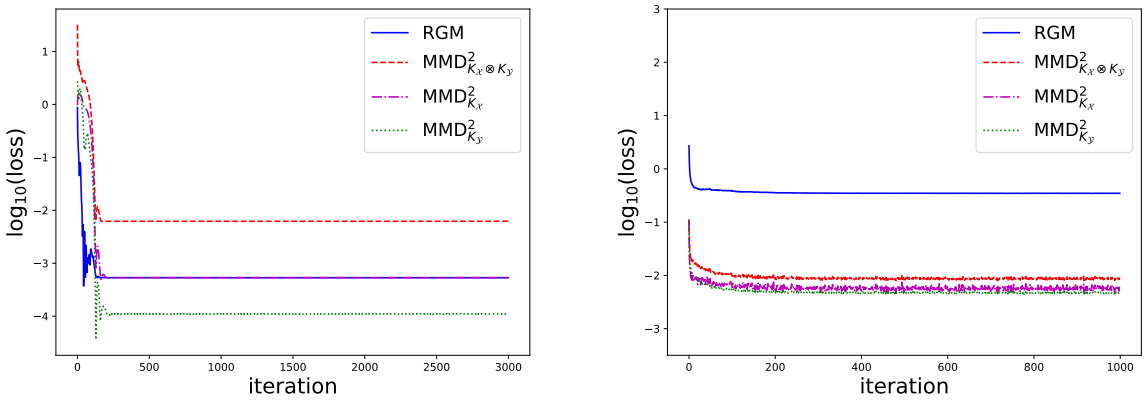
In the Gaussian experiment in Section 4.6, we minimize (4.3.3) using Adam [Kingma and Ba, 2014] for 3000 iterations. The learning rate at the initial iteration is 0.1 and we halve it after every 500 iterations. Figure 4.4(a) shows the training curve.

4.11.2 MNIST

$\mathbb{R}^2$ and MMDs For numerical stability, we encode a variant of empirical estimate (4.3.3) as our loss:

$$\lambda_1 \cdot \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(c_{\mathcal{X}}(x_i, B(y_j)) - c_{\mathcal{Y}}(F(x_i), y_j))^2 + \mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n)$$

$$+ \lambda_2 \cdot \mathrm{MMD}^2_{K_{\mathcal{X}}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) + \lambda_3 \cdot \mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n) \,. \tag{4.11.1}$$



(a) Gaussian (b) MNIST

Figure 4.4: Training curves for the experiments: (a) the Gaussian experiment, (b) the MNIST experiment with $\mathcal{X} = \mathbb{R}^2$.

We choose tuning parameters $(\lambda_1, \lambda_2, \lambda_3) = (0.01, 1, 1)$. For the Fully Connected Neural Network (FCNN) F and B, we apply the Rectified Linear Unit (ReLU) activation function $\sigma(x) = \max(x, 0)$ to all three hidden layers of F and B, and an additional tangent hyperbolic (tanh) function $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ to the output layer of F, both of which are elementwise activation functions. To put it explicitly, $y = F(x)$ is defined as

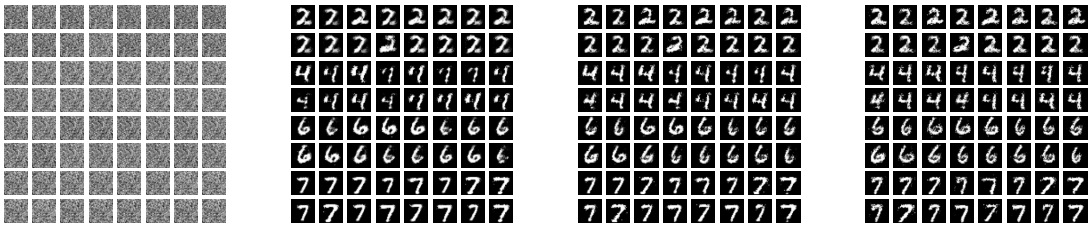
$$\begin{aligned} h_0 &= x, \quad x \in \mathbb{R}^2 \\ h_l &= \sigma(W_l h_{l-1} + b_l), \quad l = 1, 2 \\ y &= \tanh(W_3 h_2 + b_3) \end{aligned}$$

with $W_1 \in \mathbb{R}^{50\times 2}, W_2 \in \mathbb{R}^{50\times 50}, W_3 \in \mathbb{R}^{784\times 50}, b_1, b_2 \in \mathbb{R}^{50\times 1}, b_3 \in \mathbb{R}^{784\times 1}$. Similarly, $\widetilde{x} = B(\widetilde{y})$ is defined as

$$\begin{aligned} \widetilde{h}_0 &= \widetilde{y}, \quad \widetilde{y} \in \mathbb{R}^{784} \\ \widetilde{h}_l &= \sigma(\widetilde{W}_l \widetilde{h}_{l-1} + \widetilde{b}_l), \quad l = 1, 2 \\ \widetilde{x} &= \widetilde{W}_3 \widetilde{h}_2 + \widetilde{b}_3 \end{aligned}$$

with $\widetilde{W}_1 \in \mathbb{R}^{50\times 784}, \widetilde{W}_2 \in \mathbb{R}^{50\times 50}, \widetilde{W}_3 \in \mathbb{R}^{2\times 50}, \widetilde{b}_1, \widetilde{b}_2 \in \mathbb{R}^{50\times 1}, \widetilde{b}_3 \in \mathbb{R}^{2\times 1}$.

The training set is randomly divided into minibatches of size 256, for which we run Adam again for 1000 iterations. The learning rate at the initial iteration is 0.005 and we halve it after every 50 iterations. Figure 4.5 shows the generated images during the training process.



(a) Before training (b) + 20 iterations (c) + 50 iterations (d) + 1000 iterations

Figure 4.5: Generated images on MNIST data for digits $2, 4, 6, 7$ during the training process, under the experimental setup with $\mathbb{R}^2$ and MMD discrepancy.

Technical Details about Ablation Analysis As we mentioned, we construct Deep Convolutional Neural Networks (DCNNs) F and B by adapting the structures of the generator and discriminator in DCGAN [Radford et al., 2015], respectively. Concretely, F is a map from $\mathcal{X} = \mathbb{R}^d$ to $\mathcal{Y} = \mathbb{R}^{784}$, which consists of the following three parts.

1. Two fully connected layers that first map d inputs to 50 outputs, and then map 50 inputs to 128 outputs.

2. A reshaping step to transform the vector-type data to pixel-type data, i.e., a tensor with 128 channels, where each channel has only one-by-one pixel ($128 \times 1 \times 1$).

3. A series of four strided two-dimensional convolutional transpose layers with kernel size 4, and stride 1, 1, 2, 2. In four layers, we gradually decrease the number of channels to $64, 32, 16, 1$.

We pair the first three convolutional layers with two-dimensional batch normalization [Ioffe and Szegedy, 2015]. In terms of activation functions, we apply ReLU activation function to

all hidden layers (two fully connected layers and the first three convolutional layers), and an tanh function to the output layer.

On the other hand, B maps pixel-type data in $\mathcal{Y} = \mathbb{R}^{784}$ to $\mathcal{X} = \mathbb{R}^d$, which essentially reverses F in the following way.

1. A series of four two-dimensional convolutional layers with kernel size 4 and stride 2. In four layers, we gradually increase the number of channels to $16, 32, 64, 128$.

2. A reshaping step to transform the pixel-type data ($128 \times 1 \times 1$) to vector-type data.

3. One fully connected layer that maps 128 inputs to d outputs.

Again, we pair the last three convolutional layers with two-dimensional batch normalization. Also, we apply a leaky ReLU activation function $L(x) = \max(x, 0.2x)$ to all hidden layers (four convolutional layers); $L(x)$ is an elementwise activation function.

We implement the Sinkhorn divergence with squared Euclidean cost by using GeomLoss [Feydy et al., 2019]. Concretely, we first define the entropic regularized Kantorovich problem between $\widehat{\mu}_m$ and $B_{\#}\widehat{\nu}_n$ on some Euclidean space $\mathcal{X}$

$$W_{2,\varepsilon}^2(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) = \min_{\gamma \in \Pi(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n)} \sum_{i=1}^{m}\sum_{j=1}^{n} \gamma_{ij} \left(\|x_i - B(y_j)\|^2 + \varepsilon \log\big(\gamma_{ij}\big) \right)$$

where γ is a coupling matrix and γ_{ij} denotes its (i, j) element. Then the Sinkhorn divergence between two empirical measures is defined by $s_{\varepsilon,\mathcal{X}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) = W_{2,\varepsilon}^2(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) - \frac{1}{2}W_{2,\varepsilon}^2(\widehat{\mu}_m, \widehat{\mu}_m) - \frac{1}{2}W_{2,\varepsilon}^2(B_{\#}\widehat{\nu}_n, B_{\#}\widehat{\nu}_n)$. We encode our loss by replacing MMDs in (4.11.1) with Sinkhorn divergences

$$\lambda_1 \cdot \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(c_{\mathcal{X}}(x_i, B(y_j)) - c_{\mathcal{Y}}(F(x_i), y_j))^2 + s_{\varepsilon,\mathcal{X}\times\mathcal{Y}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n)$$
$$+ \lambda_2 \cdot s_{\varepsilon,\mathcal{X}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) + \lambda_3 \cdot s_{\varepsilon,\mathcal{Y}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n)\ ,$$

and choose tuning parameters $(\lambda_1, \lambda_2, \lambda_3) = (1, 1, 1)$. The Sinkhorn parameter ε is set to be 0.0001 for all three discrepancy measures. The rest of the setups, including the neural network architecture, the choice of the optimizer, the number of iterations, and batch size, are the same as the MMD case above.

Similarly, we implement the MMD-GAN by optimizing F under the MMD objective, and the rest of the setups are the same as the MMD case.

Additional Evaluation Scores To complement the results—Table 4.1 that shows the evaluation scores based on the difference between the first moments of $\widehat{F}_{\#}\mu$ and ν—in Section 4.6, we compute the evaluation scores based on the difference between higher order moments. Particularly, let us compare the second moment:

$$\|\widehat{\Sigma}_{\widehat{F}_{\#}\mu} - \widehat{\Sigma}_{\nu}\| \,,$$

where $\widehat{\Sigma}_{\widehat{F}_{\#}\mu}$ and $\widehat{\Sigma}_{\nu}$ are the sample covariance matrices estimated from $\{\widehat{F}(\tilde{x}_i)\}_{i=1}^{\tilde{m}}$ and $\{\tilde{y}_j\}_{j=1}^{\tilde{n}}$, respectively. For the same set of models as in Table 4.1, the corresponding new evaluation scores are presented in Table 4.2. We can observe results similar to those from Table 4.1. First, we can see that replacing the MMD terms with the Sinkhorn divergence leads to smaller (= better) evaluation scores (exception: DCNN with $d = 4$). Also, increasing the source dimension d shows mixed results. On the other hand, we can see that using DCNN leads to smaller (= better) evaluation scores.

Second, we follow the same evaluation strategy in Goodfellow et al. [2014] to fit a kernel density estimator to the generated samples $\{\widehat{F}(\tilde{x}_i)\}_{i=1}^{\tilde{m}}$ and report the mean log-likelihood of $\{\tilde{y}_j\}_{j=1}^{\tilde{n}}$ under this distribution. We refer readers to Section 5 of Goodfellow et al. [2014] for further details. The mean log likelihood scores are reported in Table 4.3, for which a larger score indicates a better performance. Essentially, we can observe the results consistent with those using the other evaluation scores. Again, we can see that replacing the MMD terms

d	2	4	10	20	50
FCNN					
Sinkhorn	13.90(0.03)	13.50(0.02)	13.68(0.03)	13.89(0.03)	15.09(0.03)
MMD	23.66(0.02)	21.49(0.03)	19.55(0.03)	22.59(0.03)	24.01(0.03)
MMDGAN	24.07(0.06)	15.40(0.02)	14.87(0.02)	15.32(0.03)	13.01(0.02)
DCNN					
Sinkhorn	13.33(0.02)	13.20(0.03)	13.62(0.03)	13.80(0.03)	14.66(0.03)
MMD	14.00(0.03)	12.65(0.04)	14.35(0.04)	16.77(0.04)	18.68(0.05)
MMD-GAN	14.58(0.06)	11.04(0.04)	10.70(0.04)	10.65(0.04)	10.62(0.03)
baseline	6.68				

Table 4.2: Evaluation scores (a smaller number implies better result) using the differences between the covariance matrices of the generated distribution $\widehat{F}_{\#}\mu$ and the target distribution ν; these are estimated by new samples $\{\tilde{x}_i\}_{i=1}^{2000}$ from $\mu = N(0, I_d)$ and $\{\tilde{y}_j\}_{j=1}^{2000}$ from the MNIST test set. We repeat the evaluation process for 100 times in each setting, and present the average evaluation score with the corresponding standard deviation in the parenthesis. The baseline is computed by randomly dividing the MNIST test set of 4000 samples into two subsets of 2000 samples and calculating the difference between the covariances matrices of the two subsets analogously to the evaluation score.

with the Sinkhorn divergence leads to larger (= better) evaluation scores. Also, increasing the source dimension d shows mixed results, while using DCNN leads to larger (= better) evaluation scores.

4.12 Inductive Bias of RGM and Brenier's Polar Factorization

As mentioned in Section 4.7, given two Polish probability spaces $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$, we prove that there is a pair of cost functions $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ such that the resulting network spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ are strongly isomorphic. More importantly, we discuss how these costs give rise to a specific strong isomorphism based on Brenier's polar factorization, providing a deeper insight into the inductive bias of the RGM sampler.

Preliminaries Throughout this section, we focus on Borel probability measures on $\mathbb{R}^d$ that are absolutely continuous with respect to the d-dimensional Lebesgue measure and have finite second moments; $\mathcal{P}_2^r(\mathbb{R}^d)$ denotes the collection of such measures. We can always

d	2	4	10	20	50
FCNN					
Sinkhorn	-199(4)	-150(4)	-160(4)	-154(3)	-151(3)
MMD	-384(7)	-315(6)	-294(6)	-365(6)	-365(6)
MMDGAN	-353(7)	-315(5)	-287(5)	-295(5)	-297(5)
DCNN					
Sinkhorn	-193(4)	-141(3)	-143(3)	-139(3)	-149(3)
MMD	-300(5)	-217(4)	-193(4)	-220(5)	-277(5)
MMD-GAN	-191(6)	-152(14)	-265(4)	-233(5)	-178(4)
baseline	-207				

Table 4.3: Mean log likelihood scores (a larger number implies better result) computed for the entire MNIST test set $\{\tilde{y}_j\}_{j=1}^{4000}$ with respect to the Gaussian Parzen window estimator fitted to $\{\widehat{F}(\tilde{x}_i)\}_{i=1}^{4000}$, where $\{\tilde{x}_i\}_{i=1}^{4000}$ are new samples from $\mu = N(0, I_d)$. The baseline is computed in a similar way, by replacing $\{\widehat{F}(\tilde{x}_i)\}_{i=1}^{4000}$ with a validation set $\{y_j\}_{j=1}^{4000}$ from the MNIST training set.

find a unique optimal transport map—given as the gradient of a convex function—between any two elements in $\mathcal{P}_2^r(\mathbb{R}^d)$; this is Brenier's theorem stated in Lemma 3.2.1, which we formally restate as follows.

Theorem 4.12.1 (Brenier's Theorem)**.** *Given $\lambda, \rho \in \mathcal{P}_2^r(\mathbb{R}^d)$, we can find a convex function $\psi\colon \mathbb{R}^d \to \mathbb{R}$ such that $\nabla\psi$ and $(\nabla\psi)^{-1}$ are unique optimal transport maps between them:*

$$\nabla\psi = \operatorname*{arg\,min}_{T_\#\lambda=\rho} \int_{\mathbb{R}^d} \|x - T(x)\|^2 \,\mathrm{d}\lambda(x) \quad \textit{and} \quad (\nabla\psi)^{-1} = \operatorname*{arg\,min}_{T_\#\rho=\lambda} \int_{\mathbb{R}^d} \|x - T(x)\|^2 \,\mathrm{d}\rho(x) \ .$$

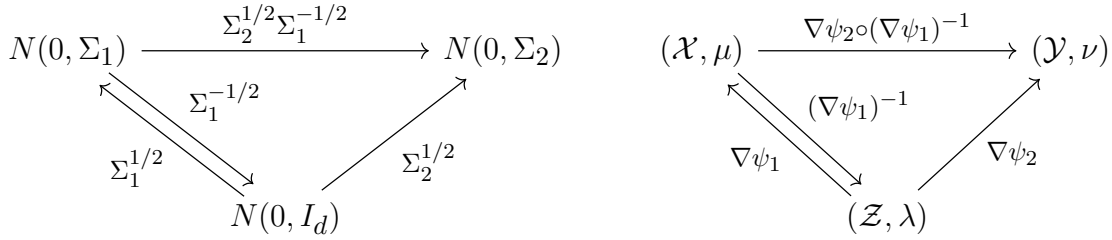
Remark 4.12.2. The proof of Brenier's theorem requires intricate convex analysis techniques. It is helpful to consider a simple case—where both λ and ρ are Gaussian—to better understand the main message of Brenier's theorem. Suppose $\lambda = N(0, I_d)$ and $\rho = N(0, \Sigma)$—assuming Σ is invertible—and focus on linear transport maps, that is, $T \in \mathbb{R}^{d\times d}$ such that $TT^\top = \Sigma$, for which the transport cost boils down to

$$\mathbb{E}_{x\sim N(0,I_d)}\|(I_d - T)x\|^2 = \mathrm{tr}((I_d - T)(I_d - T)^\top) = \mathrm{tr}(I_d) - 2\mathrm{tr}(T) + \mathrm{tr}(\Sigma) \ .$$

Therefore, the transport cost is minimized by a linear transport map T that maximizes its trace under the constraint $TT^\top = \Sigma$. Using linear algebra techniques, one can verify that $\Sigma^{1/2}$, the unique square root of Σ, is the optimal linear transport map;[12] we can indeed see that it is the gradient of a convex function $x \mapsto \frac{1}{2}x^\top \Sigma^{1/2} x$. Last but not least, notice that any admissible linear transport map $T \in \mathbb{R}^{d\times d}$, that satisfies $TT^\top = \Sigma$, can be decomposed as $T = \Sigma^{1/2} S$ for some $S \in O(d)$, where $O(d)$ is the orthogonal group in dimension d. The last fact is called the matrix factorization theorem, which will be elaborated in 4.12.2 to shed light on Brenier's polar factorization, as done here as an example of Brenier's theorem.

4.12.1 Designing Cost Functions for Strong Isomorphism

We have already seen in Example 4.9.5 that Mahalanobis distances make two Gaussian distributions strongly isomorphic. Though this constructive example seems to be a special case, it indicates a fundamental principle that carries over to general cases. We will elaborate on the general constructive principle in this section by referring to a common space Polish probability space $(\mathcal{Z}, \lambda)$ to design cost functions. Figure 4.6(a) visualizes Example 4.9.5 along with an additional base measure $\lambda := N(0, I_d)$; as we have remarked after Theorem 4.12.1, linear maps $\Sigma_1^{1/2}$ and $\Sigma_2^{1/2}$ are the optimal transport maps from λ to μ and ν, respectively.



(a) Example 4.9.5 (b) Any $\mu, \nu \in \mathcal{P}_2^r(\mathbb{R}^d)$

Figure 4.6: Optimal transport maps by Brenier's theorem

Letting $c_{\mathcal{Z}}$ be the Euclidean distance, notice that we may rewrite the Mahalanobis dis-

12. Though this argument is designed to show that $\Sigma^{1/2}$ is optimal among linear transport maps to get insights, it can be shown that $\Sigma^{1/2}$ is, in fact, optimal among all admissible transport maps.

tances in Example 4.9.5 as

$$c_{\mathcal{X}}(x, x') = c_{\mathcal{Z}}(\Sigma_1^{-1/2} x, \Sigma_1^{-1/2} x') \ ,$$
$$c_{\mathcal{Y}}(y, y') = c_{\mathcal{Z}}(\Sigma_2^{-1/2} y, \Sigma_2^{-1/2} y') \ .$$

This shows that $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are derived by properly combining the base distance $c_{\mathcal{Z}}$ with the optimal transport maps $\Sigma_1^{-1/2}$ and $\Sigma_2^{-1/2}$, respectively. Also, in this case, $\Sigma_2^{1/2}\Sigma_1^{-1/2}$, a composition of the two optimal transport maps, gives rise to a strong isomorphism under these cost functions.

The aforementioned procedure is indeed applicable to any $\mu, \nu \in \mathcal{P}_2^r(\mathbb{R}^d)$. As visualized in Figure 4.6(b), simply replace the linear maps $\Sigma_1^{1/2}$ and $\Sigma_2^{1/2}$ with the optimal transport maps $\nabla\psi_1$ and $\nabla\psi_2$ from any suitable base measure $\lambda \in \mathcal{P}_2^r(\mathbb{R}^d)$, respectively, by Brenier's theorem. Then, for any cost function $c_{\mathcal{Z}}$ on $\mathcal{Z}$, define

$$\begin{aligned} c_{\mathcal{X}}(x, x') &= c_{\mathcal{Z}}((\nabla\psi_1)^{-1}(x), (\nabla\psi_1)^{-1}(x')) \ , \\ c_{\mathcal{Y}}(y, y') &= c_{\mathcal{Z}}((\nabla\psi_2)^{-1}(y), (\nabla\psi_2)^{-1}(y')) \ . \end{aligned} \tag{4.12.1}$$

Then, the network spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$, where $\mathcal{X} = \text{supp}(\mu)$ and $\mathcal{Y} = \text{supp}(\nu)$, are strongly isomorphic; also, $\nabla\psi_2 \circ (\nabla\psi_1)^{-1}$ is a strong isomorphism. Notice that we have derived this fundamental result by simply rethinking Example 4.9.5 via Brenier's theorem and generalizing the diagram in Figure 4.6. We reiterate that the principle behind the isomorphic Gaussian example is fundamental and generalizable to generic measures in $\mathcal{P}_2^r(\mathbb{R}^d)$; it is certainly not just a toy example.

4.12.2 Identifying Strong Isomorphism

In the previous section, we have seen how to define suitable cost functions $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ that make $\mu, \nu \in \mathcal{P}_2^r(\mathbb{R}^d)$ strongly isomorphic. As pointed out in Section 4.7, the RGM sampler

brings in an inductive bias towards a strong isomorphism; indeed, we have shown that $\nabla\psi_2 \circ (\nabla\psi_1)^{-1}$ is a strong isomorphism under the cost functions in (4.12.1), leading to $\mathrm{RGM}(\mu, \nu) = 0$. In this section, we look at these results from a different angle using an insight from Brenier's polar factorization, highlighting unseen aspects of the inductive bias.

We start from Figure 4.6(b): fix $\mu, \nu \in \mathcal{P}_2^r(\mathbb{R}^d)$ and let $\mathcal{X} = \mathrm{supp}(\mu)$ and $\mathcal{Y} = \mathrm{supp}(\nu)$. Also, let λ be the Lebesgue measure on $\mathcal{Z} = [0, 1]^d$. Recall that the key ingredients of the diagram were optimal transport maps $\nabla\psi_1$ and $\nabla\psi_2$ from the base measure. It turns out that we may replace them with other transport maps—possibly not optimal—from the base measure. This observation comes from Brenier's polar factorization [Brenier, 1991], which we paraphrase as follows:

Theorem 4.12.3 (Brenier's Polar Factorization)**.** *For any transport map T_1 from λ to μ, we can find a unique map $s_1 \colon \mathcal{Z} \to \mathcal{Z}$ such that $(s_1)_{\#}\lambda = \lambda$ and $T_1 = \nabla\psi_1 \circ s_1$.*

In other words, any transport map from λ to μ is factorized into a composition of the unique optimal transport map $\nabla\psi_1$ and a (Lebesgue) measure-preserving map $s_1 \colon \mathcal{Z} \to \mathcal{Z}$. Let $S(\mathcal{Z})$ be the collection of all measure-preserving maps from $\mathcal{Z}$ to $\mathcal{Z}$, then Brenier's polar factorization essentially shows a one-to-one correspondence between $\mathcal{T}(\lambda, \mu)$ and $S(\mathcal{Z})$. Analogously, this implies a one-to-one correspondence between $\mathcal{T}(\lambda, \nu)$ and $S(\mathcal{Z})$.

Remark 4.12.4 (Matrix polar factorization)**.** To get insights into this sophisticated result, let us pause and go back to the remark below Theorem 4.12.1, where we have mentioned that in the Gaussian case, any linear transport map is decomposed as $T = \Sigma^{1/2}S$, the multiplication of the optimal transport map $\Sigma^{1/2}$ and an orthogonal matrix $S \in O(d)$,[13] which exactly correspond to $\nabla\psi_1$ and s_1 in Theorem 4.12.3, respectively. In other words, Brenier's Polar Factorization is a generalization of the matrix factorization that we have discussed earlier by restricting to the Gaussian case and linear transport maps.

13. Note that $S \in O(d)$ transports $N(0, I_d)$ to itself.

As mentioned earlier, we now replace the two arrows—the optimal transport maps $\nabla\psi_1$ and $\nabla\psi_2$ from the base measure—in Figure 4.6(b) with any transport maps, namely, $\nabla\psi_1 \circ s_1$ and $\nabla\psi_2 \circ s_2$, respectively, for any $s_1, s_2 \in S(\mathcal{Z})$. One technicality here is that we restrict our focus to bijective s_1 and s_2 to use invertibility to reverse the arrows.[14] Then, we can also define a transport map $\nabla\psi_2 \circ s_2 \circ (\nabla\psi_1 \circ s_1)^{-1}$ from $\mathcal{X}$ to $\mathcal{Y}$ by chaining the transport maps $\mu \to \lambda$ and $\lambda \to \nu$ as in Figure 4.6(b). These changes are shown in the new diagram Figure 4.7.

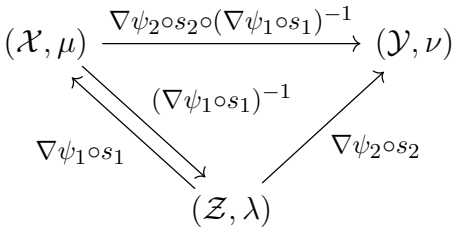


Figure 4.7: Generalization of Figure 4.6(b) via Brenier's polar factorization.

Now, let us go back to transform sampling. The arrow from μ to ν in Figure 4.7 in fact shows that there are many transport maps from μ to ν, that is,

$$\mathcal{F} = \{\nabla\psi_2 \circ s_2 \circ (\nabla\psi_1 \circ s_1)^{-1} : \text{bijective } s_1, s_2 \in S(\mathcal{Z})\} \subset \mathcal{T}(\mu, \nu) \ .$$

Though $\mathcal{F}$ is a collection of transport maps constructed in a certain way, that is, chaining $\mu \to \lambda$ and $\lambda \to \nu$, it still consists of infinitely many transport maps, reiterating that transform sampling is indeed over-identified. We show how the RGM sampler induces an inductive bias in this case, thereby choosing a strong isomorphism from the collection $\mathcal{F}$.

Consider the cost functions $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ defined by (4.12.1) with a cost $c_{\mathcal{Z}}$. As transport maps in $\mathcal{F}$ are invertible, by Lemma 4.9.3,

$$\mathcal{I}' := \{(F, F^{-1}) : F \in \mathcal{F}\} \subset \mathcal{I}(\mu, \nu).$$

14. Elements of $S(\mathcal{Z})$ may not be invertible in general. That said, a subset of $S(\mathcal{Z})$ consisting of bijective measure-preserving maps is dense in $S(\mathcal{Z})$ as discussed in Brenier and Gangbo [2003].

Recall that computing the RGM distance amounts to finding (F, B) such that $c_{\mathcal{X}}(x, B(y))$ best matches $c_{\mathcal{Y}}(F(x), y)$ on average. For $(F, B) \in \mathcal{I}'$, by (4.12.1), we have

$$c_{\mathcal{X}}(x, B(y)) = c_{\mathcal{Z}}((\nabla\psi_1)^{-1}(x), s_1 \circ s_2^{-1} \circ (\nabla\psi_2)^{-1}(y)) \ ,$$
$$c_{\mathcal{Y}}(F(x), y) = c_{\mathcal{Z}}(s_2 \circ s_1^{-1} \circ (\nabla\psi_1)^{-1}(x), (\nabla\psi_2)^{-1}(y)) \ .$$

Therefore, $c_{\mathcal{X}}(x, B(y)) = c_{\mathcal{Y}}(F(x), y)$ indeed holds provided $s_1 = s_2$, which amounts to a pair $(F, B) = (\nabla\psi_2 \circ (\nabla\psi_1)^{-1}, \nabla\psi_1 \circ (\nabla\psi_2)^{-1}) \in \mathcal{I}'$. In other words, among infinitely many possible elements in $\mathcal{F}$ characterized by s_1 and s_2, the RGM sampler with cost functions in (4.12.1) favors $\nabla\psi_2 \circ (\nabla\psi_1)^{-1}$, the strong isomorphism under these cost functions as visualized in Figure 4.6(b). In summary,

- we can characterize a collection $\mathcal{F}$ of transport maps by pairs of bijective measure-preserving maps (s_1, s_2) via Brenier's polar factorization,
- the RGM sampler with cost functions in (4.12.1) finds the one in $\mathcal{F}$ satisfying the rearrangement correspondence $s_1 = s_2$, which is exactly the strong isomorphism (inductive bias).

CHAPTER 5

A CONVEXIFIED MATCHING APPROACH TO IMPUTATION AND INDIVIDUALIZED INFERENCE

5.1 Introduction

One central topic in applied econometric research is to assess the effects of policy interventions reliably. On the one hand, to analyze nonexperimental data, advanced econometric estimates are devised to provide granular counterfactual answers by leveraging specific structural models. In an influential paper in 1986 [LaLonde, 1986], LaLonde questioned whether such sophisticated econometric estimates are credible. He compared them with the experimental benchmarks on some coarse summary—for example, the average treatment effect (ATE)—and concluded unfavorably. Subsequent works [Dehejia and Wahba, 1999, 2002] recovered some experimental benchmarks with improved methods, leaving the debate open up to this day [Imbens and Xu, 2024]. On the other hand, when experimental data, as in randomized controlled trials, are available, elementary statistical estimates can determine the ATE and eliminate confounding explanations. The ATE estimate, however, does not answer whether the treatment works for an individual. In modern applications such as personalized medicine and online marketing, the treatment effects vary across individuals; the treatment might be beneficial for some individuals but ineffective for others [Liu and Meng, 2016]. One potentially costly approach is to conduct individualized experiments in different time windows, known as N-of-1 trials [Hill, 1955, Liang and Recht, 2023]. Therefore, it is desirable to develop nonexperimental methods that conform to the coarse estimates—for example, those for ATE or the average treatment effect on the treated (ATT)—at the aggregate level, while delivering more granular, individualized inference.

Individualized inference is about imputing missing counterfactual outcomes—the outcomes that would have been observed had the subjects received the opposite treatment—and

their uncertainty quantification. Three popular approaches are available in the literature: matching, regression imputation, and synthetic control. This chapter proposes a new convexified matching method by integrating these three approaches. The central quantity we play is a coupling matrix between the treated and control data clouds, which resembles matching and is used to synthesize the counterfactual outcomes as in classic nonparametric regression.

Matching [Rubin, 2006, Rosenbaum, 2020] is a widely adopted way to impute counterfactual outcomes. For each subject, the counterfactual outcome is estimated by identifying units in the opposite treatment group similar to the subject in covariates and then averaging their outcomes. Due to its simplicity, nearest neighbor matching, which finds subjects closest under a suitable distance between the covariates, is favored in practice. Exact or approximate matching with high-dimensional covariates can be difficult, and thus propensity score matching was proposed as a remedy [Rosenbaum and Rubin, 1983]. There are more sophisticated matching methods to improve balance or forbid certain matches, often referred to as the optimal matching [Rosenbaum, 2020, Zubizarreta et al., 2023]; they require solving combinatorial optimization problems by network optimization techniques or mixed integer programming, which can be computationally expensive. However, even analyzing the aggregate ATE estimates resulting from combinatorial optimization problems is highly nontrivial, and these estimates may differ from simple ATE estimates based on propensity score weighting [Rosenbaum, 1987, Hirano and Imbens, 2001, Hirano et al., 2003]. As in matching, we solve for a coupling matrix—a convex relaxation to the combinatorial constraints—without access to outcomes, separating the design and analysis phases in an observational study. For a large-scale matching problem with N units, we devise entropic-regularized algorithms from optimal transport [Villani, 2003, Peyré and Cuturi, 2019] to solve convexified matching, calling a matrix scaling subroutine with a number of times nearly independent of N; see Section 5.4. In contrast, in a conventional matching problem, solvability within polynomial time of N may not be feasible.

Regression imputation is another approach to estimating counterfactual outcomes. The idea is simple: under the ignorability or unconfoundedness assumption [Rosenbaum and Rubin, 1983], the underlying functions that map the covariates to the responses under the treatment and the control are identifiable and estimable based on observational data. It translates a causal inference problem into a regression problem. Parametric or nonparametric regression techniques estimate the conditional response function under the treatment and the control, given the covariates, and in turn find the conditional average treatment effect (CATE), see Hahn [1998], Heckman et al. [1997], Heckman and Vytlacil [2005], MaCurdy et al. [2011]. At the aggregate level, semiparametric efficient estimation of ATE leveraging regression imputation techniques is studied in Hahn [1998]. Admittedly, if the quantity of interest is at the aggregate-level—a one-dimensional parameter such as ATE or ATT—the specific nonparametric regression technique matters less, as long as it estimates the function reasonably well. There has been a fruitful line of literature on the use of flexible machine learning methods in estimating CATE [Belloni et al., 2014, Athey and Imbens, 2016, Chernozhukov et al., 2017, Farrell, 2015, Farrell et al., 2021, Wager and Athey, 2018], in place of classic Nadaraya-Watson (NW) kernel nonparametric regression. Albeit naive, let us use the NW estimator as an example to illustrate a shared feature by regression imputation and matching: to compute a counterfactual outcome for a treated unit, NW finds local neighbors in the control group and uses convex weights based on the nonparametric kernel to synthesize the outcome. We will adopt this aggregation by convex weights feature into our convexified matching, whereas our coupling weights are optimized globally. Our coupling weights are determined to optimize imputation quality and individual uncertainty quantification; see Section 5.3.

Hybrid methods were proposed to combine matching or propensity score weighting [Rosenbaum, 1987, Hirano and Imbens, 2001, Hirano et al., 2003, Li et al., 2018] with regression imputation techniques to efficiently estimate ATE, notably the doubly robust estimators

[Van der Laan and Robins, 2003, Cattaneo, 2010, Farrell, 2015, Farrell et al., 2021, Chernozhukov et al., 2022]. The estimator compares favorably as it is valid when either the propensity score function or the regression function is consistent, robust to bias due to misspecification. Similar in spirit, bias correction using regression to improve the nearest neighbor matching was studied in Abadie and Imbens [2011, 2006].

From a different vein, synthetic control [Abadie and Gardeazabal, 2003, Abadie et al., 2010, Abadie, 2021] provides a fresh look at counterfactual imputation: a convex combination of control units may synthesize a subject under treatment better than any one of the control unit. Mathematically, a convex combination of data may simultaneously reduce the approximation error or bias, and, at the same time, reduce the variance driven by the idiosyncratic errors. This convex relaxation idea also eases the computation for the combinatorial optimization problem in matching. We adopt this convex combination idea to approximate the overall covariate information under treatment by convex combinations of the covariate information under control, defined based on a data set-to-data set coupling; see Section 5.2.

We develop a convexified matching method that solves an optimal coupling, which in turn defines convex combinations for missing value imputation. Unlike combinatorial optimization problems, where the optimization variable must be an extreme point of a certain polytope to represent an assignment, we optimize over the whole polytope as the variable represents the weights of the convex combination. Moreover, we add an entropic regularization to the optimization, where the strength of regularization controls the bias and variance tradeoff. The resulting formulation is a smooth convex optimization problem confined to the polytope that can be solved efficiently. Lastly, by properly specifying the constraints, we can pair with propensity score weighting [Rosenbaum, 1987, Hirano and Imbens, 2001, Hirano et al., 2003] estimators. We can guarantee that the aggregate summary of individual treatment effects coincides with the desired ATE or ATT estimates.

We introduce an inference procedure to construct individual confidence intervals for the estimated counterfactual outcomes based on the convexified matching method. Providing a credible confidence interval around the individual treatment effect can help decide whether to adopt the treatment for each individual. The width of the confidence interval is determined by the approximation error in covariate balancing and the entropy of the convex weights, which comprise the objective function of the optimization. Therefore, how we formulate the convexified matching is directly targeted at individualized inference. Entropic regularization plays a crucial role in inference, namely, controlling the width of the individual confidence intervals, and in computation, namely, designing fast iterative algorithms to solve the optimization.

In summary, we integrate important features from matching, regression imputation, and synthetic control. We impute counterfactual outcomes by convex combinations defined based on an optimal coupling. The coupling is a convex relaxation to optimal matching and can be solved efficiently using iterative matrix scaling subroutines called the Sinkhorn algorithm [Sinkhorn, 1967, Cuturi, 2013]. We estimate granular individual treatment effects while maintaining a desirable aggregate-level summary by properly constraining the coupling. We construct transparent, individual confidence intervals for the estimated counterfactual outcomes, where the optimization objective controls the width of the confidence intervals.

5.2 Imputation Method: Synthetic Coupling

We first introduce Kernel Synthetic Coupling (KSC), a convexified matching method for missing value imputation. Later in Section 5.3, we build individual confidence intervals around the imputed values. To streamline the exposition, we cast the KSC method following the notations in the potential outcome framework [Neyman, 1923, Rubin, 1974] for a binary treatment—a leading example for missing value imputation.

We consider N units indexed by $1, \ldots, N$, where each unit is associated with two potential

outcome variables $Y_i(1), Y_i(0) \in \mathbb{R}$ under treatment and control, respectively, and a covariate vector $x_i \in \mathbb{R}^d$. We can observe only one of $Y_i(1), Y_i(0)$ depending on the treatment assignment $Z_i = 1$ or $Z_i = 0$, denoting that unit i is treated or not. Accordingly, the observed outcome Y_i of unit i is given as $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$. We let $\mathcal{T} = \{i \in [N] : Z_i = 1\}$ and $\mathcal{C} = \{i \in [N] : Z_i = 0\}$ denote the sets of indices of the treated units and control units, respectively, so that $\mathcal{T} \cup \mathcal{C} = \{1, \ldots, N\} =: [N]$. Also, let $N_t := |\mathcal{T}|$ and $N_c := |\mathcal{C}|$ denote the numbers of the treated and control units. The individual treatment effect of unit i is defined as $\tau_i = Y_i(1) - Y_i(0)$, which is unobservable as only one of $Y_i(1), Y_i(0)$ can be observed.

5.2.1 A Simple Formulation

Without loss of generality, we focus on imputing the missing potential outcomes of the treated units; counterfactual outcomes of the control units can be similarly obtained. We estimate $Y_j(0)$ by some $\widehat{Y}_j(0)$ for each treated unit j. This, in turn, allows us to estimate the individual treatment effect τ_j by $Y_j - \widehat{Y}_j(0) = Y_j(1) - \widehat{Y}_j(0)$. The proposed convexified matching combines two ideas. First, we want to find a match between the data cloud in the treated group and that in the control using the covariate information. Second, as in the synthetic control method, the matching quality is determined by how well it approximates the covariate of each treated unit by a convex combination of the covariates of control units. As a result, this procedure involves finding a matrix, which we call a *coupling*, indexed by a pair (i, j) for $i \in \mathcal{C}$ and $j \in \mathcal{T}$ denoting the weight assigned to the covariate of control unit i to approximate the covariate of treated unit j. We solve the following optimization

to obtain the *optimal coupling*:

$$\min_{\pi \in \mathbb{R}_+^{N_c \times N_t}} \quad \frac{1}{2N_t} \sum_{j \in \mathcal{T}} \|x_j - \sum_{i \in \mathcal{C}} \frac{\pi_{ij}}{1/N_t} x_i\|^2 + \lambda \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{T}} \pi_{ij} \log \frac{\pi_{ij}}{e}, \tag{5.2.1}$$

$$\text{subject to} \quad \sum_{i \in \mathcal{C}} \pi_{ij} = \frac{1}{N_t} \quad \forall j \in \mathcal{T}, \tag{5.2.2}$$

$$\sum_{j \in \mathcal{T}} \pi_{ij} = \frac{1}{N_c} \quad \forall i \in \mathcal{C}. \tag{5.2.3}$$

The first term in the objective function (5.2.1) is the average of the squared approximation errors for covariate balancing. In the first constraint (5.2.2), for each treated unit j, the synthetic weights $(\frac{\pi_{ij}}{1/N_t})_{i \in \mathcal{C}}$ sum to 1 and thus the term $\sum_{i \in \mathcal{C}} \frac{\pi_{ij}}{1/N_t} x_i$ in (5.2.1) is a convex combination to approximate x_j.

The second term in (5.2.1) is an entropic regularization with regularization parameter $\lambda > 0$. Increasing the parameter λ encourages the solution π to be uniform, namely, $\pi_{ij} = \frac{1}{N_c \cdot N_t}$ for all (i, j)'s, which essentially corresponds to increasing the number of neighbors in the nearest neighbor matching.

The second constraint (5.2.3) enforces that all control units contribute equally and thus seeks a matching between treated and control data sets; this constraint can be generalized to incorporate inverse propensity scores, to be shown in the next section. If this constraint is dropped, the optimization can be decoupled into N_t independent optimization of the same type, each with one treated unit $j \in \mathcal{T}$, also solvable by our optimization algorithm detailed in Section 5.4. We shall show in a second that this constraint will be crucial. It enables the convex program to synthesize granular information while enforcing, at the coarse level, that the answers agree with typical estimators for the ATE or ATT.

Finally, the optimization variable $\pi \in \mathbb{R}_+^{N_c \times N_t}$ is a matrix whose entries are indexed by (i, j) for $i \in \mathcal{C}$ and $j \in \mathcal{T}$, instead of the usual indexing by $i \in \{1, \ldots, N_c\}$ and $j \in \{1, \ldots, N_t\}$, to denote that π_{ij} is the weight between control unit i and treated unit j. We

call $\pi \in \mathbb{R}_+^{N_c \times N_t}$ satisfying the constraints (5.2.2) and (5.2.3) a *coupling*, which generalizes doubly stochastic matrices to non-square matrices with prescribed row and column sums.

Upon obtaining the solution, denoted by $\widehat{\pi}$, we impute the counterfactual outcome $Y_j(0)$ of treated unit j by the convex combination of the outcomes of control units with the weights $(N_t \widehat{\pi}_{ij})_{i \in \mathcal{C}}$:

$$\widehat{Y}_j(0) := \sum_{i \in \mathcal{C}} \frac{\widehat{\pi}_{ij}}{1/N_t} Y_i, \tag{5.2.4}$$

which in turn leads to the individual treatment effect estimate $\widehat{\tau}_j := Y_j - \widehat{Y}_j(0)$. By the constraint (5.2.3), the average of the estimated individual treatment effects coincides with the difference in the mean outcomes between the treatment and control groups denoted as $\widehat{\tau}^{\mathrm{DiM}}$:

$$\frac{1}{N_t} \sum_{j \in \mathcal{T}} \widehat{\tau}_j = \frac{1}{N_t} \sum_{j \in \mathcal{T}} (Y_j - \sum_{i \in \mathcal{C}} \frac{\widehat{\pi}_{ij}}{1/N_t} Y_i) = \frac{1}{N_t} \sum_{j \in \mathcal{T}} Y_j - \frac{1}{N_c} \sum_{i \in \mathcal{C}} Y_i =: \widehat{\tau}^{\mathrm{DiM}}. \tag{5.2.5}$$

Therefore, the proposed convexified matching allows for estimating individual treatment effects while maintaining the desired ATT estimate $\widehat{\tau}^{\mathrm{DiM}}$.

The resulting optimization is a smooth convex optimization problem. To see this, notice that the first term of the objective function (5.2.1) is the following quadratic function of π:

$$\frac{N_t}{2} \langle \pi, K_{cc} \pi \rangle - \langle \pi, K_{ct} \rangle + \frac{\mathrm{tr}(K_{tt})}{2N_t}, \tag{5.2.6}$$

where $K_{cc} = (\langle x_i, x_{i'} \rangle)_{i,i' \in \mathcal{C}} \in \mathbb{R}^{N_c \times N_c}$, $K_{ct} = (\langle x_i, x_j \rangle)_{(i,j) \in \mathcal{C} \times \mathcal{T}} \in \mathbb{R}^{N_c \times N_t}$, $K_{tt} = (\langle x_j, x_{j'} \rangle)_{j,j' \in \mathcal{T}} \in \mathbb{R}^{N_t \times N_t}$ are the Gram matrices whose entries are inner products of the covariates. This quadratic function is convex as the Gram matrix K_{cc} is positive semidefinite. The second term—the entropy term—is convex. The constraint set resulting from (5.2.2) and (5.2.3) is a convex polytope consisting of couplings. In Section 5.4, we introduce and analyze efficient algorithms to solve this optimization problem based on a simple iterative

matrix scaling procedure called the Sinkhorn algorithm [Sinkhorn, 1967].

Lastly, we discuss several existing methods and concepts related to the proposed method. Abadie and L'hour [2021] considers multiple treated units ($N_t > 1$) and proposes running synthetic controls separately for treated units with a different regularization function. The proposed method differs from running synthetic controls separately because of the constraint (5.2.3). Another way of applying synthetic control to multiple treated units is to average these units and find the weights for the average [Kreif et al., 2016, Robbins et al., 2017, Ben-Michael et al., 2022]. If there is only one treated unit ($N_t = 1$) and no constraint (5.2.3), the proposed method is equivalent to the standard synthetic control with entropic regularization as in Hainmueller [2012] which proposes entropic regularization for covariate balancing. Also, notice that the first term of (5.2.1), under the constraint (5.2.2), is upper bounded as follows:

$$\frac{1}{2N_t}\sum_{j\in\mathcal{T}}\|x_j - \sum_{i\in\mathcal{C}}\frac{\pi_{ij}}{1/N_t}x_i\|^2 = \frac{1}{2N_t}\sum_{j\in\mathcal{T}}\|\sum_{i\in\mathcal{C}}\frac{\pi_{ij}}{1/N_t}(x_j - x_i)\|^2 \leq \frac{1}{2N_t}\sum_{j\in\mathcal{T}}\sum_{i\in\mathcal{C}}\frac{\pi_{ij}}{1/N_t}\|x_j - x_i\|^2,$$

where the inequality uses Jensen's inequality. This upper bound analytically shows that the average approximation error based on convex combinations is smaller than that based on pairwise distances, which aligns with the motivation mentioned in Section 5.1. Replacing the first term of the objective function (5.2.1) by the above upper bound—while maintaining the constraints (5.2.2), (5.2.3)—leads to the optimal transport problem [Villani, 2003] with entropic regularization [Cuturi, 2013].

5.2.2 Extensions

We present two extensions of the simple formulation introduced in Section 5.2.1. The first is extending the main insights to cover the nonlinear case when the potential outcomes can be nonlinear functions of covariates, using a kernel trick [Shawe-Taylor and Cristianini, 2004,

Steinwart and Christmann, 2008]. The second is incorporating notions of propensity scores [Rosenbaum and Rubin, 1983] into our synthetic coupling formulation.

Kernel Synthetic Coupling We can kernelize the simple formulation to extend to nonlinear cases. The kernel trick is widely used in machine learning to modify methods that originated in the linear setting to accommodate nonlinearity, such as ridge regression, support vector machines, and principal component analysis. Recall that (5.2.1) relies on the Gram matrices as shown in (5.2.6). Therefore, to kernelize the simple formulation, we can replace the Gram matrices with the kernel matrices, namely, letting $K_{cc} = (k(x_i, x_{i'}))_{i,i' \in \mathcal{C}}$, $K_{ct} = (k(x_i, x_j))_{(i,j) \in \mathcal{C} \times \mathcal{T}}$, and $K_{tt} = (k(x_j, x_{j'}))_{j,j' \in \mathcal{T}}$ for some kernel function $k \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. The kernelized formulation is essentially the same convex optimization problem. It can be tackled by the algorithms we introduce in Section 5.4.

Applying the above kernel trick is equivalent to replacing the first term of (5.2.1) with the approximation error in the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with the kernel k, namely, $\frac{1}{2N_t} \sum_{j \in \mathcal{T}} \|\phi_{x_j} - \sum_{i \in \mathcal{C}} \frac{\pi_{ij}}{1/N_t} \phi_{x_i}\|_{\mathcal{H}}^2$, where $\|\cdot\|_{\mathcal{H}}$ is the norm in $\mathcal{H}$ and $\phi_x = k(x, \cdot) \in \mathcal{H}$ denotes the canonical feature of $x \in \mathbb{R}^d$. In other words, we generalize the matching of covariates in the Euclidean space to the matching of embedded canonical features in $\mathcal{H}$, which leads to the following kernelized formulation:

$$\begin{aligned} \min_{\pi \in \mathbb{R}_+^{N_c \times N_t}} \quad & \frac{1}{2} \sum_{j \in \mathcal{T}} v_j \|\phi_{x_j} - \sum_{i \in \mathcal{C}} \frac{\pi_{ij}}{v_j} \phi_{x_i}\|_{\mathcal{H}}^2 + \lambda \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{T}} \pi_{ij} \log \frac{\pi_{ij}}{e}, \\ \text{subject to} \quad & \sum_{i \in \mathcal{C}} \pi_{ij} = v_j \quad \forall j \in \mathcal{T} \quad \text{and} \quad \sum_{j \in \mathcal{T}} \pi_{ij} = w_i \quad \forall i \in \mathcal{C}, \end{aligned} \tag{5.2.7}$$

where $v = (v_j)_{j \in \mathcal{T}} \in \Delta_{N_t}^+$ and $w = (w_i)_{i \in \mathcal{C}} \in \Delta_{N_c}^+$ are strictly positive probability vectors that generalize the constraints (5.2.2) and (5.2.3) to allow treated and control units to have different weights. With the optimal coupling $\widehat{\pi}$, we impute using convex combination same as in (5.2.4).

Propensity Score Using arbitrary weights $v = (v_j)_{j\in\mathcal{T}} \in \Delta^+_{N_t}$ and $w = (w_i)_{i\in\mathcal{C}} \in \Delta^+_{N_c}$ allows for various average treatment effect estimators, particularly including those incorporating propensity scores. To estimate the ATT, we can set $v = \frac{1}{N_t} 1_{N_t}$ as before, and the average of the estimated individual treatment effects of the treated leads to $\frac{1}{N_t}\sum_{j\in\mathcal{T}} \widehat{\tau}_j = \frac{1}{N_t}\sum_{j\in\mathcal{T}} Y_j - \sum_{i\in\mathcal{C}} w_i Y_i$ by the constraints.

Now, let $\widehat{p}\colon \mathbb{R}^d \to (0,1)$ be a suitable estimator of the propensity score, $x \mapsto \mathbb{P}(Z = 1 \,|\, X = x)$. If the weights w are chosen based on the estimated propensity scores as

$$w_i = \frac{\widehat{p}(x_i)}{1-\widehat{p}(x_i)} \Big/ \sum_{i'\in\mathcal{C}} \frac{\widehat{p}(x_{i'})}{1-\widehat{p}(x_{i'})} \quad \forall i \in \mathcal{C}, \tag{5.2.8}$$

the aggregate effects match the normalized Inverse Probability Weighting (IPW) estimator [Hirano and Imbens, 2001]:

$$\frac{1}{N_t}\sum_{j\in\mathcal{T}} \widehat{\tau}_j = \frac{\sum_{j\in\mathcal{T}} Y_j}{N_t} - \frac{\sum_{i\in\mathcal{C}} \frac{\widehat{p}(x_i)}{1-\widehat{p}(x_i)} Y_i}{\sum_{i'\in\mathcal{C}} \frac{\widehat{p}(x_{i'})}{1-\widehat{p}(x_{i'})}} =: \widehat{\tau}^{\mathrm{IPW}}_{\mathrm{ATT}}. \tag{5.2.9}$$

Namely, the proposed KSC method is guaranteed to yield individual treatment effect estimates that conform to (5.2.9) at the aggregate level. Similarly, solving (5.2.7) by setting

$$v_j = \frac{1}{\widehat{p}(x_j)} \Big/ \sum_{j'\in\mathcal{T}} \frac{1}{\widehat{p}(x_{j'})} \quad \forall j \in \mathcal{T} \quad \text{and} \quad w_i = \frac{1}{1-\widehat{p}(x_i)} \Big/ \sum_{i'\in\mathcal{C}} \frac{1}{1-\widehat{p}(x_{i'})} \quad \forall i \in \mathcal{C},$$

we obtain an coupling $\widehat{\pi}$ that matches the normalized IPW estimator for ATE:

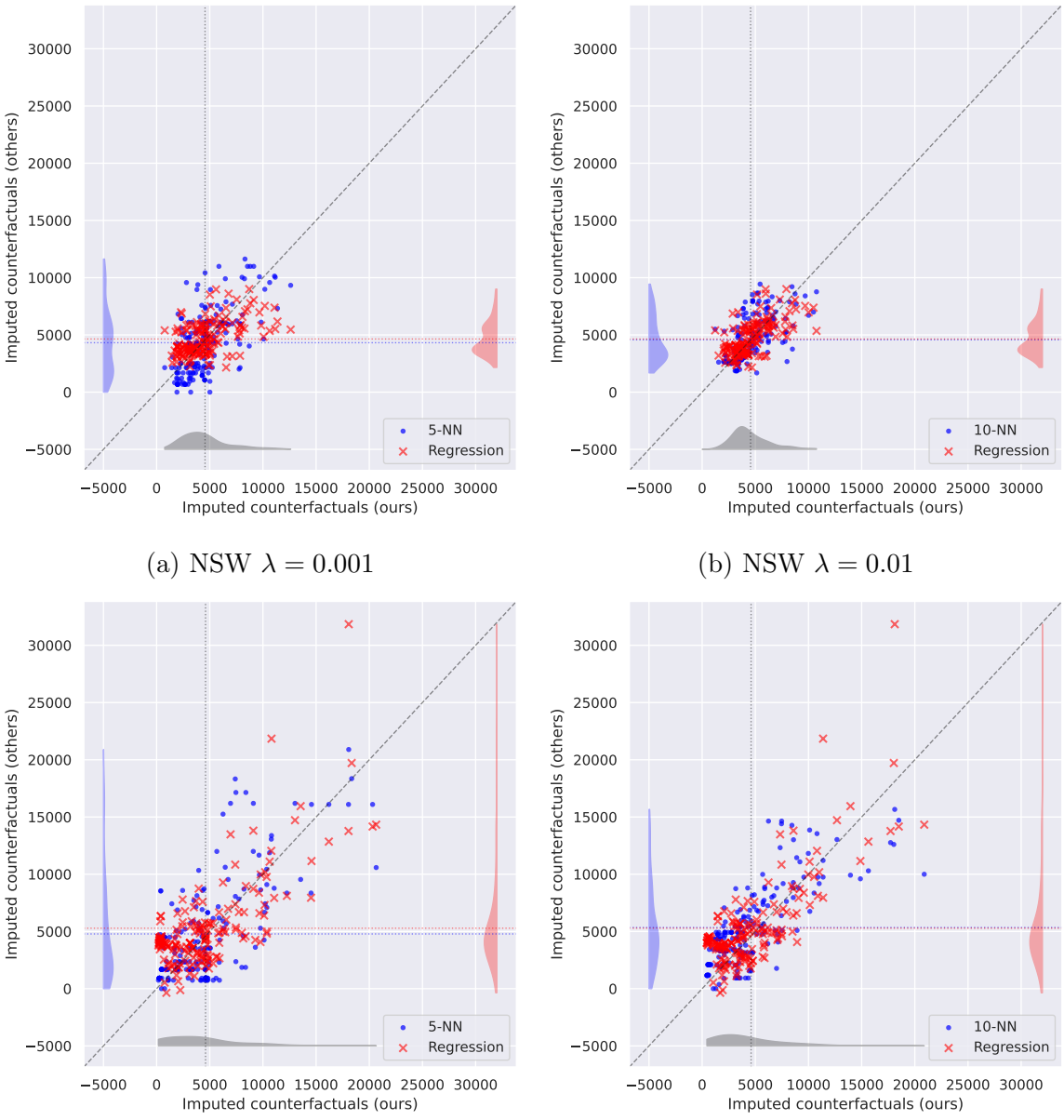
$$\sum_{j\in\mathcal{T}} v_j \widehat{\tau}_j = \frac{\sum_{j\in\mathcal{T}} \frac{1}{\widehat{p}(x_j)} Y_j}{\sum_{j'\in\mathcal{T}} \frac{1}{\widehat{p}(x_{j'})}} - \frac{\sum_{i\in\mathcal{C}} \frac{1}{1-\widehat{p}(x_i)} Y_i}{\sum_{i'\in\mathcal{C}} \frac{1}{1-\widehat{p}(x_{i'})}} =: \widehat{\tau}^{\mathrm{IPW}}_{\mathrm{ATE}}.$$

5.2.3 Numeric Illustration

Now, we demonstrate our imputation method, KSC, using a real-world data set. This section aims to concisely present the features of KSC and compare the results with other imputation methods; further details and robust evaluations are provided in Section 5.7 together with the inference procedure presented in Section 5.3.

We apply the proposed method to evaluate the National Supported Work (NSW) demonstration program, first analyzed by LaLonde [1986]. This data set has been widely studied in the program evaluation literature. The NSW program was conducted from 1975 to 1978 in the United States, which aimed at providing a job training program to disadvantaged workers, where the treatment, namely, the training, was randomly assigned to some of them. The data we consider here is a specific subset of the experimental data used in Dehejia and Wahba [1999], which consists of $N = 445$ subjects with $N_t = 185$ treated units and $N_c = 260$ control units. The outcome of interest Y_i is the post-treatment earnings recorded in 1978. Each individual is associated with six variables: age, years of education, and four indicator variables denoting whether the individual is black, Hispanic, married, and a high school dropout. Following Dehejia and Wahba [1999], the pretreatment outcomes, the earnings in 1974 and 1975, and two indicator variables denoting whether these earnings are zero, are included as covariates. Accordingly, we have $x_i \in \mathbb{R}^d$ with $d = 10$.

Figure 5.1(a) and Figure 5.1(b) show the scatter plots of the imputed counterfactual outcomes of the treated units, where the x-coordinate is $\widehat{Y}_j(0)$ imputed by the proposed method, while the y-coordinate is the imputed counterfactual outcomes by the nearest neighbor matching with replacement or the regression imputation method. The proposed method relies on the linear kernel following the simple formulation in Section 5.2.1 with $\lambda \in \{0.001, 0.01\}$, the nearest neighbor matching uses 5 and 10 nearest neighbors, and the regression method fits a linear regression model to the observed outcome using all the covariates and the treatment indicator. At first glance, the scatter plots visually suggest that



(a) NSW $\lambda = 0.001$

(b) NSW $\lambda = 0.01$

(c) PSID $\lambda = 0.001$

(d) PSID $\lambda = 0.01$

Figure 5.1: Scatter plots of the imputed counterfactual outcomes of the treated along with the histograms of the marginal distributions. The x-coordinate is $\widehat{Y}_j(0)$ imputed by the proposed method with $\lambda \in \{0.001, 0.01\}$, while the y-coordinate is the imputed counterfactual outcomes by the nearest neighbor matching with replacement or the regression method. (a) and (b) are based on the NSW experimental data, while (c) and (d) are based on the trimmed PSID data. The proposed method uses the linear kernel for both data. For the NSW data, we use uniform weights v, w in (5.2.7), while for the PSID data, we set w to be the propensity score-based weights following (5.2.8).

all three imputation methods provide comparable imputed counterfactual outcomes. One crucial difference is that the average of the imputed counterfactual outcomes and thus the estimated individual treatment effects vary across the methods. The ATT estimate computed by the proposed method always matches—regardless of the choice of λ—the mean difference between the treatment and control groups, which is roughly 1794.3, as shown in (5.2.5)—the unbiased estimator of the ATT in the experimental setup. In contrast, the k-nearest neighbor matching with $k = 5$, with $k = 10$, and the regression imputation method produce the ATT estimates 2030.5, 1776.6, and 1706.2, respectively. While the average of the imputed counterfactual outcomes does not depend on the choice of λ, we can see that it determines the level of individualization. The histograms along the x-axis show that the imputed counterfactual outcomes based on the proposed method with $\lambda = 0.001$ are more dispersed than those based on $\lambda = 0.01$; the latter is based on larger regularization leading to more uniform weights, which is similar to the fact that $k = 5$ is more dispersed—along the y-axis—than $k = 10$ in the nearest neighbor matching.

Since LaLonde [1986], there has been a debate on the credibility of the average treatment effect estimation using observational control groups. Since such control groups can be significantly different from the control group of the experimental data, the resulting ATE estimates can deviate much from the experimental benchmark, say, the mean difference of 1794.3 computed earlier, which has been the central argument of LaLonde [1986]. Later, several methods [Dehejia and Wahba, 1999, 2002] have been proposed to recover the experimental benchmark estimate using observational control groups. Here, we focus on one such method based on the normalized inverse probability weighting (IPW) estimator (5.2.9) and apply the KSC method with the weights based on the propensity scores. To this end, we consider a control group from the Panel Study of Income Dynamics (PSID) data, consisting of 2490 units, with which we estimate the propensity scores by logistic regression. The resulting ATT estimate by $\widehat{\tau}_{\mathrm{ATT}}^{\mathrm{IPW}}$ is 2579.7, which is far from the benchmark of 1794.3,

rooted in the stark difference between the experimental and observational control groups. To remedy this, we trim the PSID data by only taking the units whose propensity scores are in $[0.05, 0.95]$, yielding 214 units. The resulting ATT estimate by $\widehat{\tau}^{\text{IPW}}_{\text{ATT}}$ is then 1748.0, which resembles the experimental benchmark. Taking this trimmed PSID data as a control group, we apply the proposed method with the weights w_i based on the propensity scores as in (5.2.8) and compare with the matching and regression methods. The results are shown in Figure 5.1(c) and Figure 5.1(d). Again, though visually similar in distributions, the ATT estimates differ across the methods, where the proposed method with the propensity score-based weights provides the ATT estimate 1748.0, regardless of λ. In contrast, the k-nearest neighbor matching with $k = 5$, with $k = 10$, and the regression method produce the ATT estimates 1567.3, 1003.0, and 1065.5, respectively. This example illustrates that the proposed imputation method allows for producing granular information, namely, the individual treatment effects, from a coarse level estimate by $\widehat{\tau}^{\text{IPW}}_{\text{ATT}}$. The coarse level estimate is closer to the experimental benchmark estimate than inherently imputation-based methods such as the nearest neighbor matching or regression.

5.3 Inference: Individual Confidence Intervals

Constructing confidence intervals for the estimated individual treatment effects is conducive to reliable decision-making. This section introduces and analyzes the method to construct individual confidence intervals based on the proposed convexified matching. To this end, we build a confidence interval for the missing counterfactual outcome $Y_j(0)$ for each treated unit $j \in \mathcal{T}$. We first introduce appropriate model assumptions. Then, we elucidate the construction of individual confidence intervals, followed by a numerical example, and discuss the coverage and efficiency of the proposed method.

5.3.1 Setup

The Model Let $\mathcal{H}$ be the reproducing kernel Hilbert space (RKHS) associated with a kernel function $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the RKHS inner product, $\| \cdot \|_{\mathcal{H}}$ denotes the RKHS norm, and $\phi_x := k(x, \cdot) \in \mathcal{H}$ is the corresponding reproducing kernel function. We postulate the following model for the potential outcomes: for each $i = 1, \ldots, N$, letting $x_i \in \mathbb{R}^d$ be the covariate of unit i, we have

$$\begin{aligned} Y_i(0) &= f_0(x_i) + \varepsilon_{0,i},\ \varepsilon_{0,i} \sim \mathcal{N}(0, \sigma_0^2), \\ Y_i(1) &= f_1(x_i) + \varepsilon_{1,i},\ \varepsilon_{1,i} \sim \mathcal{N}(0, \sigma_1^2), \end{aligned}$$

where $f_0, f_1 \in \mathcal{H}$ are some functions in the RKHS and $\sigma_0, \sigma_1 \geq 0$. Let $\mathbf{X} := [x_1, \ldots, x_N]$ denote the fixed design matrix and let $\mathbf{Z} := [Z_1, \ldots, Z_N]$ denote the possibly random treatment assignment vector.

Assumption 5.3.1. $\{\varepsilon_{0,i}, \varepsilon_{1,i}\}_{i=1}^N$ are independent and

$$\mathbf{Z} \perp\!\!\!\perp \{\varepsilon_{0,i}, \varepsilon_{1,i}\}_{i=1}^N \mid \mathbf{X}, \tag{5.3.1}$$

that is, the treatment assignment and the errors are independent given the covariates.

Remark 5.3.2. Typical assumptions in causal inference involve the following: (i) $(Y_i(0), Y_i(1), X_i)_{i=1}^N$ are independently drawn from some population distribution defined on $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ and (ii) the treatment assignment Z_i is independent of $(Y_i(0), Y_i(1))$ conditional on X_i. Unlike (i), our model postulates a signal-plus-noise structure for the potential outcomes, where the covariates are from a fixed design matrix. All our results are conditioned on the fixed design $\mathbf{X}$, which we highlight by using the lowercase $\{x_i\}$ notation instead of $\{X_i\}$. On the one hand, Assumption 5.3.1 serves the same purpose as (ii) in spirit but is stronger in the sense that the errors are assumed to be Gaussian. On the other hand, we do not require $\mathbf{X}$ nor $\mathbf{Z}$ to be i.i.d., allowing Z_i's to have arbitrary dependence across units or even to be adversarially

chosen conditioned on the fixed design $\mathbf{X}$.

Under this setup, for each treated unit $j \in \mathcal{T}$, the goal is to build an individual confidence interval for its missing counterfactual outcome $Y_j(0)$. More specifically, we aim to construct a confidence interval $[\widehat{L}_j, \widehat{U}_j]$ that contains the conditional expectation $f_0(x_j) = \mathbb{E}[Y_j(0) \,|\, x_j] = \mathbb{E}[Y_j(0) \,|\, \mathbf{X}, \mathbf{Z}]$ with a desired level of confidence as follows:

$$\mathbb{P}\left(f_0(x_j) \in [\widehat{L}_j, \widehat{U}_j] \mid \mathbf{X}, \mathbf{Z}\right) \geq 1 - \alpha, \tag{5.3.2}$$

namely, the probability that the interval $[\widehat{L}_j, \widehat{U}_j]$ contains the conditional expectation $f_0(x_j)$ is at least $1-\alpha$, where the probability is conditional on the design matrix $\mathbf{X}$ and the treatment assignments $\mathbf{Z}$. We mainly adopt a fixed-design perspective here and restrict the inference to the given data, rather than considering the generalization to new units as noted in Remark 5.3.2. The motivation stems from an individual's perspective: for each unit of the study, such as a participant in the NSW program in Section 5.2.3, the primary interest lies in knowing the missing counterfactual outcome; we are less concerned about how the results generalize to new samples with different covariates. By adopting this perspective, constructing intervals around the imputed counterfactual outcomes is a transparent approach that speaks directly to the primary interest. Therefore, we focus on this setup and design an intuitive method that practitioners can easily implement.

5.3.2 Individual Confidence Intervals

Construction We consider the following Kernel Synthetic Coupling (KSC) discussed in Section 5.2.2. To construct an individual interval for the counterfactual control outcome of each treated unit, we solve (5.2.7) with uniform weights on the treated ($v_j = \frac{1}{N_t}$), which we rewrite in the matrix form as follows (dropping the constant $\frac{\mathrm{tr}(K_{tt})}{2N_t}$ in (5.2.6)): let $(w_i)_{i \in \mathcal{C}}$

be positive numbers such that $\sum_{i\in\mathcal{C}} w_i = 1$,

$$\begin{aligned} \min_{\pi\in\mathbb{R}_+^{N_c\times N_t}} \quad & \tfrac{N_t}{2}\langle \pi, K_{cc}\pi\rangle - \langle \pi, K_{ct}\rangle + \lambda h(\pi), \\ \text{subject to} \quad & \sum_{i\in\mathcal{C}} \pi_{ij} = \tfrac{1}{N_t} \quad \forall j \in \mathcal{T} \quad \text{and} \quad \sum_{j\in\mathcal{T}} \pi_{ij} = w_i \quad \forall i \in \mathcal{C}, \end{aligned} \tag{5.3.3}$$

where $h(\pi)$ is the entropy term of π; see Section 5.4.

Let $\widehat{\pi}$ be the solution of the above optimization problem and define a probability transition matrix $\widehat{P} = (\widehat{p}_{ij})_{(i,j)\in\mathcal{C}\times\mathcal{T}}$, where $\widehat{p}_{ij} = N_t\widehat{\pi}_{ij}$ so that $\sum_{i\in\mathcal{C}} \widehat{p}_{ij} = 1$ for any $j \in \mathcal{T}$. For each treated unit j, recall the imputed value $\widehat{Y}_j(0)$ defined in (5.2.4), for which we construct a confidence interval $[\widehat{L}_j, \widehat{U}_j]$, where $\widehat{L}_j$ and $\widehat{U}_j$ are defined as follows:

$$\widehat{L}_j := \widehat{Y}_j(0) - \widehat{\theta}\cdot\sqrt{\left(K_{tt} + \widehat{P}^\top K_{cc}\widehat{P} - 2K_{ct}^\top\widehat{P}\right)_{jj}} - z_{1-\frac{\alpha}{2}}\cdot\widehat{\sigma}_0\sqrt{\left(\widehat{P}^\top\widehat{P}\right)_{jj}}, \tag{5.3.4}$$

$$\widehat{U}_j := \widehat{Y}_j(0) + \widehat{\theta}\cdot\sqrt{\left(K_{tt} + \widehat{P}^\top K_{cc}\widehat{P} - 2K_{ct}^\top\widehat{P}\right)_{jj}} + z_{1-\frac{\alpha}{2}}\cdot\widehat{\sigma}_0\sqrt{\left(\widehat{P}^\top\widehat{P}\right)_{jj}}. \tag{5.3.5}$$

Here, $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution and $\widehat{\theta}, \widehat{\sigma}_0$ are suitable estimates of the norm $\|f_0\|_{\mathcal{H}}$ and the standard deviation σ_0, respectively.

We apply kernel ridge regression to the control group data to estimate $\|f_0\|_{\mathcal{H}}$ and σ_0. We can estimate $f_0 \in \mathcal{H}$ by $\widehat{f}_0(x) = \sum_{i\in\mathcal{C}} \widehat{\beta}_i k(x_i, x)$ with $\widehat{\beta} = (K_{cc} + \rho I_{N_c})^{-1} Y_c \in \mathbb{R}^{N_c}$, where $\rho > 0$ is a ridge regularization parameter and $Y_c = (Y_i)_{i\in\mathcal{C}} \in \mathbb{R}^{N_c}$ is a vector of the control outcomes. Then, we estimate $\|f_0\|_{\mathcal{H}}$ and the variance σ_0^2 by $\widehat{\theta} := \|\widehat{f}_0\|_{\mathcal{H}} = \sqrt{\langle\widehat{\beta}, K_{cc}\widehat{\beta}\rangle}$ and

$$\widehat{\sigma}_0^2 := \frac{1}{N_c}\sum_{i\in\mathcal{C}}(Y_i - \widehat{f}_0(x_i))^2 = \frac{1}{N_c}\langle Y_c - K_{cc}\widehat{\beta}, Y_c - K_{cc}\widehat{\beta}\rangle.$$

Before diving into each component of the construction, we first look at a simple numerical example to illustrate the performance of the proposed method.

Numerical Example We consider fixed one-dimensional covariates $x_1, \ldots, x_N \in [0,1]$ with $N = 500$ and randomly choose $N_t = 200$ treatment units. Let $\mathcal{H}$ be the RKHS associated with the Gaussian kernel $k(x, x') = \exp\bigl(-\gamma \cdot |x - x'|^2\bigr)$ with $\gamma = 2.5$, and let $f_0(x) = k(0.5, x)$, which yields $\|f_0\|_{\mathcal{H}} = (k(0.5, 0.5))^{1/2} = 1$. Then, we generate the potential outcomes by $Y_i(0) = f_0(x_i) + \varepsilon_{0,i}$, where $\varepsilon_{0,1}, \ldots, \varepsilon_{0,N}$ are independently drawn from $N(0, \sigma_0^2)$. As $x_1, \ldots, x_N$ and $Z_1, \ldots, Z_N$ are fixed, the randomness is solely from the residuals $\varepsilon_{0,i}$'s, which we sample 1,000 times.

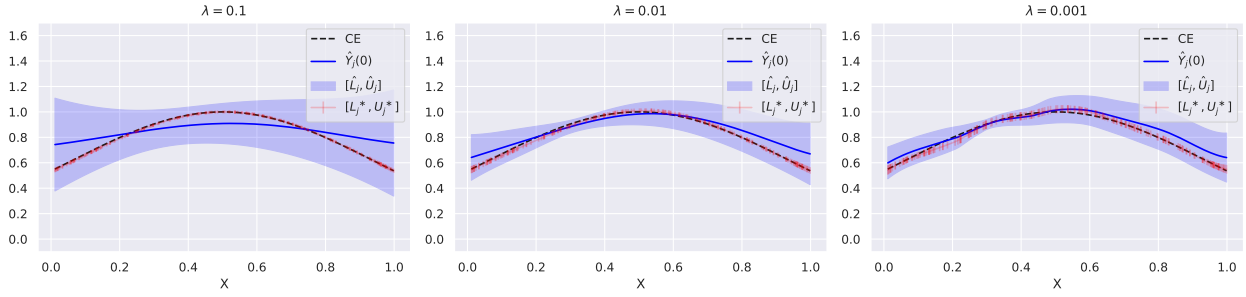
Figure 5.2 shows the constructed confidence intervals $[\widehat{L}_j, \widehat{U}_j]$ for each treated unit j for different values of σ_0, with $\alpha = 0.05$. For each $\sigma_0 \in \{0.1, 1, 3\}$, we solve (5.3.3) for $\lambda \in \{0.1, 0.01, 0.001\}$ with uniform weights w_i's, run the kernel ridge regression to obtain $\widehat{\theta}, \widehat{\sigma}_0$, and produce $[\widehat{L}_j, \widehat{U}_j]$ which is shown as blue shaded regions in the plots. We compare this interval with an "oracle" interval $[L_j^*, U_j^*]$, which—as we will explain in the next section—is guaranteed to have the exact $1 - \alpha$ coverage, namely,

$$\mathbb{P}\Big(f_0(x_j) \in [L_j^*, U_j^*] \mid \mathbf{X}, \mathbf{Z}\Big) = 1 - \alpha. \tag{5.3.6}$$

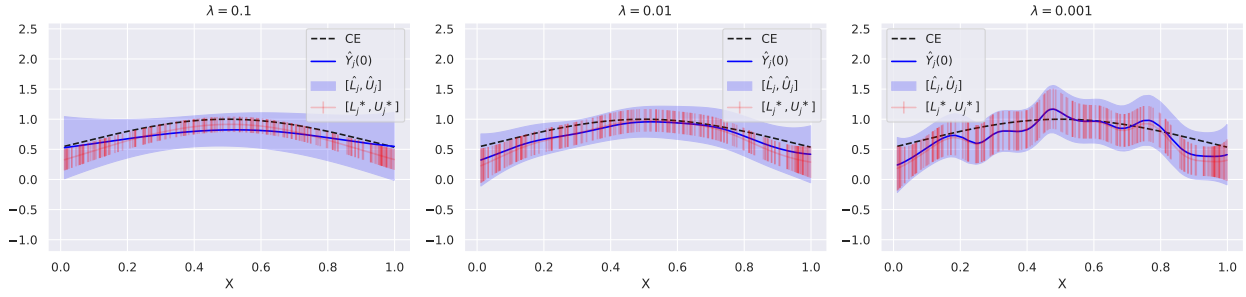
By construction, if the estimates $\widehat{\theta}, \widehat{\sigma}_0$ are accurate enough, the interval $[\widehat{L}_j, \widehat{U}_j]$ must contain the oracle interval $[L_j^*, U_j^*]$, which can be verified in Figure 5.2. For each σ_0, decreasing λ results in smaller approximation errors, namely, the bias correction reflected in $\Big(K_{tt} + \widehat{P}^\top K_{cc} \widehat{P} - 2K_{ct}^\top \widehat{P}\Big)_{jj}^{1/2}$, and the estimated counterfactual outcomes (solid blue curves) that are more wiggly, as the obtained weights $\widehat{P}$ have more variations. As we will see, $[\widehat{L}_j, \widehat{U}_j]$ is wider than $[L_j^*, U_j^*]$ by the twice of the bias correction, namely,

$$(\widehat{U}_j - \widehat{L}_j) - (U_j^* - L_j^*) = 2 \cdot \widehat{\theta} \cdot \Big(K_{tt} + \widehat{P}^\top K_{cc} \widehat{P} - 2K_{ct}^\top \widehat{P}\Big)_{jj}^{1/2}.$$

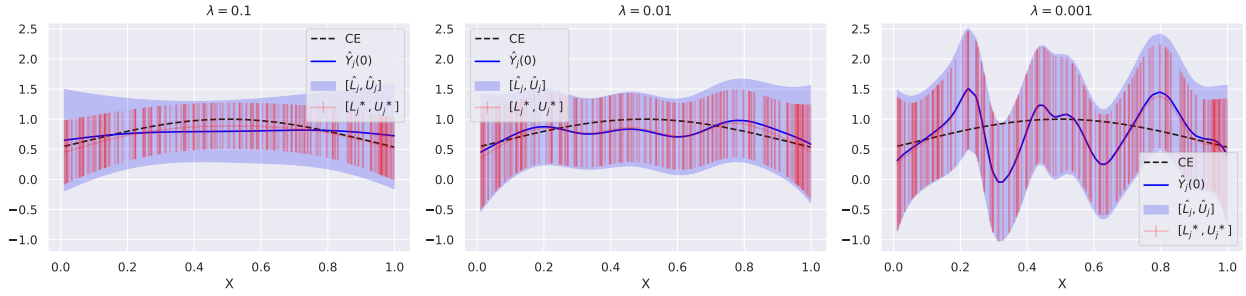
We can verify from Figure 5.2 that this gap between the two intervals gets smaller as λ decreases, consistent with the above observation that the bias decreases.



(a) $\sigma_0 = 0.1$



(b) $\sigma_0 = 1$

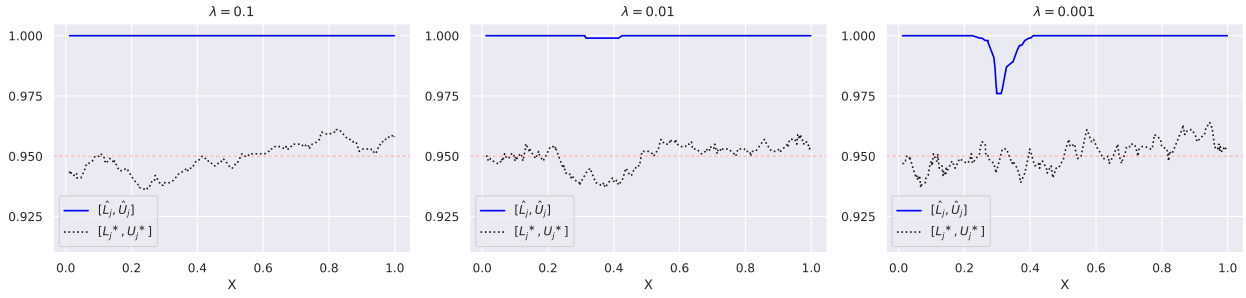


(c) $\sigma_0 = 3$

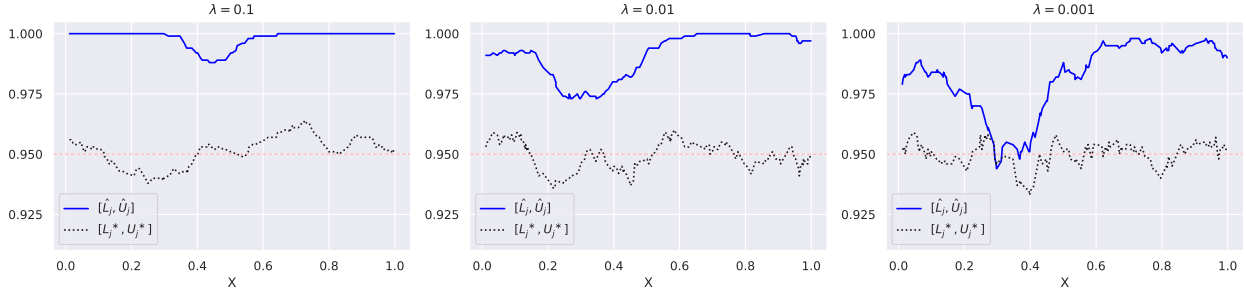
Figure 5.2: Confidence intervals of potential outcomes generated by $Y_i(0) = f_0(x_i) + \varepsilon_{0,i}$, where $\varepsilon_{0,1}, \ldots, \varepsilon_{0,N}$ are i.i.d. from $N(0, \sigma_0^2)$, with $\alpha = 0.05$. The black dashed curve is the ground truth conditional expectation function $f_0(x) = \exp\big(-2.5 \times |x - 0.5|^2\big)$. The blue solid curve is the estimated counterfactual outcome $\widehat{Y}_j(0) = \sum_{i \in \mathcal{C}} \widehat{p}_{ij} Y_i$ for each treated unit j. The blue shaded region shows $[\widehat{L}_j, \widehat{U}_j]$ defined in (5.3.4), (5.3.5), where $\widehat{\theta}, \widehat{\sigma}_0$ are estimated by the kernel ridge regression as explained in Section 5.3.2. The oracle interval $[L_j^*, U_j^*]$ defined in (5.3.9), (5.3.10) is shown using the red error bars.

Figure 5.3 plots the coverage of $[\widehat{L}_j, \widehat{U}_j]$ and $[L_j^*, U_j^*]$, namely, the probability that they contain $f_0(x_j)$, estimated using the 1,000 samples. When $\lambda = 0.1$, we can see that the coverage of $[\widehat{L}_j, \widehat{U}_j]$ is almost 1, meaning that the interval is too conservative, which results from the fact that the bias term is not small enough compared to the variance of the residuals. As λ decreases, we can see that the coverage of $[\widehat{L}_j, \widehat{U}_j]$ gets closer to $1 - \alpha = 0.95$, which is

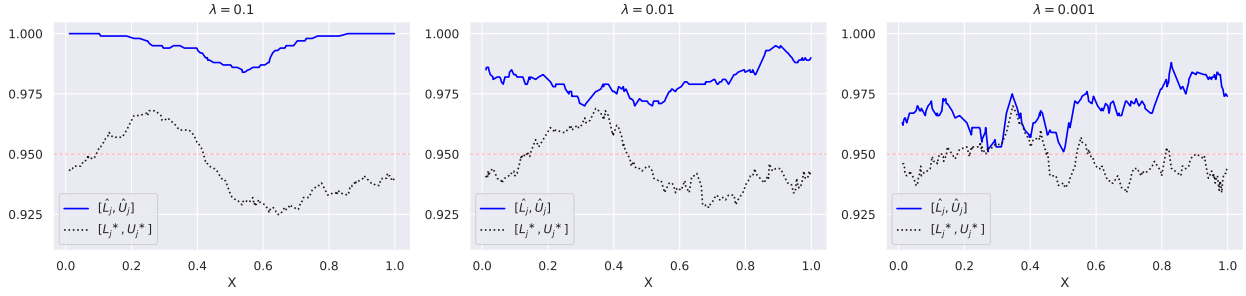
consistent with the fact that the bias decreases. Particularly, when $\sigma_0 = 3$ and $\lambda = 0.001$, the coverage of $[\widehat{L}_j, \widehat{U}_j]$ is closest to the coverage of the oracle interval $[L_j^*, U_j^*]$, suggesting that the bias is dominated by the variance. Meanwhile, for $(\sigma_0, \lambda) = (1, 0.01)$ and $(\sigma_0, \lambda) = (3, 0.1)$, the coverage of $[\widehat{L}_j, \widehat{U}_j]$ is larger than the coverage of the oracle interval $[L_j^*, U_j^*]$ by approximately a constant, which suggests that the bias term and the variance term are nearly balanced. In the next section, we provide theoretical guidance to choose the regularization parameter λ to balance the bias and variance tradeoff.



(a) $\sigma_0 = 0.1$



(b) $\sigma_0 = 1$



(c) $\sigma_0 = 3$

Figure 5.3: Coverage of the confidence intervals $[\widehat{L}_j, \widehat{U}_j]$ and $[L_j^*, U_j^*]$ estimated using the 1,000 samples. The red dashed line shows the nominal coverage $1 - \alpha = 0.95$.

5.3.3 Coverage and Efficiency

Now, we dive into each component in constructing individual confidence intervals to dissect the coverage and efficiency. We discuss how the entropic regularization parameter λ drives the bias and variance tradeoff, as seen in the numerical simulations. Furthermore, we show that optimizing the entropic-regularized convexified matching problem (5.3.3) with the desired λ chosen above is closely connected to optimizing the average of the squared lengths of the individual confidence intervals.

Point Estimate For each treated unit $j \in \mathcal{T}$, the center of the confidence interval $[\widehat{L}_j, \widehat{U}_j]$ is the imputed counterfactual outcome $\widehat{Y}_j(0) = \sum_{i \in \mathcal{C}} \widehat{p}_{ij} Y_i(0)$. This weighted average estimator leverages all the information in $\mathbf{X}$ to anchor an optimal coupling between the treated and control units for synthesizing the controls. It uses convex combinations of the control units to synthesize the granular counterfactual outcomes for the treated units, while balancing the aggregate level statistics, the ATT. We remark that the optimal coupling $\widehat{\pi}$—and thus $\widehat{P} = N_t \widehat{\pi}$ —is calculated based solely on $\mathbf{X}$ and $\mathbf{Z}$, not the outcomes Y_i's. Now, recall the following bias-variance decomposition of the point estimate $\widehat{Y}_j(0)$:

$$\widehat{Y}_j(0) = \mathbb{E}[Y_j(0) \,|\, \mathbf{X}, \mathbf{Z}] + \underbrace{\widehat{Y}_j(0) - \mathbb{E}[\widehat{Y}_j(0) \,|\, \mathbf{X}, \mathbf{Z}]}_{:=\mathcal{V}_j} + \underbrace{\mathbb{E}[\widehat{Y}_j(0) \,|\, \mathbf{X}, \mathbf{Z}] - \mathbb{E}[Y_j(0) \,|\, \mathbf{X}, \mathbf{Z}]}_{:=\mathcal{B}_j}.$$

Bias-Awareness The point estimate could be biased as

$$\mathcal{B}_j = \mathbb{E}[\widehat{Y}_j(0) \,|\, \mathbf{X}, \mathbf{Z}] - \mathbb{E}[Y_j(0) \,|\, \mathbf{X}, \mathbf{Z}] = \sum_{i \in \mathcal{C}} \widehat{p}_{ij} f_0(x_i) - f_0(x_j) = \langle f_0, \sum_{i \in \mathcal{C}} \widehat{p}_{ij} \phi_{x_i} - \phi_{x_j} \rangle_{\mathcal{H}},$$

where the last equality uses the reproducing property of the RKHS. Since f_0 is unknown and estimating such a function uniformly on the support of the covariate vector can be difficult,

we rely on a bias-awareness correction by invoking the Cauchy-Schwarz inequality:

$$|\mathcal{B}_j| \le \|f_0\|_{\mathcal{H}} \cdot \|\phi_{x_j} - \sum_{i\in\mathcal{C}} \widehat{p}_{ij}\phi_{x_i}\|_{\mathcal{H}} = \|f_0\|_{\mathcal{H}} \cdot \left(K_{tt} + \widehat{P}^\top K_{cc}\widehat{P} - 2K_{ct}^\top \widehat{P}\right)_{jj}^{1/2}. \quad (5.3.7)$$

This explains subtracting and adding the term $\widehat{\theta} \cdot \left(K_{tt} + \widehat{P}^\top K_{cc}\widehat{P} - 2K_{ct}^\top \widehat{P}\right)_{jj}^{1/2}$ in (5.3.4) and (5.3.5), respectively, where $\widehat{\theta}$ estimates the unknown quantity $\|f_0\|_{\mathcal{H}}$.

Variance Under the model,

$$\mathcal{V}_j = \widehat{Y}_j(0) - \sum_{i\in\mathcal{C}} \widehat{p}_{ij} f_0(x_i) = \sum_{i\in\mathcal{C}} \widehat{p}_{ij}\varepsilon_{0,i} \sim N(0, \sigma_0^2 \sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2). \quad (5.3.8)$$

If we knew the bias $\mathcal{B}_j$, we could construct the following oracle confidence interval $[L_j^*, U_j^*]$:

$$L_j^* := \widehat{Y}_j(0) - \mathcal{B}_j - z_{1-\frac{\alpha}{2}} \cdot \sigma_0 \sqrt{\sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2}, \quad (5.3.9)$$

$$U_j^* := \widehat{Y}_j(0) - \mathcal{B}_j + z_{1-\frac{\alpha}{2}} \cdot \sigma_0 \sqrt{\sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2}, \quad (5.3.10)$$

Due to (5.3.8), the oracle interval $[L_j^*, U_j^*]$ is guaranteed to have the exact $1-\alpha$ coverage, namely, (5.3.6) holds. However, since $\mathcal{B}_j$ is unknown, we rely on the aforementioned bias-awareness correction to construct the confidence interval $[\widehat{L}_j, \widehat{U}_j]$, obtained by inflating the oracle interval using (5.3.7).

Efficiency: Interval Length We now explain how the width of the confidence interval is related to our convexified matching objective. As a result, the explanation will shed light on why the entropic regularization serves as a tuning parameter to trade off bias and variance for overall interval length efficiency. First, we need the following simple facts on the Kullback-Leibler divergence, Hellinger distance, and χ^2 distance.

Proposition 5.3.3. *For any probability vector $p \in \Delta_n$,*

$$\frac{1}{2(n\|p\|_\infty + 1)} \leq \frac{\sum_{i=1}^n p_i \log(p_i) + \log(n)}{n\sum_{i=1}^n p_i^2 - 1} \leq 1.$$

The above fact will help us relate the entropy term and the standard Euclidean norm of the weights. For a fixed $j \in \mathcal{T}$, suppose that $(\widehat{p}_{ij})_{i\in\mathcal{C}} \in \Delta_{N_c}$, the j-th column of $\widehat{P}$, is delocalized in the sense $\max_{i\in\mathcal{C}} \widehat{p}_{ij} \leq (M/2-1)/N_c$ with some constant $M > 4$. The interval length is as follows:

$$\mathrm{Len}_j := 2 \cdot \|f_0\|_{\mathcal{H}} \cdot \|\phi_{x_j} - \sum_{i\in\mathcal{C}} \widehat{p}_{ij}\phi_{x_i}\|_{\mathcal{H}} + 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \sigma_0 \sqrt{\sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2}.$$

From $\widehat{\pi}_{ij} = \widehat{p}_{ij}/N_t$ and Proposition 5.3.3, we deduce the following bounds:

$$\mathrm{Len}_j^2 \leq 8\left\{\|f_0\|_{\mathcal{H}}^2 \|\phi_{x_j} - \sum_{i\in\mathcal{C}} \frac{\widehat{\pi}_{ij}}{1/N_t}\phi_{x_i}\|_{\mathcal{H}}^2 + \frac{M z_{1-\frac{\alpha}{2}}^2 \sigma_0^2}{N_c/N_t}\left[\sum_{i\in\mathcal{C}} \widehat{\pi}_{ij} \log\frac{\widehat{\pi}_{ij}}{e} + \frac{\frac{1}{M}+1+\log(N_cN_t)}{N_t}\right]\right\},$$

$$\mathrm{Len}_j^2 \geq 4\left\{\|f_0\|_{\mathcal{H}}^2 \|\phi_{x_j} - \sum_{i\in\mathcal{C}} \frac{\widehat{\pi}_{ij}}{1/N_t}\phi_{x_i}\|_{\mathcal{H}}^2 + \frac{z_{1-\frac{\alpha}{2}}^2 \sigma_0^2}{N_c/N_t}\left[\sum_{i\in\mathcal{C}} \widehat{\pi}_{ij} \log\frac{\widehat{\pi}_{ij}}{e} + \frac{2+\log(N_cN_t)}{N_t}\right]\right\}.$$

Averaging over $j \in \mathcal{T}$, and barring the constant M, the upper and lower bounds on the right-hand sides remind us of the entropic regularized convexified matching objective in (5.2.7), with the regularization parameter

$$\lambda \asymp \frac{z_{1-\frac{\alpha}{2}}^2}{N_c}\frac{\sigma_0^2}{\|f_0\|_{\mathcal{H}}^2}.$$

The above tradeoff confirms the empirical findings: for a high signal-to-noise problem, to balance bias and variance, we need a smaller λ for better approximation error; for a low signal-to-noise problem, a larger λ is preferred to reduce variance. Such a phenomenon confirms the observations in Figure 5.2. Searching for a coupling minimizing the objective

of convexified matching is closely related to minimizing the interval length, provided the regularization parameter λ is appropriately chosen. Therefore, optimizing the entropic regularized convexified matching program (5.2.7) with the desired λ chosen above is closely connected to optimizing the average of the squared lengths of the individual confidence intervals $\frac{1}{N_t}\sum_{j\in\mathcal{T}}\mathrm{Len}_j^2$.

5.4 Optimization: Algorithms and Analysis

This section introduces and analyzes optimization algorithms to solve the convexified matching problem. For a cleaner analysis, we focus on (5.3.3), a version of (5.2.7) with uniform weights $v = \frac{1}{N_t}$ on the treated. Notice that it is an instance of the following constrained nonlinear optimization problem: for $n, m \in \mathbb{N}$, $a \in \Delta_n^+$, $b \in \Delta_m^+$, $\lambda > 0$, and $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$,

$$\min_{\pi\in\Pi_{a,b}} \; g(\pi) + \lambda h(\pi), \tag{5.4.1}$$

where $\Pi_{a,b} = \{\pi \in \mathbb{R}_+^{n\times m} : \pi 1_m = a \text{ and } \pi^\top 1_n = b\}$ and $h\colon \mathbb{R}_+^{n\times m} \to \mathbb{R}$ is the entropy function defined by $h(\pi) = \sum_{i=1}^n \sum_{j=1}^m \pi_{ij} \log \frac{\pi_{ij}}{e}$. The convexified matching amounts to the case where g is a convex quadratic function, in which (5.4.1) is a convex program taking the form (5.3.3). It can be solved using the interior point method in poly(nm) time, possibly slow in practice for coupling matrices with nm large.

In this section, we devise a faster algorithm that requires an almost *dimension-free* number of oracle calls to a subroutine—a simple iterative matrix scaling procedure called the Sinkhorn algorithm [Sinkhorn, 1967]. When λ is sufficiently large, the algorithm compares favorably to the interior point method, requiring only a $\log(1/\varepsilon)$ number of Sinkhorn subroutines. Later, we modify and extend the algorithm to cover all $\lambda \in \mathbb{R}_+$ by connecting to the steepest descent method under the Kullback-Leibler divergence. The convergence is admittedly slower for small λ: $\log(nm)/\varepsilon$ number of Sinkhorn subroutines is required, but

it still compares favorably to the interior point method.

Additional Notation For a matrix $A \in \mathbb{R}^{n\times m}$, let $\|A\|_1 := \sum_{i=1}^n \sum_{j=1}^m |A_{ij}|$ denote its L^1 norm, and $\|A\|_\infty := \max_{1\le i\le n} \max_{1\le j\le m} |A_{ij}|$ denote its L^∞ norm. Let h keep its definition as the entropy function, and let $\nabla h(A)$ denote the gradient of h, which is $\log(A)$, the entrywise logarithm of A, provided all entries of A are positive. For any $a \in \Delta_n$ and $b \in \Delta_m$, let $\Pi_{a,b}^+ := \{A \in \Pi_{a,b} : A_{ij} > 0 \ \forall i, j\}$. Let $\Delta_{n,m} = \{A \in \mathbb{R}_+^{n\times m} : \sum_{i=1}^n \sum_{j=1}^m A_{ij} = 1\}$. For vectors $u, v \in \mathbb{R}^n$, let $\frac{u}{v}$ denote the entrywise division provided all the entries of v are nonzero.

5.4.1 Optimality Conditions and the Sinkhorn Algorithm

When g is convex and $\lambda > 0$, the strong convexity of h on $\Pi_{a,b}$ implies that (5.4.1) admits a unique minimizer. Moreover, as h prevents the minimizer from having a zero entry, (5.4.1) admits a unique minimizer on $\Pi_{a,b}^+$, which is the interior of $\Pi_{a,b}$. It turns out that this unique minimizer of (5.4.1) is the fixed point of an operator related to the entropic regularized optimal transport problem [Cuturi, 2013], as we shall show in Proposition 5.4.2. Later, in Theorem 5.4.3, we derive convergence to the minimizer via the iterative matrix scaling subroutine. This subroutine is referred to as the Sinkhorn algorithm [Sinkhorn, 1967], which we introduce below.

Definition 5.4.1 (Sinkhorn [Sinkhorn, 1967, Cuturi, 2013])**.** Fix $\lambda > 0$. The following operator $\Phi_\lambda \colon \mathbb{R}^{n\times m} \to \Pi_{a,b}^+$ is well-defined:

$$\Phi_\lambda(C) := \operatorname*{arg\,min}_{\pi\in\Pi_{a,b}} \left(\langle C, \pi\rangle + \lambda h(\pi)\right) \quad \forall C \in \mathbb{R}^{n\times m},$$

where we often call the right-hand side the entropic regularized optimal transport problem given a cost matrix C. Moreover, for any $C \in \mathbb{R}^{n\times m}$, one can obtain $\Phi_\lambda(C)$ by iteratively scaling the rows and columns of the matrix $\exp(-C/\lambda)$ using the Sinkhorn algorithm

summarized in Algorithm 1, namely, $\Phi_\lambda(C)$ is the limit of the sequence produced from Sinkhorn$(e^{-C/\lambda}, a, b)$.

Algorithm 1 Sinkhorn(P, a, b)

Require: A matrix $P \in \mathbb{R}^{n\times m}$ with positive entries, $a \in \Delta_n^+$, and $b \in \Delta_m^+$.

1: Initialize $k \leftarrow 0$, $x^{(k)} \leftarrow 1_n$, and $y^{(k)} \leftarrow \frac{b}{P^\top x^{(k)}}$.

2: **repeat**

3:
$$x^{(k+1)} \leftarrow \frac{a}{Py^{(k)}} \quad \text{and} \quad y^{(k+1)} \leftarrow \frac{b}{P^\top x^{(k+1)}}.$$

4: $P^{(k+1)} = \mathrm{diag}(x^{(k+1)})P\mathrm{diag}(y^{(k+1)})$.

5: $k \leftarrow k + 1$.

6: **until** Discrepancy between $P^{(k)}1_m$ and a reaches a desired level.

7: **return** $P^{(k)}$.

For simplicity, Algorithm 1 states the Sinkhorn algorithm such that the column sum of $P^{(k)}$ matches with b, namely, $(P^{(k)})^\top 1_n = b$ for any $k \geq 0$. Then, to assess the accuracy of the scaled matrix $P^{(k)}$, we may only need to check a suitable discrepancy between a and the row sum of $P^{(k)}$, that is, $P^{(k)}1_m$.

The following proposition shows that when g is convex, the unique minimizer of (5.4.1) is the fixed point of an operator given by the composition of Φ_λ and ∇g.

Proposition 5.4.2. *Fix $\lambda > 0$. For a function $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$ such that ∇g exists, define an operator $T_\lambda\colon \mathbb{R}^{n\times m} \to \Pi_{a,b}^+$ by $T_\lambda = \Phi_\lambda \circ \nabla g$, that is, for any $A \in \mathbb{R}^{n\times m}$,*

$$T_\lambda(A) = \operatorname*{arg\,min}_{\pi\in\Pi_{a,b}} (\langle \nabla g(A), \pi\rangle + \lambda h(\pi)). \tag{5.4.2}$$

If g is convex, (5.4.1) admits a unique minimizer contained in $\Pi_{a,b}^+$, say, $\pi^\star \in \Pi_{a,b}^+$, and $\pi^\star$ is the unique fixed point of T_λ, namely, $\pi^\star = T_\lambda(\pi^\star)$.

Based on the established connection between the optimality condition of (5.4.1) and the Sinkhorn algorithm, we propose two algorithms to solve (5.4.1) in the following sections.

5.4.2 Fixed-Point Algorithm

By Proposition 5.4.2, finding the fixed point of T_λ is equivalent to solving (5.4.1). Therefore, we propose solving (5.4.1) by the fixed-point iterations of the operator $T_\lambda \colon \mathbb{R}^{n\times m} \to \Pi^+_{a,b}$, which is simply choosing an initial point $\pi^{(0)} \in \Pi_{a,b}$ and iterating $\pi^{(k+1)} \leftarrow T_\lambda(\pi^{(k)})$ for $k \geq 0$, where $T_\lambda(\pi^{(k)}) = \Phi_\lambda(\nabla g(\pi^{(k)}))$ is approximated by $\mathrm{Sinkhorn}(e^{-\nabla g(\pi^{(k)})/\lambda}, a, b)$ as explained in Definition 5.4.1. Algorithm 2 summarizes this procedure.

Algorithm 2 $\mathrm{FixedPoint}(g, \lambda, a, b)$

Require: A map $g \colon \mathbb{R}^{n\times m} \to \mathbb{R}$, a parameter $\lambda > 0$, $a \in \Delta^+_n$, and $b \in \Delta^+_m$.
1: Pick any $\pi^{(0)} \in \Pi_{a,b}$ and set $k \leftarrow 0$.
2: **repeat**
3: $\quad \pi^{(k+1)} \leftarrow \mathrm{Sinkhorn}(e^{-\nabla g(\pi^{(k)})/\lambda}, a, b)$.
4: $\quad k \leftarrow k + 1$.
5: **until** Discrepancy between $\pi^{(k)}$ and $\pi^{(k-1)}$ reaches a desired level.
6: **return** $\pi^{(k)}$.

We show that T_λ is a contraction if the gradient of g is Lipschitz and λ is larger than the Lipschitz constant. In such a case, Algorithm 2 converges to the fixed point quickly. When g is a quadratic function, we can write the Lipschitz constant in terms of the L^∞ norm of the Hessian matrix, which is independent of the input dimension nm. We do not require g to be convex for this result; T_λ admits a unique fixed point due to the contraction argument, which is independent of Proposition 5.4.2. The role of Proposition 5.4.2 is to translate this convergence result into the convergence to the minimizer of (5.4.1) for convex g by equating the fixed point to the minimizer.

Theorem 5.4.3. *Let $g \colon \mathbb{R}^{n\times m} \to \mathbb{R}$ be a quadratic function given by $g(\pi) = \frac{1}{2}\langle \pi, H\pi\rangle + \langle C, \pi\rangle$ for some $H \in \mathbb{R}^{n\times n}$ that is symmetric and $C \in \mathbb{R}^{n\times m}$. Assume $\lambda > \|H\|_\infty$. Then, the operator T_λ defined in* (5.4.2) *is a contraction under the distance defined by $\|\cdot\|_1$ and has a unique fixed-point, say, $\pi^\star \in \Pi^+_{a,b}$. Moreover, Algorithm 2, assuming the inner loop*

Sinkhorn *is always exact, outputs a sequence* $(\pi^{(k)})_{k\geq 0}$ *such that for any* $T \in \mathbb{N}$,

$$\|\pi^{(T)} - \pi^\star\|_1 \leq \frac{(\|H\|_\infty/\lambda)^T}{1-(\|H\|_\infty/\lambda)}\|\pi^{(1)} - \pi^{(0)}\|_1. \tag{5.4.3}$$

Remark 5.4.4. Theorem 5.4.3 shows that to obtain an ε-approximate fixed point $\|\pi^{(T)} - \pi^\star\|_1^2 \leq \varepsilon$ when $\lambda > \|H\|_\infty$, we need $\log(1/\varepsilon)/2\log(\lambda/\|H\|_\infty)$ number of Sinkhorn routines, which is a dimension-free quantity.

Theorem 5.4.3 is based on the following lemma that analyzes the Lipschitz property of the operator Φ_λ—defined in Proposition 5.4.1—w.r.t. the L^∞ metric.

Lemma 5.4.5. *Fix* $\lambda > 0$. *Define a map* $\Phi_\lambda \colon \mathbb{R}^{n\times m} \to \Pi_{a,b}^+$ *by*

$$\Phi_\lambda(C) = \operatorname*{arg\,min}_{\pi\in\Pi_{a,b}} \left(\langle C, \pi\rangle + \lambda h(\pi)\right).$$

Then, for any $C_1, C_2 \in \mathbb{R}^{n\times m}$, *we have*

$$\|\Phi_\lambda(C_1) - \Phi_\lambda(C_2)\|_1 \leq \frac{\|C_1 - C_2\|_\infty}{\lambda}. \tag{5.4.4}$$

Proof of Theorem 5.4.3. By (5.4.4) of Lemma 5.4.5, we have for any $A_1, A_2 \in \mathbb{R}^{n\times m}$,

$$\begin{aligned}
\|T_\lambda(A_1) - T_\lambda(A_2)\|_1 &= \|\Phi_\lambda(\nabla g(A_1)) - \Phi_\lambda(\nabla g(A_2))\|_1 \\
&\leq \frac{\|\nabla g(A_1) - \nabla g(A_2)\|_\infty}{\lambda} \\
&= \frac{\|H(A_1 - A_2)\|_\infty}{\lambda} \\
&\leq \frac{\|H(A_1 - A_2)\|_\infty}{\lambda} \\
&\leq \frac{\|H\|_\infty}{\lambda}\|A_1 - A_2\|_1,
\end{aligned}$$

where the last inequality follows from $\|H(A_1 - A_2)\|_\infty \leq \|H\|_\infty\|A_1 - A_2\|_1$. Therefore, T_λ is a contraction on $\Pi_{a,b}$ equipped with $\|\cdot\|_1$ if $\lambda > \|H\|_\infty$. By the Banach-Caccioppoli

theorem, T_λ has a unique fixed point, and we have (5.4.3). □

5.4.3 Local Algorithm with Kullback-Leibler Geometry

The fixed-point algorithm may not converge for sufficiently small λ. To fill in the gap when $\lambda \in [0, \|H\|_\infty)$, which is left unanswered by Theorem 5.4.3, we employ the steepest descent method under Bregman divergence to solve (5.4.1) as follows: let $f = g + \lambda h$,

$$\pi^{(k+1)} = \operatorname*{arg\,min}_{\pi \in \Pi_{a,b}} \left(\langle \nabla f(\pi^{(k)}), \pi \rangle + \frac{D_h(\pi, \pi^{(k)})}{\tau_k} \right), \tag{5.4.5}$$

where $D_h(\pi, \pi^{(k)}) = h(\pi) - h(\pi^{(k)}) - \langle \nabla h(\pi^{(k)}), \pi - \pi^{(k)} \rangle$ is the Bregman divergence, which equals the Kullback-Leibler divergence on $\Pi_{a,b}$, and $\tau_k > 0$ is a suitable step size. With the entropic regularization h, we implement the steepest descent over the polytope $\Pi_{a,b}$ under the Kullback-Leibler divergence. We emphasize that this can be solved using the Sinkhorn algorithm as well since the right-hand side of (5.4.5) is equivalent to minimizing $\langle \nabla g(\pi^{(k)}) + (\lambda - \tau_k^{-1}) \nabla h(\pi^{(k)}), \pi \rangle + \tau_k^{-1} h(\pi)$, which leads to

$$\pi^{(k+1)} = \Phi_{\tau_k^{-1}} \left(\nabla g(\pi^{(k)}) + (\lambda - \tau_k^{-1}) \nabla h(\pi^{(k)}) \right).$$

This can be approximated by applying Algorithm 1 to $e^{-(\nabla g(\pi^{(k)}) + (\lambda - \tau_k^{-1}) \nabla h(\pi^{(k)}))/\tau_k^{-1}}$. Algorithm 3 summarizes this procedure. Note that letting $\tau_k = \lambda^{-1}$ for all $k \geq 0$ in Algorithm 3 recovers Algorithm 2.

We show that Algorithm 3 outputs a sequence that converges to the minimum of (5.4.1) provided the step size is below a certain threshold.

Theorem 5.4.6. *Let $g \colon \mathbb{R}^{n \times m} \to \mathbb{R}$ be a convex quadratic function given by $g(\pi) = \frac{1}{2}\langle \pi, H\pi \rangle + \langle C, \pi \rangle$ for some $H \in \mathbb{R}^{n \times n}$ that is symmetric and positive semidefinite and $C \in \mathbb{R}^{n \times m}$. Then, for any $\lambda \geq 0$, if we run Algorithm 3, assuming the inner loop* Sinkhorn

Algorithm 3 SteepestDescentKL($g, \lambda, \{\tau_k\}_{k\geq 0}, a, b$)

Require: A map $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$, a parameter $\lambda > 0$, $a \in \Delta_n^+$, and $b \in \Delta_m^+$.
Require: Step sizes $\{\tau_k\}_{k\geq 0}$.
1: Pick any $\pi^{(0)} \in \Pi_{a,b}$ and set $k \leftarrow 0$.
2: **repeat**
3: $\quad \pi^{(k+1)} \leftarrow \mathrm{Sinkhorn}(e^{-(\nabla g(\pi^{(k)}) + (\lambda - \tau_k^{-1})\nabla h(\pi^{(k)}))/\tau_k^{-1}}, a, b)$.
4: $\quad k \leftarrow k + 1$.
5: **until** Discrepancy between $\pi^{(k)}$ and $\pi^{(k-1)}$ reaches a desired level.
6: **return** $\pi^{(k)}$.

is always exact, with constant step size $\tau_k = \tau$ for all $k \geq 0$, where $\tau^{-1} \geq \|H\|_\infty + \lambda$, it outputs a sequence $(\pi^{(k)})_{k\geq 0}$ such that for any $T \in \mathbb{N}$,

$$g(\pi^{(T)}) + \lambda h(\pi^{(T)}) - \min_{\pi \in \Pi_{a,b}} (g(\pi) + \lambda h(\pi)) \leq \frac{1}{T}\frac{\log(nm)}{\tau}. \tag{5.4.6}$$

When $\lambda > 0$, the following holds as well:

$$\|\pi^{(T)} - \pi^\star\|_1^2 \leq \frac{1}{T}\frac{2\log(nm)}{\lambda\tau}, \tag{5.4.7}$$

where $\pi^\star \in \Pi_{a,b}^+$ is the unique minimizer of (5.4.1).

Remark 5.4.7. The $\lambda \in [\|H\|_\infty, \infty)$ case has been studied in Theorem 5.4.3. Theorem 5.4.6 shows that for any $\lambda \in [0, \|H\|_\infty)$, with the choice $\tau^{-1} = 2\|H\|_\infty$, we need $\frac{1}{\varepsilon}\cdot 2\|H\|_\infty \log(nm)$ number of Sinkhorn subroutines to get ε-close to the objective value. Further more, when λ is strictly positive, we can assure that after $\frac{1}{\varepsilon}\cdot 4\|H\|_\infty \log(nm)\cdot\frac{1}{\lambda}$ number of Sinkhorn subroutines, $\|\pi^{(T)} - \pi^\star\|_1^2 \leq \varepsilon$, with a logarithmic dependence on the problem dimension.

Proof of Theorem 5.4.6. We first show the following from (5.4.5): for any $k \geq 0$ and $\pi \in \Pi_{a,b}$,

$$\langle \nabla f(\pi^{(k)}), \pi^{(k+1)} - \pi\rangle \leq \frac{\langle \nabla h(\pi^{(k)}) - \nabla h(\pi^{(k+1)}), \pi^{(k+1)} - \pi\rangle}{\tau_k}. \tag{5.4.8}$$

To see this, pick $\pi \in \Pi_{a,b}$ and define $q(\varepsilon) = f((1-\varepsilon)\pi^{(k+1)} + \varepsilon\pi) + \tau_k^{-1} D_h((1-\varepsilon)\pi^{(k+1)} +$

$\varepsilon\pi, \pi^{(k)})$ for $\varepsilon \in [0,1]$. As $q(\varepsilon) \geq q(0)$ for $\varepsilon \in [0,1]$, the right derivative of q at $\varepsilon = 0$ must be nonnegative, which leads to (5.4.8). Next, letting D_f be the Bregman divergence with respect to f, we have

$$\begin{aligned}
&f(\pi^{(k+1)}) - f(\pi) \\
&= f(\pi^{(k+1)}) - f(\pi^{(k)}) + f(\pi^{(k)}) - f(\pi) \\
&= D_f(\pi^{(k+1)}, \pi^{(k)}) + \langle \nabla f(\pi^{(k)}), \pi^{(k+1)} - \pi^{(k)} \rangle + f(\pi^{(k)}) - f(\pi) \\
&\leq D_f(\pi^{(k+1)}, \pi^{(k)}) + \langle \nabla f(\pi^{(k)}), \pi^{(k+1)} - \pi^{(k)} \rangle + \langle \nabla f(\pi^{(k)}), \pi^{(k)} - \pi \rangle \\
&\leq D_f(\pi^{(k+1)}, \pi^{(k)}) + \frac{\langle \nabla h(\pi^{(k)}) - \nabla h(\pi^{(k+1)}), \pi^{(k+1)} - \pi \rangle}{\tau_k},
\end{aligned}$$

where the first inequality uses the convexity of f and the second inequality follows from (5.4.8). For the last term on the right-hand side, we have

$$\begin{aligned}
&\langle \nabla h(\pi^{(k)}) - \nabla h(\pi^{(k+1)}), \pi^{(k+1)} - \pi \rangle \\
&= \langle \log \tfrac{\pi^{(k)}}{\pi^{(k+1)}}, \pi^{(k+1)} - \pi \rangle \\
&= \langle \log \tfrac{\pi}{\pi^{(k)}} - \log \tfrac{\pi}{\pi^{(k+1)}}, \pi \rangle - \langle \log \tfrac{\pi^{(k+1)}}{\pi^{(k)}}, \pi^{(k+1)} \rangle \\
&= D_h(\pi, \pi^{(k)}) - D_h(\pi, \pi^{(k+1)}) - D_h(\pi^{(k+1)}, \pi^{(k)}).
\end{aligned}$$

Hence,

$$f(\pi^{(k+1)}) - f(\pi) \leq D_f(\pi^{(k+1)}, \pi^{(k)}) + \tfrac{D_h(\pi,\pi^{(k)}) - D_h(\pi,\pi^{(k+1)}) - D_h(\pi^{(k+1)},\pi^{(k)})}{\tau_k}. \quad (5.4.9)$$

Meanwhile,

$$\begin{aligned}
D_f(\pi^{(k+1)}, \pi^{(k)}) &= \tfrac{\langle \pi^{(k+1)} - \pi^{(k)}, H(\pi^{(k+1)} - \pi^{(k)}) \rangle}{2} + \lambda D_h(\pi^{(k+1)}, \pi^{(k)}) \\
&\leq \frac{\|H\|_\infty}{2} \|\pi^{(k+1)} - \pi^{(k)}\|_1^2 + \lambda D_h(\pi^{(k+1)}, \pi^{(k)}) \\
&\leq (\|H\|_\infty + \lambda) D_h(\pi^{(k+1)}, \pi^{(k)}),
\end{aligned}$$

where the first inequality follows from Hölder's inequality and the second inequality uses Pinsker's inequality $\|\pi^{(k+1)} - \pi^{(k)}\|_1^2 \leq 2D_h(\pi^{(k+1)}, \pi^{(k)})$; see, Lemma 2.5 of Tsybakov [2009]. Now, as $\tau_k^{-1} = \tau^{-1} \geq \|H\|_\infty + \lambda$, we have

$$D_f(\pi^{(k+1)}, \pi^{(k)}) \leq \frac{D_h(\pi^{(k+1)}, \pi^{(k)})}{\tau},$$

which, together with (5.4.9), leads to

$$f(\pi^{(k+1)}) - f(\pi) \leq \frac{D_h(\pi, \pi^{(k)}) - D_h(\pi, \pi^{(k+1)})}{\tau}$$

for any $\pi \in \Pi_{a,b}$. By letting $\pi = \pi^{(k)}$, we have $f(\pi^{(k+1)}) \leq f(\pi^{(k)})$ for all $k \geq 0$. Hence,

$$f(\pi^{(T)}) - f(\pi) \leq \frac{1}{T}\sum_{k=0}^{T-1}(f(\pi^{(k+1)}) - f(\pi)) \leq \frac{D_h(\pi, \pi^{(0)}) - D_h(\pi, \pi^{(T)})}{T\tau} \leq \frac{D_h(\pi, \pi^{(0)})}{T\tau}.$$

As f always admits a minimizer on $\Pi_{a,b}$ (even when $\lambda = 0$), plugging a minimizer to π and using the fact that $D_h(\pi, \pi^{(0)}) \leq \log(nm)$, we obtain (5.4.6). Lastly,

$$\begin{aligned} f(\pi^{(T)}) - f(\pi^\star) &= \langle \nabla f(\pi^\star), \pi^{(T)} - \pi^\star \rangle + D_f(\pi^{(T)}, \pi^\star) \\ &= \langle \nabla f(\pi^\star), \pi^{(T)} - \pi^\star \rangle + \tfrac{\langle \pi^{(T)} - \pi^{(k)}, H(\pi^{(T)} - \pi^{(k)}) \rangle}{2} + \lambda D_h(\pi^{(T)}, \pi^\star) \\ &\geq \frac{\lambda}{2}\|\pi^{(T)} - \pi^\star\|_1^2, \end{aligned}$$

where the last inequality uses Pinsker's inequality and $\langle \nabla f(\pi^\star), \pi^{(T)} - \pi^\star \rangle \geq 0$, which holds as $\pi^\star$ is a minimizer of f. Hence, we have (5.4.7). □

5.5 Discussion

We proposed a convexified matching method for missing value imputation and individualized inference, integrating favorable features from optimal matching, regression imputation, and

synthetic control. We impute counterfactual outcomes based on convex combinations of observed outcomes, defined by an optimal coupling between the treated and control data sets. Finding an optimal coupling is a convex relaxation to the combinatorial optimal matching problem, for which we propose efficient algorithms based on matrix scaling. Unlike existing imputation methods, we begin with a desirable aggregate-level summary and estimate granular-level individual treatment effects by properly constraining the coupling so that the estimated individual effects are consistent with the aggregate summary. We provided a method to construct individual confidence intervals for the estimated counterfactual outcomes, along with a simulation study to demonstrate the effectiveness of our method. The simulation confirms our theoretical insight on the level of entropic regularization needed. We establish that entropic regularization plays a crucial role in both efficiency in inference and computation: first, the convexified matching objective is gauged for minimizing the width of the individual confidence intervals, trading off bias and variance; second, the entropic regularization enables us to design fast algorithms. We demonstrated the empirical performance of our method on the NSW data, using both experimental and nonexperimental data sets, with a comparison to existing methods and robustness checks.

5.6 Supplemental Proofs

5.6.1 Proof of Proposition 5.3.3

Proof of Proposition 5.3.3. For any $p = (p_1, \ldots, p_n) \in \Delta_n$, we have

$$\sum_{i=1}^{n} p_i \log(p_i) + \log(n) \leq n \sum_{i=1}^{n} p_i^2 - 1,$$

where the left-hand side is the Kullback-Leibler (KL) divergence of p from $\frac{1}{n}1_n$ which is known to be bounded by the χ^2 distance of p from $\frac{1}{n}1_n$ on the right-hand side; see Lemma

2.7 of Tsybakov [2009]. Meanwhile, the KL divergence is bounded below by the Hellinger distance; see Lemma 2.4 of Tsybakov [2009]. Therefore, we have

$$\sum_{i=1}^{n} p_i \log(p_i) + \log(n) \geq \sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{1/n})^2,$$

where the right-hand side is the Hellinger distance between p and $\frac{1}{n}1_n$. The right-hand side can be further lower bounded,

$$\begin{aligned}\sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{1/n})^2 &= \sum_{i=1}^{n} \frac{(p_i - 1/n)^2}{(\sqrt{p_i} + \sqrt{1/n})^2} \\ &\geq \frac{n \sum_{i=1}^{n} p_i^2 - 1}{2(n\|p\|_\infty + 1)}.\end{aligned}$$

Combining these inequalities, we have

$$\frac{1}{2(n\|p\|_\infty + 1)} \leq \frac{\sum_{i=1}^{n} p_i \log(p_i) + \log(n)}{n \sum_{i=1}^{n} p_i^2 - 1} \leq 1.$$

□

5.6.2 Proof of Proposition 5.4.2

Proof of Proposition 5.4.2. We have already discussed that (5.4.1) must admit a unique minimizer $\pi^\star \in \Pi_{a,b}^{+}$. We show that $\pi^\star$ is a unique fixed point of T_λ by using the Karush-Kuhn-Tucker (KKT) conditions. To see this, let us first rewrite (5.4.1) as

$$\begin{aligned}\min_{\pi \in \mathbb{R}_+^{n \times m}} \quad & (g(\pi) + \lambda h(\pi)) \\ \text{subject to} \quad & \pi 1_m = a \quad \text{and} \quad \pi^\top 1_n = b.\end{aligned}$$

Then, define the Lagrangian $L\colon \mathbb{R}_+^{n\times m} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ by

$$\begin{aligned} &L(\pi, \mu, \nu) \\ &\quad = g(\pi) + \lambda h(\pi) + \langle \mu, \pi 1_m - a \rangle + \langle \nu, \pi^\top 1_n - b \rangle. \end{aligned}$$

The KKT conditions for $(\pi, \mu, \nu) \in \mathbb{R}_+^{n\times m} \times \mathbb{R}^n \times \mathbb{R}^m$ are $\pi \in \Pi_{a,b}$ and π is a minimizer of $L(\cdot, \mu, \nu)$ over $\mathbb{R}_+^{n\times m}$. Due to h, we can see that $L(\cdot, \mu, \nu)$ must admit a unique minimizer at the interior of $\mathbb{R}_+^{n\times m}$, namely, all the entries of the unique minimizer are strictly positive. Accordingly, the minimizer must satisfy the first-order condition:

$$\begin{aligned} \nabla_\pi L(\pi, \mu, \nu) &= \nabla g(\pi) + \lambda \nabla h(\pi) + \mu 1_m^\top + 1_n \nu^\top \\ &= 0, \end{aligned}$$

which leads to

$$\pi = \exp\left(-\frac{\mu 1_m^\top + 1_n \nu^\top + \nabla g(\pi)}{\lambda} \right). \tag{5.6.1}$$

Hence, $(\pi, \mu, \nu) \in \mathbb{R}_+^{n\times m} \times \mathbb{R}^n \times \mathbb{R}^m$ satisfies the KKT conditions if $\pi \in \Pi_{a,b}$ and (5.6.1) holds. Meanwhile, for any $(\pi, \mu, \nu) \in \mathbb{R}_+^{n\times m} \times \mathbb{R}^n \times \mathbb{R}^m$, if the right-hand side of (5.6.1) belongs to $\Pi_{a,b}$, it must be $\Phi_\lambda(\nabla g(\pi)) = T_\lambda(\pi)$ which can be deduced from Theorem 1 of Sinkhorn [1967].

In summary, $(\pi, \mu, \nu) \in \mathbb{R}_+^{n\times m} \times \mathbb{R}^n \times \mathbb{R}^m$ satisfies the KKT conditions if

$$\pi = \exp\left(-\frac{\mu 1_m^\top + 1_n \nu^\top + \nabla g(\pi)}{\lambda} \right) = T_\lambda(\pi).$$

Meanwhile, for (5.4.1), we can check that the Slater condition—see for instance, Section 5.2.3 of Boyd and Vandenberghe [2004]—is satisfied, and thus we have an optimizer $(\mu^\star, \nu^\star) \in \mathbb{R}^n \times \mathbb{R}^m$ of the dual problem with strong duality. Accordingly, the KKT conditions are satisfied by $(\pi^\star, \mu^\star, \nu^\star)$. This implies that $\pi^\star = T_\lambda(\pi^\star)$. Hence, T_λ has a fixed point $\pi^\star$. For

uniqueness, suppose $\pi = T_\lambda(\pi)$ holds for some $\pi \in \Pi_{a,b}^+$. Again, by Theorem 1 of Sinkhorn [1967], the matrix $T_\lambda(\pi) = \Phi_\lambda(\nabla g(\pi))$ must take the form of the right-hand side of (5.6.1) for some $(\mu, \nu) \in \mathbb{R}^n \times \mathbb{R}^m$, meaning that the KKT conditions are satisfied by (π, μ, ν) and thus π is a minimizer of (5.4.1). As we have already shown that (5.4.1) admits a unique minimizer $\pi^\star$, we conclude $\pi = \pi^\star$. Hence, $\pi^\star$ is the unique fixed point of T_λ. □

5.6.3 Proof of Lemma 5.4.5

Proof of Lemma 5.4.5. We first claim that for any $\pi \in \Pi_{a,b}$, we have

$$\langle C_1 + \lambda \nabla h(\Phi_\lambda(C_1)), \pi - \Phi_\lambda(C_1) \rangle \geq 0. \tag{5.6.2}$$

Suppose not, namely, there exists $\pi \in \Pi_{a,b}$ such that

$$\langle C_1 + \lambda \nabla h(\Phi_\lambda(C_1)), \pi - \Phi_\lambda(C_1) \rangle < 0.$$

In other words, letting $q(\pi) := \langle C_1, \pi \rangle + \lambda h(\pi)$, we have $\pi \in \Pi_{a,b}$ such that

$$\langle \nabla q(\Phi_\lambda(C_1)), \pi - \Phi_\lambda(C_1) \rangle < 0.$$

As q is differentiable at $\Phi_\lambda(C_1)$ and q is convex on $\Pi_{a,b}$, this means that we can find a point π' on the line segment between $\Phi_\lambda(C_1)$ and π such that $q(\pi') < q(\Phi_\lambda(C_1))$, which contradicts that $\Phi_\lambda(C_1)$ is a minimizer of q over $\Pi_{a,b}$. Hence, (5.6.2) must hold for any $\pi \in \Pi_{a,b}$. Letting $\pi = \Phi_\lambda(C_2)$ in (5.6.2), we have

$$\langle C_1 + \lambda \nabla h(\Phi_\lambda(C_1)), \Phi_\lambda(C_2) - \Phi_\lambda(C_1) \rangle \geq 0.$$

Changing the role of C_1 and C_2,

$$\langle C_2 + \lambda \nabla h(\Phi_\lambda(C_2)), \Phi_\lambda(C_1) - \Phi_\lambda(C_2) \rangle \geq 0.$$

Combining the above two inequalities, we have

$$\begin{aligned} &\langle \nabla h(\Phi_\lambda(C_1)) - \nabla h(\Phi_\lambda(C_2)), \Phi_\lambda(C_1) - \Phi_\lambda(C_2) \rangle \\ &\quad \leq -\frac{1}{\lambda} \langle C_1 - C_2, \Phi_\lambda(C_1) - \Phi_\lambda(C_2) \rangle. \end{aligned} \tag{5.6.3}$$

By applying Hölder's inequality to the right-hand side of (5.6.3), we have

$$\begin{aligned} &-\frac{1}{\lambda} \langle C_1 - C_2, \Phi_\lambda(C_1) - \Phi_\lambda(C_2) \rangle \\ &\quad \leq \frac{\|C_1 - C_2\|_\infty \cdot \|\Phi_\lambda(C_1) - \Phi_\lambda(C_2)\|_1}{\lambda}. \end{aligned} \tag{5.6.4}$$

Also, as h is 1-strongly convex with respect to $\|\cdot\|_1$ on $\Delta_{n,m}$, we have for any $P_1, P_2 \in \Delta_{n,m}$,

$$\langle \nabla h(P_1) - \nabla h(P_2), P_1 - P_2 \rangle \geq \|P_1 - P_2\|_1^2,$$

which allows us to lower bound the left-hand side of (5.6.3) as follows:

$$\begin{aligned} &\langle \nabla h(\Phi_\lambda(C_1)) - \nabla h(\Phi_\lambda(C_2)), \Phi_\lambda(C_1) - \Phi_\lambda(C_2) \rangle \\ &\quad \geq \|\Phi_\lambda(C_1) - \Phi_\lambda(C_2)\|_1^2. \end{aligned} \tag{5.6.5}$$

Combining (5.6.3), (5.6.4), and (5.6.5), we have (5.4.4). □

5.7 Applications: Revisit the NSW Data

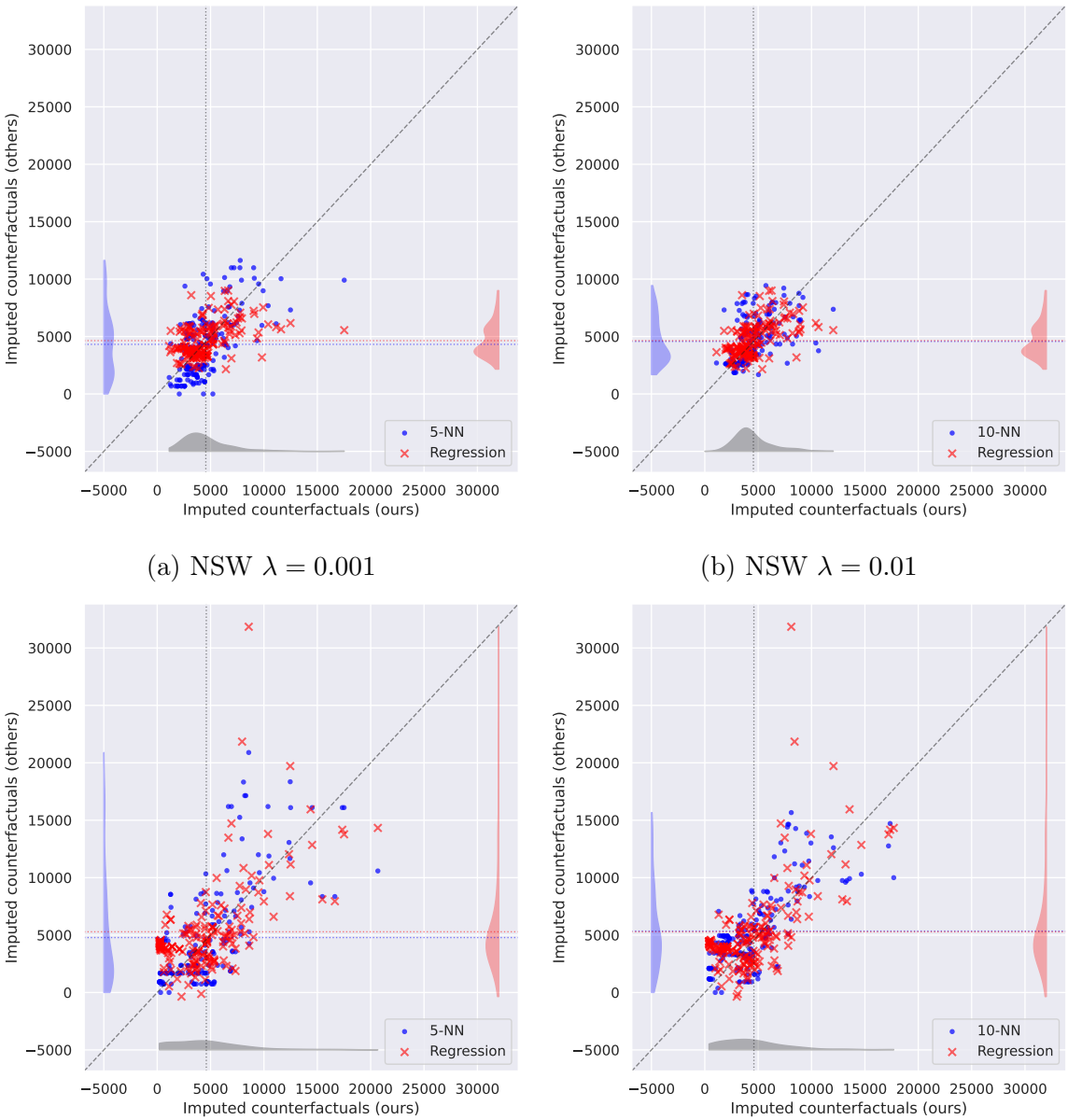
In this section, we numerically investigate our imputation and individualized inference method using the NSW data and compare it with other methods, continued from Section 5.2.3.

5.7.1 Imputation

We conduct a robustness check of the imputation results in Section 5.2.3 on the NSW data by specifying different nonlinear kernels. We consider the RBF kernel and the polynomial kernel of degree two. The results are shown in Figure 5.4 and Figure 5.5. Though the overall patterns are similar to those using the linear kernel, we can observe some differences. First, for the experimental data, the imputed counterfactual outcomes via the RBF kernel with $\lambda = 0.001$ show the largest variability. In Figure 5.4(a), there are two treated units whose imputed counterfactual outcomes are above 15000, which do not occur in the results via the linear or polynomial kernel; as a result, the right tail of the histogram along the x-axis is longer than the other cases. For the PSID data, we can see that using the polynomial kernel leads to the most dispersed imputed outcomes, visualized by the right tails of the histograms along the x-axis in (a) and (b) of Figure 5.5 which are longer than the other cases. We point out that unlike the NN-matching method and our KSC method, whose counterfactual distributions are supported on $\mathbb{R}_+$ even for the nonexperimental PSID data, the regression imputation method presents a wider support for imputed values, even with negative outcomes as seen across Figures 5.1, 5.4 and 5.5.

5.7.2 Inference

We apply the method presented in Section 5.3 to construct confidence intervals around the imputed counterfactual outcomes. Figure 5.6 (a) shows the results for the experimental data using the linear kernel with $\lambda \in \{0.001, 1\}$, which plots the estimated individual treatment effects (ITEs), $\widehat{\tau}_j = Y_j - \widehat{Y}_j(0)$, against the logarithm of earnings in 1975, where the confidence intervals are shown as error bars around the estimated ITEs. Here, only 74 treated units (out of 185 treated units) with positive earnings in 1975 are shown. We pick this index merely for visualization of individualization. For comparison, we also show the blue points representing the ITEs based on imputing all the treated units with the mean of the control



(a) NSW $\lambda = 0.001$

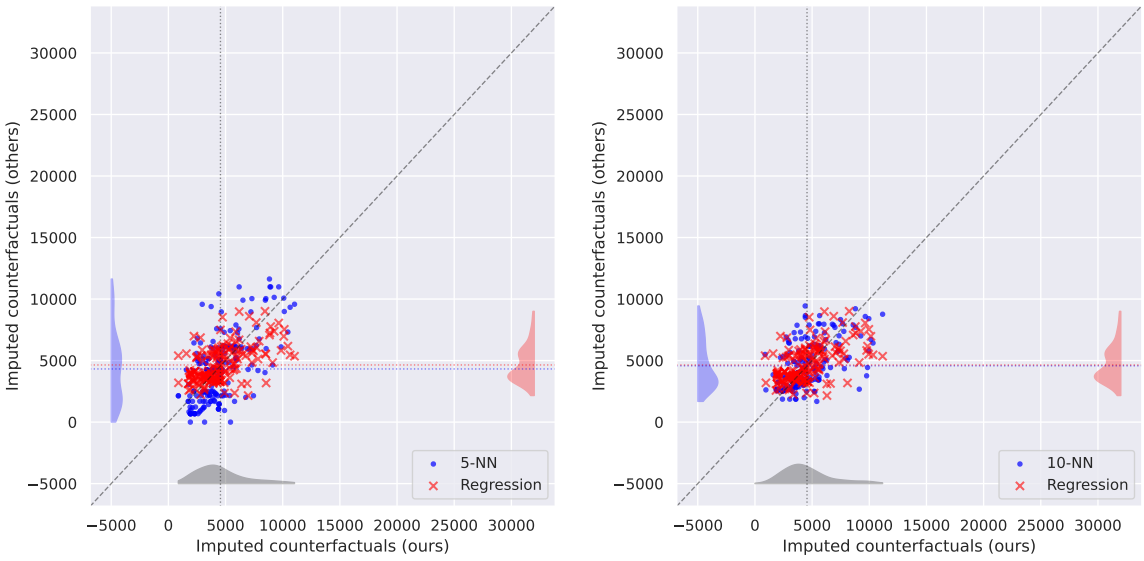
(b) NSW $\lambda = 0.01$

(c) PSID $\lambda = 0.001$

(d) PSID $\lambda = 0.01$

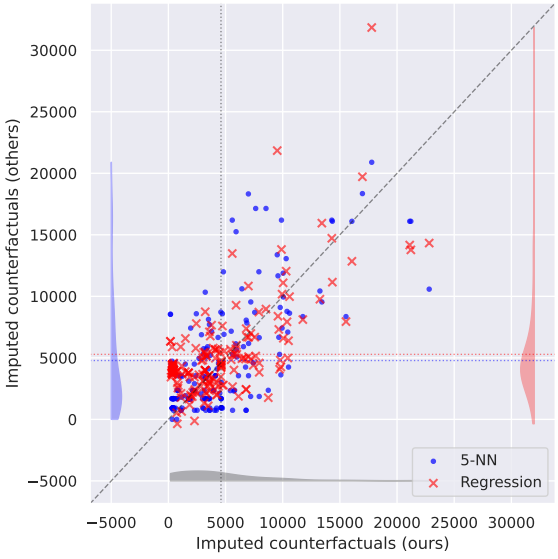
Figure 5.4: Scatter plots of the imputed counterfactual outcomes of the treated along with the histograms of the marginal distributions. The setup is the same as in Figure 5.1, but the results are based on the RBF kernel for both NSW and PSID data.

group outcomes, along with the confidence intervals based on the standard error of the mean; the case where there is no individualization in the imputed control outcomes, and thus no meaningful individual confidence intervals addressing the bias. Let us first focus on

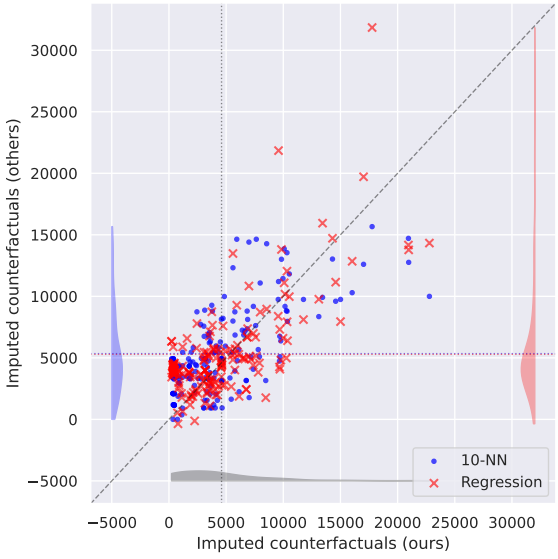


(a) NSW $\lambda = 0.001$

(b) NSW $\lambda = 0.01$



(c) PSID $\lambda = 0.001$



(d) PSID $\lambda = 0.01$

Figure 5.5: Scatter plots of the imputed counterfactual outcomes of the treated along with the histograms of the marginal distributions. The setup is the same as in Figure 5.1, but the results are based on the polynomial kernel of degree two for both NSW and PSID data.

the point estimates of the ITEs. We can see that most of the estimated ITEs fall in the interval $(-10000, 10000)$, seemingly uncorrelated with the earnings in 1975, while there are seven individuals whose ITEs are above 10000. These seven units with the largest ITEs are

likely to be the ones who benefit most from the treatment, standing out from the rest of the treatment group.

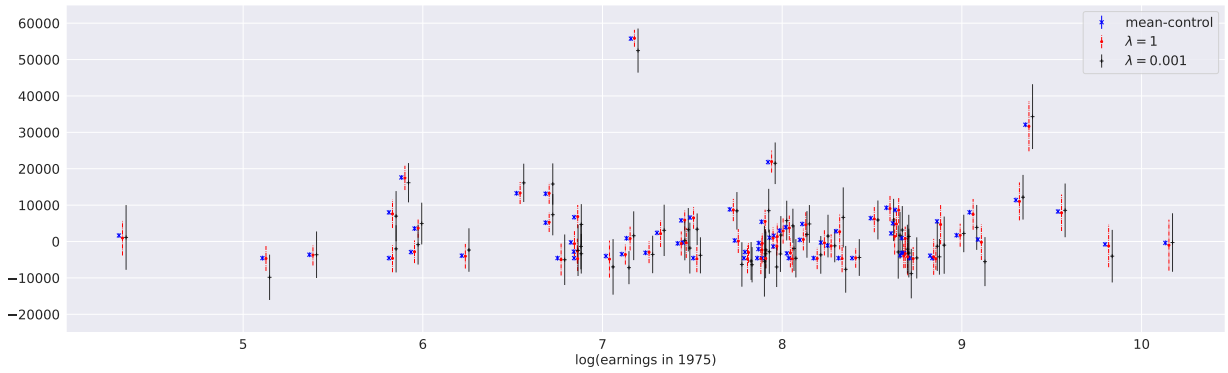
The confidence intervals help us decide how sure we are about these seemingly significant ITEs. We can confidently say that the top two ITEs are extremely large compared to the rest of the ITEs since the lower ends of their individual confidence intervals are larger than the upper ends of the rest of the ITEs, regardless of the choice of λ. Now, looking at the confidence intervals, rather than only looking at the point estimates, the four units whose ITEs are between 10000 and 20000 are less distinguishable from the rest. The error bars do help to suggest which units benefit more here. For the third largest ITE, slightly above 20000, the conclusion may vary depending on λ. If we read the confidence intervals corresponding to $\lambda = 1$ (red, dashed), this unit seems to be more significant than the rest, while this conclusion is less assertive for $\lambda = 0.001$ as the confidence interval of this unit overlaps with the interval $(-15000, 20000)$ which contains the rest of the confidence intervals. In other words, using $\lambda = 0.001$ leads to the most conservative conclusion regarding who benefits the most from the treatment as the confidence intervals are wider. Unlike the confidence intervals by the proposed method, the intervals based on the mean imputation (shown in blue) are extremely narrow, which does not take into account individual heterogeneity, overlooking a potentially large bias; the confidence interval here is useless for individual decision making and may result in overly optimistic conclusions, for instance, essentially the majority benefits significantly from the program.

(b) and (c) of Figure 5.6 repeat the same analysis using the RBF kernel and the polynomial kernel of degree two, respectively. The overall patterns are similar to those using the linear kernel, but the confidence intervals tend to be wider. Accordingly, even the second largest ITE above 30000 is not as distinguishable from the rest as in the linear kernel case.

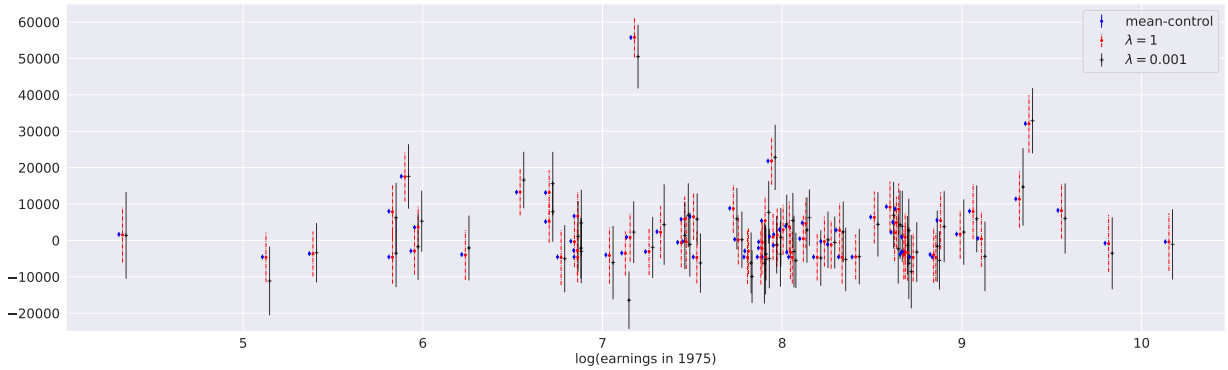
Figure 5.7 shows the results using the PSID data. Notice that the intervals are generally much wider than those of the NSW experimental data. One possible explanation is that the

PSID data, even after trimmed, differs significantly from the experimental data, leading to the matching with more conservative error estimates; particularly, the standard deviations of the observed outcomes of the experimental and PSID are 5483.8 and 10700.9, respectively. Accordingly, the confidence intervals are wider, which makes it harder to reach a decisive conclusion with confidence.

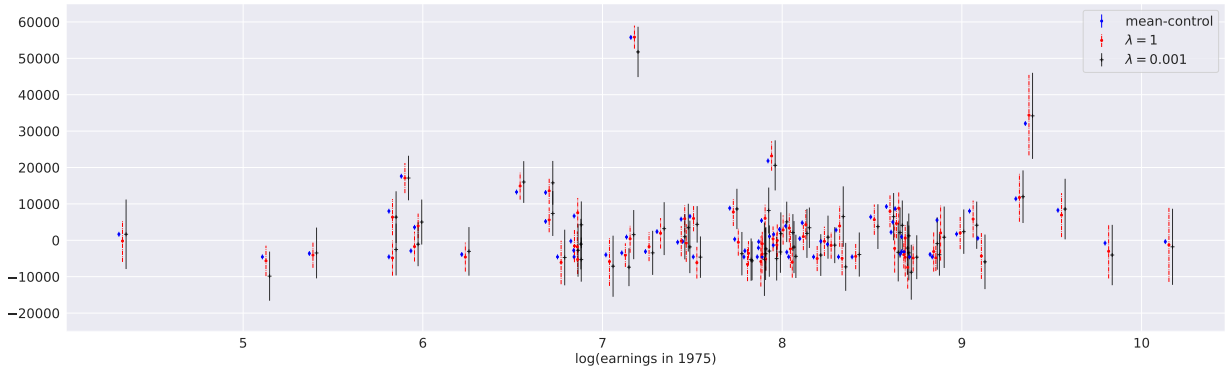
Lastly, we briefly comment on the choice of ρ, the regularization parameter of the kernel ridge regression mentioned in Section 5.3.2. Increasing ρ leads to a smaller estimate of $\|f_0\|_{\mathcal{H}}$, leading to a smaller bias correction. Increasing ρ will typically result in a larger estimate of σ_0, implying a large variability. The correct choice of ρ depends on the underlying signal-to-noise ratio for the data-generating process in the control outcomes, namely, the complexity of the function vs. the amount of noise. In all examples of this section, we choose ρ using 5-fold cross-validation.



(a) Linear

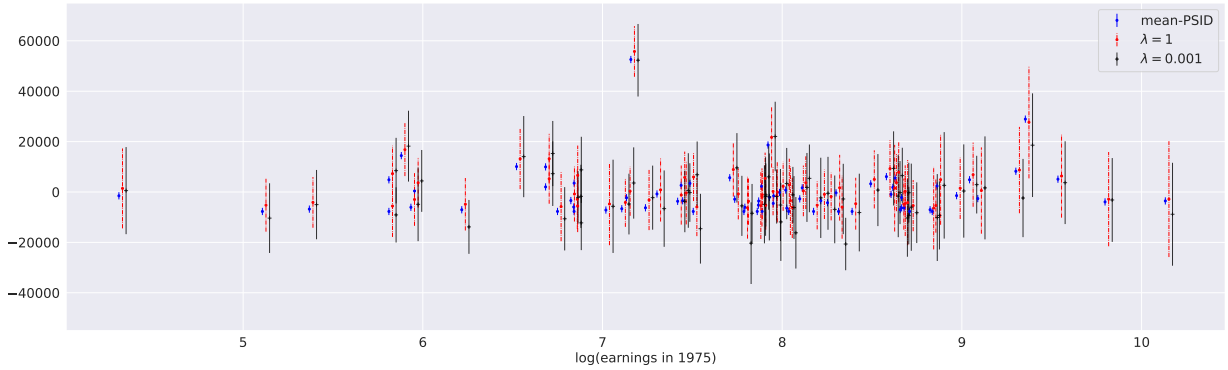


(b) RBF

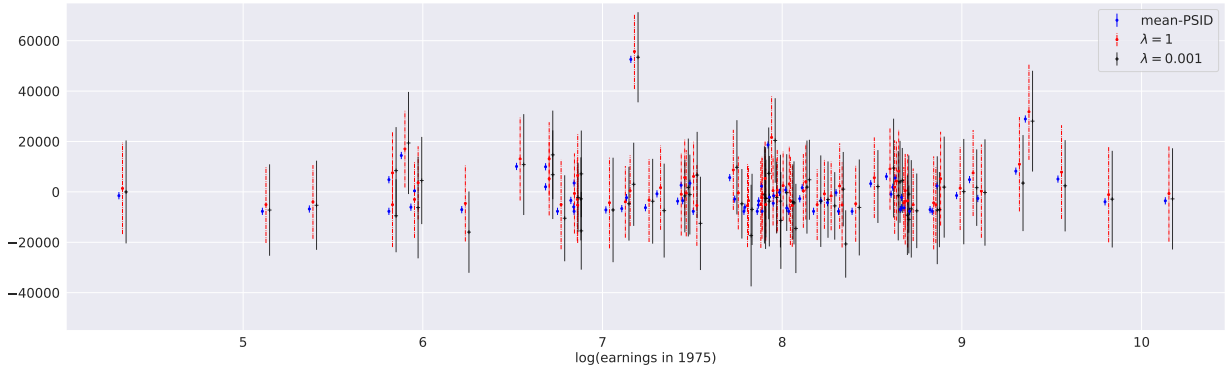


(c) Polynomial

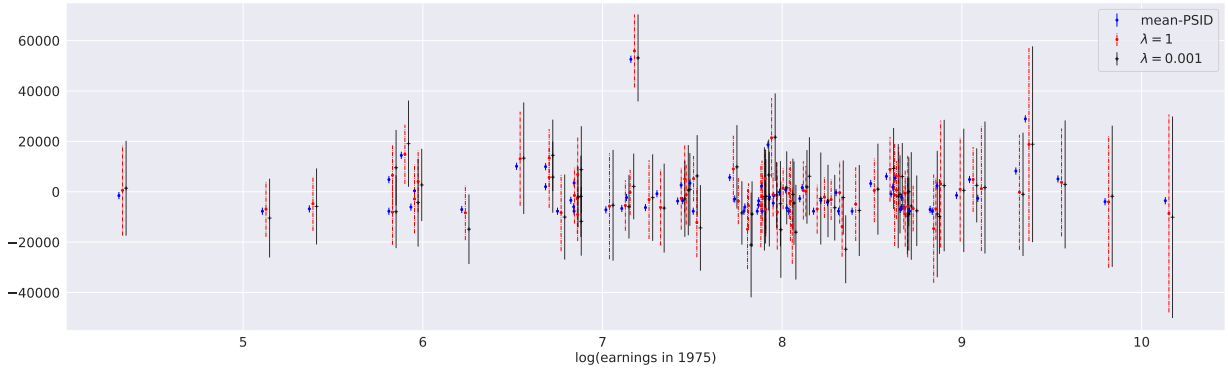
Figure 5.6: NSW experimental data: estimated individual treatment effects (ITEs) of the treated along the logarithm of earnings in 1975 on the x-axis, with the 95% confidence intervals shown as error bars around the estimated ITEs. 74 treated units (among 185 treated units) with positive earnings in 1975 are shown. The results are based on the experimental data using uniform weights v, w in (5.2.7) with three different kernels: linear, RBF, and polynomial of degree two. The blue points are imputing all the treated units with the mean of the control group outcomes, where the blue error bars show the confidence interval based on the standard error of the mean.



(a) Linear



(b) RBF



(c) Polynomial

Figure 5.7: PSID nonexperimental data: estimated individual treatment effects (ITEs) of the treated along the logarithm of earnings in 1975 on the x-axis, with the 95% confidence intervals shown as error bars around the estimated ITEs. The setup is the same as in Figure 5.6, but the results are based on the trimmed PSID data. For the weights v, w in (5.2.7), we let v be uniform, while w is the propensity score-based weights as in (5.2.8). The blue points and error bars are obtained as in Figure 5.6, but with the trimmed PSID data instead of the NSW control group.

REFERENCES

Alberto Abadie. Using Synthetic Controls: Feasibility, Data requirements, and Methodological Aspects. *Journal of Economic Literature*, 59(2):391–425, 2021.

Alberto Abadie and Javier Gardeazabal. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 93(1):113–132, 2003.

Alberto Abadie and Guido W. Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267, 2006.

Alberto Abadie and Guido W. Imbens. Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics*, 29(1):1–11, 2011.

Alberto Abadie and Jérémy L'hour. A Penalized Synthetic Control Estimator for Disaggregated Data. *Journal of the American Statistical Association*, 116(536):1817–1834, 2021.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the Dark: An Efficient Algorithm for Bandit Linear Optimization. In *Conference on Learning Theory*, 2008.

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-Linear Time Approximation Algorithms for Optimal Transport via Sinkhorn Iteration. In *Advances in Neural Information Processing Systems*, 2017.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2005.

T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193 – 212, 1952.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning*, 2017.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and Equilibrium in Generative Adversarial Nets (GANs). In *International Conference on Machine Learning*, 2017.

Susan Athey and Guido Imbens. Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Yu Bai, Tengyu Ma, and Andrej Risteski. Approximability of Discriminators Implies Diversity in GANs. *arXiv preprint arXiv:1806.10586*, 2018.

Raef Bassily, Mikhail Belkin, and Siyuan Ma. On Exponential Convergence of SGD in Non-Convex Over-Parametrized Learning. *arXiv preprint arXiv:1811.02564*, 2018.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on Treatment Effects after Selection among High-Dimensional Controls. *Review of Economic Studies*, 81(2):608–650, 2014.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. Synthetic Controls with Staggered Adoption. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2): 351–381, 2022.

Jean-David Benamou and Yann Brenier. A Computational Fluid Mechanics Solution to the Monge-Kantorovich Mass Transfer Problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Andrew J Blumberg, Mathieu Carriere, Michael A Mandell, Raul Rabadan, and Soledad Villar. MREC: A Fast and Versatile Framework for Aligning and Matching Point Clouds with Applications to Single Cell Molecular Data. *arXiv preprint arXiv:2001.01666*, 2020.

Sergey Bobkov and Michel Ledoux. *One-Dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances*. American Mathematical Society, 2019.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Claire Brécheteau. A Statistical Test of Isomorphism Between Metric-Measure Spaces Using the Distance-To-A-Measure Signature. *Electronic Journal of Statistics*, 13(1):795–849, 2019.

Yann Brenier. Décomposition Polaire et Réarrangement Monotone des Champs de Vecteurs. *Comptes Rendus des Séances de l'Académie des Sciences. Série I. Mathématique*, 305(19): 805–808, 1987.

Yann Brenier. The Least Action Principle and the Related Concept of Generalized Flows for Incompressible Perfect Fluids. *Journal of the American Mathematical Society*, 2(2): 225–255, 1989.

Yann Brenier. Polar Factorization and Monotone Rearrangement of Vector-Valued Functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Yann Brenier and Wilfrid Gangbo. L^p Approximation of Maps by Diffeomorphisms. *Calculus of Variations and Partial Differential Equations*, 16(2):147–164, 2003.

Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka. Learning Generative Models across Incomparable Spaces. In *International Conference on Machine Learning*, 2019.

T.T. Cai and Yihong Wu. Optimal Detection of Sparse Mixtures Against a Given Null Distribution. *IEEE Transactions on Information Theory*, 60(4):2217–2232, 2014.

T.T. Cai, X Jessie Jeng, and Jiashun Jin. Optimal Detection of Heterogeneous and Heteroscedastic Mixtures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(5):629–662, 2011.

Matias D Cattaneo. Efficient Semiparametric Estimation of Multi-Valued Treatment Effects under Ignorability. *Journal of Econometrics*, 155(2):138–154, 2010.

E. Çela. *The Quadratic Assignment Problem: Theory and Algorithms*. Springer, 1998.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Statistical Guarantees of Generative Adversarial Networks for Distribution Estimation. *arXiv preprint arXiv:2002.03938*, 2020.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review*, 107(5):261–265, 2017.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally Robust Semiparametric Estimation. *Econometrica*, 90(4):1501–1535, 2022.

Raj Chetty, John N Friedman, Michael Stepner, and the Opportunity Insights Team. The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data. *The Quarterly Journal of Economics*, 139(2):829–889, 2024.

Samir Chowdhury and Facundo Mémoli. The Gromov–Wasserstein Distance Between Networks and Stable Network Invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.

Samir Chowdhury, David W. Miller, and Tom Needham. Quantized Gromov-Wasserstein. In *ECML PKDD*, 2021.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

S Csörgő. Weighted Correlation Tests for Location-Scale Families. *Mathematical and Computer Modelling*, 38(7-9):753–762, 2003.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, 2013.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with Optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Steven De Rooij, Tim Van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the Leader If You Can, Hedge If You Must. *Journal of Machine Learning Research*, 15(1): 1281–1316, 2014.

T de Wet. Goodnes-Of-Fit Tests for Location and Scale Families Based on a Weighted L_2-Wasserstein Distance Measure. *Test*, 11:89–107, 2002.

Rajeev H Dehejia and Sadek Wahba. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

Rajeev H Dehejia and Sadek Wahba. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics*, 84(1):151–161, 2002.

Eustasio Del Barrio, Evarist Giné, and Frederic Utzet. Asymptotics for L_2 Functionals of the Empirical Quantile Process, with Applications to Tests of Fit Based on Weighted Wasserstein Distances. *Bernoulli*, 11(1):131–189, 2005.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the Gap Between Constant Step Size Stochastic Gradient Descent and Markov Chains. *The Annals of Statistics*, 48 (3):1348 – 1382, 2020.

Carl Doersch. Tutorial on Variational Autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

David Donoho and Jiashun Jin. Higher Criticism for Detecting Sparse Heterogeneous Mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

David Donoho and Jiashun Jin. Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects. *Statistical Science*, 30(1):1–25, 2015.

David L. Donoho and Alon Kipnis. Higher Criticism to Compare Two Large Frequency Tables, with sensitivity to Possible Rare and Weak Differences. *The Annals of Statistics*, 50(3):1447–1472, 2022.

Théo Dumont, Théo Lacombe, and François-Xavier Vialard. On the Existence of Monge Maps for the Gromov-Wasserstein Distance. *Foundations of Computational Mathematics*, pages 1–48, 2024.

Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training Generative Neural Networks via Maximum Mean Discrepancy Optimization. In *Uncertainty in Artificial Intelligence*, 2015.

Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2010.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the Intrinsic Dimension of Datasets by a Minimal Neighborhood Information. *Scientific Reports*, 7(1): 1–8, 2017.

Max H Farrell. Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations. *Journal of Econometrics*, 189(1):1–23, 2015.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep Neural Networks for Estimation and Inference. *Econometrica*, 89(1):181–213, 2021.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating Between Optimal Transport and MMD using Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Roland G Fryer, Steven D Levitt, John A List, et al. Parental Incentives and Early Childhood Achievement: A Field Experiment in Chicago Heights. Technical report, National Bureau of Economic Research, 2015.

Manuela Gellert, Md Faruq Hossain, Felix Jacob Ferdinand Berens, Lukas Willy Bruhn, Claudia Urbainsky, Volkmar Liebscher, and Christopher Horst Lillig. Substrate Specificity of Thioredoxins and Glutaredoxins — Towards a Functional Classification. *Heliyon*, 5(12): e02943, 2019.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic Optimization for Large-scale Optimal Transport. In *Advances in Neural Information Processing Systems*, 2016.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning Generative Models with Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2014.

Geoffrey J Gordon. Regret Bounds for Prediction Problems. In *Conference on Computational Learning Theory*, 1999.

Christian Gouriéroux and Alain Monfort. *Simulation-Based Econometric Methods*. Oxford University Press, 1997.

Wenxuan Guo, YoonHaeng Hur, Tengyuan Liang, and Chris Ryan. Online Learning to Transport via the Minimal Selection Principle. In *Conference on Learning Theory*, pages 4085–4109, 2022.

Jinyong Hahn. On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, pages 315–331, 1998.

Jens Hainmueller. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20(1):25–46, 2012.

Marc Hallin, Gilles Mordant, and Johan Segers. Multivariate Goodness-Of-Fit Tests Based on Wasserstein Distance. *Electronic Journal of Statistics*, 15(1):1328 – 1371, 2021.

Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-Tight VC-Dimension Bounds for Piecewise Linear Neural Networks. In *Conference on Learning Theory*, 2017.

Megan L Head, Luke Holman, Rob Lanfear, Andrew T Kahn, and Michael D Jennions. The Extent and Consequences of p-Hacking in Science. *PLoS biology*, 13(3):e1002106, 2015.

James J Heckman and Edward Vytlacil. Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73(3):669–738, 2005.

James J Heckman, Hidehiko Ichimura, and Petra E Todd. Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies*, 64(4):605–654, 1997.

Austin Bradford Hill. *Principles of Medical Statistics*. Oxford University Press, 1955.

Keisuke Hirano and Guido W Imbens. Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology*, 2:259–278, 2001.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, 71(4):1161–1189, 2003.

Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

YoonHaeng Hur and Yuehaw Khoo. Robust Point Matching with Distance Profiles. *arXiv preprint arXiv:2312.12641*, 2023.

YoonHaeng Hur and Tengyuan Liang. A Convexified Matching Approach to Imputation and Individualized Inference. *arXiv preprint arXiv:2407.05372*, 2024.

YoonHaeng Hur and Tengyuan Liang. Detecting Weak Distribution Shifts via Displacement Interpolation. *Journal of Business & Economic Statistics*, 43(1):178–190, 2025.

YoonHaeng Hur, Wenxuan Guo, and Tengyuan Liang. Reversible Gromov–Monge Sampler for Simulation-Based Inference. *SIAM Journal on Mathematics of Data Science*, 6(2): 283–310, 2024.

Jan-Christian Hütter and Philippe Rigollet. Minimax Estimation of Smooth Optimal Transport Maps. *The Annals of Statistics*, 49(2):1166 – 1194, 2021.

Guido Imbens and Yiqing Xu. LaLonde (1986) after Nearly Four Decades: Lessons Learned. *arXiv preprint arXiv:2406.00827*, 2024.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jyrki Kivinen and Manfred K. Warmuth. Exponentiated Gradient versus Gradient Descent for Linear Predictors. *Information and Computation*, 132(1):1–63, 1997.

Andrej N Kolmogorov. Sulla Determinazione Empirica di Una Legge di Distribuzione. *Giornale dell'Istituto Italiano degli Attuar*, 4:89–91, 1933.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K. Rohde. Optimal Mass Transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

Tjalling C. Koopmans and Martin Beckmann. Assignment Problems and the Location of Economic Activities. *Econometrica*, 25(1):53–76, 1957.

Noémi Kreif, Richard Grieve, Dominik Hangartner, Alex James Turner, Silviya Nikolova, and Matt Sutton. Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units. *Health Economics*, 25(12):1514–1528, 2016.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From Word Embeddings to Document Distances. In *International Conference on Machine Learning*, 2015.

Robert J LaLonde. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review*, pages 604–620, 1986.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, third edition, 2005.

Qi Lei, Jason D Lee, Alexandros G Dimakis, and Constantinos Daskalakis. SGD Learns One-Layer Networks in WGANs. *arXiv preprint arXiv:1910.07030*, 2019.

Steven D Levitt, John A List, Susanne Neckermann, and Sally Sadoff. The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. *American Economic Journal: Economic Policy*, 8(4):183–219, 2016.

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative Moment Matching Networks. In *International Conference on Machine Learning*, 2015.

Tengyuan Liang. Estimating Certain Integral Probability Metric (IPM) Is as Hard as Estimating under the IPM. *arXiv preprint arXiv:1911.00730*, 2019.

Tengyuan Liang. How Well Generative Adversarial Networks Learn Distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021.

Tengyuan Liang and Benjamin Recht. Randomization Inference When N Equals One. *arXiv preprint arXiv:2310.16989*, 2023.

Tengyuan Liang and James Stokes. Interaction Matters: A Note on Non-Asymptotic Local Convergence of Generative Adversarial Networks. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Keli Liu and Xiao-Li Meng. There Is Individualized Treatment. Why Not Individualized Inference? *Annual Review of Statistics and Its Application*, 3:79–111, 2016.

Thomas MaCurdy, Xiaohong Chen, and Han Hong. Flexible Estimation of Treatment Effect Parameters. *American Economic Review*, 101(3):544–551, 2011.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal Transport Mapping via Input Convex Neural Networks. In *International Conference on Machine Learning*, 2020.

Robert J. McCann. A Convexity Principle for Interacting Gases. *Advances in Mathematics*, 128(1):153–179, 1997.

Daniel McFadden. A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. *Econometrica*, 57(5):995–1026, 1989.

H. Brendan McMahan and Matthew Streeter. Adaptive Bound Optimization for Online Convex Optimization. *arXiv preprint arXiv:1002.4908*, 2010.

Facundo Mémoli. Gromov-Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11:417–487, 2011.

Facundo Mémoli and Tom Needham. Distance Distributions and Inverse Problems for Metric Measure Spaces. *Studies in Applied Mathematics*, 149(4):943–1001, 2022.

Facundo Mémoli and Tom Needham. Comparison Results for Gromov–Wasserstein and Gromov–Monge Distances. *ESAIM: Control, Optimisation and Calculus of Variations*, 30:78, 2024.

Gaspard Monge. Mémoire sur la Théorie des Déblais et des Remblais. *Mémoires de l'Académie Royale des Sciences*, pages 666–704, 1781.

Youssef Mroueh. Wasserstein Style Transfer. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. *arXiv preprint arXiv:1711.04894*, 2017.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Axel Munk and Claudia Czado. Nonparametric Validation of Similar Distributions and Assessment of Goodness of Fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(1):223–241, 1998.

Jerzy Neyman. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Translated from Polish to English with discussion in *Statistical Science*, 5(4):465–472, 1990, 1923.

Francesco Orabona. A Modern Introduction to Online Learning. *arXiv preprint arXiv:1912.13213*, 2021.

Francesco Orabona and Dávid Pál. Scale-Free Algorithms for Online Linear Optimization. In *Conference on Algorithmic Learning Theory*, 2015.

Francesco Orabona and Dávid Pál. Scale-Free Online Learning. *Theoretical Computer Science*, 716:50–69, 2018.

Felix Otto. The Geometry of Dissipative Evolution Equations: The Porous Medium Equation. *Communications in Partial Differential Equations*, 26(1&2):101–174, 2001.

Ariel Pakes and David Pollard. Simulation and the Asymptotics of Optimization Estimators. *Econometrica*, 57(5):1027–1057, 1989.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In *International Conference on Machine Learning*, 2016.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein Barycenter and Its Application to Texture Mixing. In *Scale Space and Variational Methods in Computer Vision*, pages 435–446, 2012.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1511.06434*, 2015.

Alexander Rakhlin and Karthik Sridharan. Statistical Learning and Sequential Prediction, 2014. Available at `https://www.mit.edu/~rakhlin/courses/stat928/stat928_notes.pdf`.

Michael W Robbins, Jessica Saunders, and Beau Kilmer. A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention. *Journal of the American Statistical Association*, 112(517):109–126, 2017.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 1999.

Paul R Rosenbaum. Model-Based Direct Adjustment. *Journal of the American statistical Association*, 82(398):387–394, 1987.

Paul R. Rosenbaum. Modern Algorithms for Matching in Observational Studies. *Annual Review of Statistics and Its Application*, 7:143–176, 2020.

Paul R Rosenbaum and Donald B Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55, 1983.

H. L. Royden. *Real Analysis*. Macmillan, third edition, 1988.

Donald B Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Donald B. Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, 2006.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

Adil Salim, Anna Korba, and Giulia Luise. The Wasserstein Proximal Gradient Algorithm. In *Advances in Neural Information Processing Systems*, 2020.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-Time Gromov Wasserstein Distances Using Low Rank Couplings and Costs. *arXiv preprint arXiv:2106.01128*, 2021.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-Scale Optimal Transport and Mapping Estimation. In *International Conference on Learning Representations*, 2018.

Shai Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Shai Shalev-Shwartz and Yoram Singer. A Primal-Dual Perspective of Online Learning Algorithms. *Machine Learning*, 69(2):115–142, 2007.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Galen R. Shorack and Jon A. Wellner. *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics, 2009.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.

Shashank Singh and Barnabás Póczos. Minimax Distribution Estimation in Wasserstein Distance. *arXiv preprint arXiv:1802.08855*, 2018.

Richard Sinkhorn. Diagonal Equivalence to Matrices with Prescribed Row and Column Sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

Nickolay Smirnov. Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948.

Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic Metric Alignment for Correspondence Problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the Empirical Estimation of Integral Probability Metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Charles J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

Matthew Streeter and H. Brendan McMahan. Less Regret via Online Conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*, 2012.

Gábor J. Székely and Maria L. Rizzo. Energy Statistics: A class of Statistics Based on Distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Amirhossein Taghvaei and Amin Jalali. 2-Wasserstein Approximation via Restricted Convex Potentials with Application to Improved Training for GANs. *arXiv preprint arXiv:1902.07197*, 2019.

Vayer Titouan, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. Sliced Gromov-Wasserstein. In *Advances in Neural Information Processing Systems*, 2019.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Mark J Van der Laan and James M Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2003.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

Tim van Erven, Peter Grunwald, Wouter M. Koolen, and Steven de Rooij. Adaptive Hedge. In *Advances in Neural Information Processing Systems*, 2011.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

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.

Jonathan Weed and Quentin Berthet. Estimation of Smooth Densities in Wasserstein Distance. In *Conference on Learning Theory*, 2019.

Christoph Alexander Weitkamp, Katharina Proksch, Carla Tameling, and Axel Munk. Distribution of Distances Based Object Matching: Asymptotic Inference. *Journal of the American Statistical Association*, 119(545):538–551, 2024.

Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching. In *Advances in Neural Information Processing Systems*, 2019.

Hongyi Zhang and Suvrit Sra. First-Order Methods for Geodesically Convex Optimization. In *Conference on Learning Theory*, 2016.

Martin Zinkevich. Online Convex Programming and Generalized Infinitesimal Gradient Ascent. In *International Conference on Machine Learning*, 2003.

José R Zubizarreta, Elizabeth A Stuart, Dylan S Small, and Paul R Rosenbaum. *Handbook of Matching and Weighting Adjustments for Causal Inference*. CRC Press, 2023.